

Scalable Computing: Practice and Experience

Scientific International Journal
for Parallel and Distributed Computing

ISSN: 1895-1767



Volume 25(5)

September 2024

EDITOR-IN-CHIEF

Dana Petcu

West University of Timisoara, Romania

SENIOR EDITOR

Marcin Paprzycki

Systems Research Institute of the Polish Academy of Sciences, Poland

EXECUTIVE EDITOR

Katarzyna Wasielewska-Michniewska

Systems Research Institute of the Polish Academy of Sciences, Poland

TECHNICAL EDITOR

Silviu Panica

Institute e-Austria Timisoara, Romania

EDITORIAL BOARD

Peter Arbenz, Swiss Federal Institute of Technology,

Giacomo Cabri, University of Modena and Reggio Emilia,

Philip Church, Deakin University,

Frederic Desprez, INRIA Grenoble Rhône-Alpes and LIG laboratory,

Yakov Fet, Novosibirsk Computing Center,

Giancarlo Fortino, University of Calabria,

Gianluca Frasca-Caccia, University of Salerno,

Fernando Gonzalez, Florida Gulf Coast University,

Dalvan Griebler, Pontifical Catholic University of Rio Grande do Sul,

Frederic Loulergue, University of Orleans,

Svetozar Margenov, Institute for Parallel Processing and Bulgarian Academy of Science,

Fabrizio Marozzo, University of Calabria,

Gabriele Mencagli, University of Pisa,

Viorel Negru, West University of Timisoara,

Wiesław Pawłowski, University of Gdańsk,

Shahram Rahimi, Mississippi State University,

Wilson Rivera-Gallego, University of Puerto Rico,

SUBSCRIPTION INFORMATION: please visit <http://www.scp.e.org>

Scalable Computing: Practice and Experience

Volume 25, Number 5, September 2024

TABLE OF CONTENTS

PAPERS IN THE SPECIAL ISSUE ON MACHINE LEARNING FOR SMART SYSTEMS: SMART BUILDING, SMART CAMPUS, AND SMART CITY:

- The Trajectory Data Mining Model for College Students in Campus Life and Academic Management** **3225**
Wugang Liu

PAPERS IN THE SPECIAL ISSUE ON DEEP LEARNING-BASED ADVANCED RESEARCH TRENDS IN SCALABLE COMPUTING:

- Multi Channel Electronic Communication Signal Parameters based on Nonlinear Phase Principle Modulation and Deep Learning** **3241**
Xiaoqing Yan

- Application of Artificial Intelligence Technology and Deep Learning in Laboratory Intelligent Management Platform** **3251**
Xing Lu

- A Machine Intelligence Evaluation System Based on Internet Automation Technology and Deep Learning** **3259**
Hongchuan Liu

- Research on Intelligent Building Integrated Cabling System Based on Internet of Things and Deep Learning** **3268**
Rong Zhou

- The Application of Intelligent Robots and Deep Learning in the Construction Management Platform System of Construction Engineering** **3277**
Yandong Zhou

- Obstacle Avoidance Path Planning for Power Inspection Robots based on Deep Learning Algorithms** **3288**
Yuxin Liu, Xiaoxi Ge, Haowei Jia, Lin Yuan, Min Zhou

- Analysis of Frozen Data Anomaly and Update Method of Electromechanical Energy Meter Terminal based on Deep Learning** **3296**
Fang Yao, Libin Tan

The Application of Deep Learning Intelligent Robots in the Design and Implementation of Information Retrieval Systems	3305
<i>Yuqi Miao</i>	
Application of Measurement Robots based on Deep Learning in Building Tilt Stability Monitoring	3314
<i>Wei Zhang, Junhua Li</i>	
Application of Software Robots and Deep Learning in Real time Processing of E-commerce Orders	3322
<i>Wenbo Niu, Yibo Hu, Wei Zhang</i>	
Intelligent Algorithm Operation and Data Management of Electromechanical Engineering Power Communication Network based on the Internet of Things	3330
<i>Ying Li, Wenjing Qu, Zhenqiang Zhang</i>	
Analysis of Abnormal Freezing Data and Updating Algorithm for Electromechanical Energy Meter Terminals	3342
<i>Shuzhi Zhao, Yue Du, Shanshan He, Jiao Bian, Jiabo Shi</i>	
Process Testing and Algorithm Detection Analysis of Mechanical Strength of Electromechanical Coupling in the Main Drive of Rolling Mill	3355
<i>Dongbao Zeng, Hai Yao</i>	
Research on Power Line Communication Based on Deep Learning for Electromechanical Equipment Electricity Acquisition Terminals	3366
<i>Chengfei Qi, Xiaobo Yang, Xiaokun Yang, Chaoran Bi, Wenwen Li</i>	
A Fault Monitoring System for Mechanical and Electrical Equipment of Subway Vehicles Based on Big Data Algorithms	3376
<i>Geng Li, Ya Li, Hongxue Bi</i>	
 PAPERS IN THE SPECIAL ISSUE ON SCALABLE COMPUTING IN ONLINE AND BLENDED LEARNING ENVIRONMENTS: CHALLENGES AND SOLUTIONS:	
Leveraging Emotions in Student Feedback to Improve Course Content and Delivery	3388
<i>Abid Hussain Wani</i>	
Feature Extraction of Gymnastics Images Based on Multi-scale Feature Fusion Algorithm	3394
<i>Kun Tian, Qionghua Xia</i>	

PAPERS IN THE SPECIAL ISSUE ON SCALABLE DEW COMPUTING FOR FUTURE GENERATION
IOT SYSTEMS :

IoT-Driven Hybrid Deep Collaborative Transformer with Federated Learning for Personalized E-Commerce Recommendations: An Optimized Approach 3408

Abdulmajeed Alqhatani, Surbhi Bhatia Khan

Brain Tumor Classification using Region-based CNN with Chicken Swarm Optimization 3427

A Sravanthi Peddinti, Suman Maloji, Kasiprasad Mannepalli

Improving Data Security and Scalability in Healthcare System using Blockchain Technology 3440

K.R. Rohini, P.S. Rajakumar, S. Geetha

Enhanced Throttled Load Balancing for Virtual Machine Allocation in Multiple Data Centers 3453

P. Hanumantha Rao, P.S. Rajakumar

Design and Development of an Unmanned/Autonomous Ocean Surface Vehicle using Self-Sustaining Dual Renewable Energy Harvesting System 3468

Kamalahasan M, Manivannan S, Swapna B

PAPERS IN THE SPECIAL ISSUE ON GRAPH POWERED BIG AEROSPACE DATA PROCESSING
:

Hybrid Electric Vehicle Energy Management Strategy based on Genetic Algorithm 3476

Yingzhe Luo, Chaoxiong Fan

Data Protection and Privacy Protection of Advertising based on Cloud Computing Platform 3484

Zhishe Chen

Design of 0-day Vulnerability Monitoring and Defense Architecture based on Artificial Intelligence Technology 3491

Jian Hu, Zhenhong Zhang, Feilu Hang, Linjiang Xie

A Hybrid Image Fusion and Denoising Algorithm based on Multi-scale Transformation and Signal Sparse Representation 3500

Dajun Sheng

Digital Media Internet Modeling System under Computer Artificial Intelligence Technology 3507

Miaojun Li, Qi Li, Changrong Peng, Xiaodong Zhang

E-commerce Data Mining Analysis based on User Preferences and Association Rules	3515
<i>Yun Zhang</i>	
High-resolution Holographic Image Reconstruction based on Deep Learning	3523
<i>Yun Zhang</i>	
Optimization of E-commerce Product Recommendation Algorithm Based on User Behavior	3531
<i>Yifan Ji, Lan Chen, Rui Xiong</i>	
Application of Multi-objective Optimization Algorithm based on Artificial Fish School Algorithm in Financial Investment Portfolio Problems	3540
<i>Hongxing Zhang</i>	
Optimization Algorithm for Green Environment Design Based on Artificial Intelligence	3547
<i>Kai Qian</i>	
Image Recognition Technology Based on Deep Learning in Automation Control Systems	3554
<i>Jingjing Wang</i>	
The Construction and Application of Residential Building Information Model Based on Deep Learning Algorithms	3563
<i>Shuang Zhao, Yu Yang</i>	
Human-computer Interaction Interface Design in the Cab of New Energy Vehicles	3572
<i>Yanxue Zhang, Nanmei Zhang, Hui Yang, Yanyu Ren</i>	
Application of Cluster Analysis Algorithm in Supply Chain Risk Identification	3580
<i>Qingping Zhang, Yi He</i>	
Transformer Fault Diagnosis and Location Method Based on Fault Tree Analysis	3587
<i>Zhiwu Wu, Tianfu Huang, Chunguang Wang, Xiang Wu, Yanzhao Tu</i>	
Ethical Evaluation and Optimization of Artificial Intelligence Algorithms Based on Self-supervised Learning	3594
<i>Ruoyu Deng, Yang Zhao</i>	
Machine Learning Algorithms in Supply Chain Coordination Simulation and Optimization	3603
<i>Qingping Zhang, Yi He</i>	

Simulation of Segmented Clustering of Cloud Storage Data Based on Neural Network Models and Python 3614

Guoqing Xia, Huazhen Chen

Plateau Altitude Disaster Prevention and Reduction Platform based on Beidou System 3626

Guoqing Xia, Huazhen Chen

State Monitoring and Anomaly Detection Algorithms for Electricity Meters Based on IoT Technology 3633

Chunguang Wang, Tianfu Huang, Zhiwu Wu, Ying Zhang, Hanbin Huang

Optimization of Logistics Distribution Network based on Ant Colony Optimization Neural Network Algorithm 3641

Jing Yang

PAPERS IN THE SPECIAL ISSUE ON SOFT COMPUTING AND ARTIFICIAL INTELLIGENCE FOR WIRE/WIRELESS HUMAN-MACHINE INTERFACE :

Review of Automated Test Case Generation, Optimization, and Prioritization using UML Diagrams: Trends, Limitations, and Future Directions 3651

Srinivasa Rao Kongarana, A Ananda Rao, P Radhika Raju

High Speed Low Power Analysis of 12 Transistors 2×4 line Decoder using 45GPDK Technology 3674

Sruthi Pavani Javvadi, C R S Hanuman, Sivadurgarao Parasa, Sannajaji Naraganeni

Multi Objective Data Transformation in Hybrid Clouds Networks for Offloading Data 3691

V Sridhar Reddy, N. Jayanthi, Sharon Rose Victor Juvvanapudi, Srinivas Bachu, Madipalli Sumalatha

Optimizing Task Scheduling: Exploring Advanced Machine Learning in Dew-Powered Cloud Environments 3701

A. Ganesh, K Sree Divya, Chinthakunta Sasikala, E.Poornima, Nidamanuru Srinivasa Rao, A.V.L.N Sujith, G.Ramesh

A Secure Data Storage Approach for Online Examination Platform using Cloud DBAAS Services 3715

Srinu Banothu, G. Janardhan, G. Sirisha, Srinivasulu Shepuri, Madhavi Karnam, Allam Balaram

Machine Learning based Tool Wear Prediction from Variability of Acoustic Sound Emission Signals 3725

N V. Krishnamoorthy, Joseph Vijay

Retrieval of Telugu Word from Hand Written Text using Densenet-CNN **3741**

Rajasekhar Boddu, Edara Sreenivasa Reddy

PAPERS IN THE SPECIAL ISSUE ON INTERNET OF THINGS AND AUTONOMOUS UNMANNED AERIAL VEHICLE TECHNOLOGIES FOR SMART AGRICULTURE RESEARCH AND PRACTICE :

UAV Path Planning Model Leveraging Machine Learning and Swarm Intelligence for Smart Agriculture **3752**

Roberto E. Roque-Claros, Deivi P. Flores-Llanos, Abel R. Maquera-Humpiri, Vijaya Krishna Sonthi, Sudhakar Sengan, Rajasekar Rangasamy

Smart Fertilizing Using IoT Multi-Sensor and Variable Rate Sprayer Integrated UAV **3766**

Hayder M. A. Ghanimi, R. Suguna, Josephine Pon Gloria Jeyaraj, K Sreekanth, Rajasekar Rangasamy, Sudhakar Sengan

PAPERS IN THE SPECIAL ISSUE ON RECENT ADVANCE SECURE SOLUTIONS FOR NETWORK IN SCALABLE COMPUTING :

Recurrent Neural Network based Incremental model for Intrusion Detection System in IoT **3778**

Himanshu Sharma, Prabhat Kumar, Kavita Sharma

DiffCRNN: A Novel Approach for Detecting Sound Events in Smart Home Systems Using Diffusion-Based Convolutional Recurrent Neural Network **3796**

Maryam M. Al Dabel

PAPERS IN THE SPECIAL ISSUE ON EVOLUTIONARY COMPUTING FOR AI-DRIVEN SECURITY AND PRIVACY: ADVANCING THE STATE-OF-THE-ART APPLICATIONS :

Scalable and Distributed Mathematical Modeling Algorithm Design and Performance Evaluation in Heterogeneous Computing Clusters **3812**

Zhouding Liu, Jia Li

Copyright Protection and Risk Assessment Based on Information Extraction and Machine Learning: The Case of Online Literary Works **3822**

Xudong Lin

Research on the Application of Node Importance Assessment based on HITs Algorithm in Power Grid Planning **3832**

Gaoshan Fu, Xiang Yin, Yue Gao, Dan Meng, Liang Che

Implementation and Optimization of Probabilistic and Mathematical Statistical Algorithms under Distributive Architecture	3841
<i>Shengbiao Li, Jiankui Peng</i>	
Missing Data Imputation for Health Care Big Data using Denoising Autoencoder with Generative Adversarial Network	3850
<i>Yinbing Zhang</i>	
Educational Big Data Analytics Using Sentiment Analysis for Student Requirement Analysis on Courses	3858
<i>Meida Wang, Qingfeng Yang</i>	
Green Plant Landscape Design for Urban Air Quality Purification with Computer Image Processing in Cloud, Grid, and Cluster Computing	3867
<i>Jingjing Ni</i>	
Learners Behaviour prediction and analysis model for smart learning platform using Deep Learning Approach	3876
<i>Liyuan Feng, Yunfeng Ji</i>	
Application of Intelligent Analysis based on Engineering Management and Decision Making for Economic Development of Regional Enterprise	3886
<i>Qianzhen Song, Tong Yao, Yuhong Dai</i>	
Blockchain-based E-commerce Marketing Strategy for Agricultural Supply Chain	3895
<i>Yingzi Xu, Li Yu</i>	
Next-Generation Connectivity: A Holistic Review of Cooperative NOMA in Dynamic Vehicular Networks for Intelligent Transportation Systems	3903
<i>Potula Sravani, Ijjada Sreenivasa Rao</i>	
PA Fuzzy-noise Removal in Wireless Sensors Networks	3925
<i>B Harish Goud, Raju Anitha</i>	
Machine Learning-based Risk Prediction and Safety Management for Outdoor Sports Activities	3934
<i>Yan Lu</i>	
Research on Planning and Path Optimization of Leisure Sports Activities based on Multi-objective Genetic Algorithm	3942
<i>Xu Yang</i>	
Research on Visualization and Interactivity of Virtual Reality Technology and Digital Media in Interior Space Design	3952
<i>Ke Zhang, Ratanachote Thienmongkol</i>	

Research on the Influencing Factors of Commercial Pension Insurance for Rural Residents in the Context of Population Aging Based on Big Data Analysis	3962
<i>Shitang Feng</i>	
Optimization of Weighting Algorithm in Enterprise HRMS based on Cloud Computing and Hadoop Platform	3970
<i>Genliang Zhao</i>	
Optimization of Computer Network Security System Based on Improved Neural Network Algorithm and Data Searching	3979
<i>Chongfeng Tian, Zhihao Chen, Yi Zhu Hongfei Lu Guoxiao Li Rongquan Li Wei Pan</i>	
Hand-drawn Illustration Design in National Wave Style Based on Deep Learning and Image Super-Resolution Reconstruction	3989
<i>Miaomiao Yu, Siti Salmi Binti Jamali, Adzira Binti Husain</i>	
Research on Grid Data Analysis and Intelligent Recommendation System by Introducing Neural Tensor Network Model	3996
<i>Rui Zhou, Kangqian Huang, Dejun Xiang Xin Hu</i>	
Research on the Design of a System based on Machine Learning Algorithms for Automatic Scoring of English Writing Ability	4005
<i>Shan Zhao</i>	
Research on Cryptography-based Data Security and Trustworthiness in Digital Construction of Water Resources and Hydropower	4014
<i>Chao Yue, Wei Liu, Licheng Chen, Chong Zuo</i>	
Research on Deep Learning-based Algorithm for Digital Image Combination and Target Detection	4023
<i>Shanlu Huang, Jialin Lai</i>	
Research on Learning Efficiency Improvement Strategies of Public English Perspective Based on Ant Colony Algorithm	4032
<i>Qingzhu Li</i>	
Research on the Application of MOOCs Based on Reinforcement Learning in College English Teaching	4041
<i>Yu Gu</i>	
Smart Fish Passage Design and Application of Hydroacoustic Communication Technology in Aquatic Ecosystem Restoration	4052
<i>Chao Yue, Menggen Zhu, Lei Yang, Lei Li</i>	
Design of Financial Data Analysis and Decision Support System based on Big Data	4061
<i>Sufang Zheng</i>	

PAPERS IN THE SPECIAL ISSUE ON DATA-DRIVEN OPTIMIZATION ALGORITHMS FOR SUSTAINABLE AND SMART CITY:

Research on the Construction of Intelligent System of Landscape in Science and Innovation Park of Smart City – based on the Concept of Smart Garden Design 4070

Ke Xie

Performance Analysis of Smart City Landscape Design and Planning Based on The Internet of Things 4083

Chao Kang, Yanting He, Jingjing Xu

Research on Data-driven Urban Intelligent Monitoring and Old City Reconstruction 4095

Yi Wang

Research and Application of a Dual Filtering Music Hybrid Recommendation Model Based on CatBoost Algorithm and DCN 4113

Juncai Hou

PAPERS IN THE SPECIAL ISSUE ON EFFICIENT SCALABLE COMPUTING BASED ON IOT AND CLOUD COMPUTING:

Design of Test Turntable Based on Fuzzy PID Algorithm and its Error Correction 4128

Li Tang, Zhou Liangfu

A Visual Webpage Information Extraction Framework for Competitive Intelligence System 4138

Zhiwei Zhang, Wenbo Qin, Haifeng Xu

PAPERS IN THE SPECIAL ISSUE ON UNLEASHING THE POWER OF EDGE AI FOR SCALABLE IMAGE AND VIDEO PROCESSING:

Adaptation of Scalable Neural Style Transfer to Improve Alzheimer's Disease Detection Accuracy 4153

Eid Albalawi

Brain Tumor Classification on MRI Images by using Classical Local Binary Patterns and Histograms of Oriented Gradients 4165

Srinivas Babu Gottipati, Gowri Thumbur

A Novel Hybrid Model to Detect and Classify Arrhythmia Using ECG and Bio-Signals 4177

Manjesh B N, Raja Praveen N

Optimizing Waste Reduction in Manufacturing Processes Utilizing IoT Data with Machine Learning Approach for Sustainable Production	4192
<i>Faisal Altarazi</i>	

PAPERS IN THE SPECIAL ISSUE ON HIGH-PERFORMANCE COMPUTING ALGORITHMS FOR MATERIAL SCIENCES:

A Vision-Based Analog Meter Reading Method for Inspection Robots	4205
<i>Jiacheng Li, Honglei Wang, Xishuo Zhu, Sijian Liu, Junsheng Zhang</i>	

Prediction Method of Rate of Penetration based on Fuzzy Support Vector Regression	4218
<i>Li Yang, Lishen Wang, Lili Bai, Wenfeng Sun</i>	

Research on Distributed Scheduling Algorithm for Virtual City Power Plants Based on Blockchain Technology	4228
<i>Qing Zhu, Yufeng Zhang, Jize Sun</i>	

A Nonlinear Convolutional Neural Network Algorithm for Autonomous Vehicle Lane Line Detection	4237
<i>Kanhui Lyu</i>	

Dimension Extraction of Remote Sensing Images in Topographic Surveying Based on Nonlinear Feature Algorithm	4246
<i>Yani Wang, Yinpeng Zhou, Bo Wang</i>	

Research on Intelligent Agriculture Based on Artificial Intelligence and Embedded Perception Algorithms	4255
<i>Xinhuan Zhao, Fang Zhang, Na Gao</i>	

The Application of Intelligent Welding Robots and Visual Detection Algorithms in Building Steel Structures	4265
<i>Wei Zhang, Junhua Li</i>	

Application of Physical Modeling and Virtual Simulation Technology in Measuring the Performance of Subway Train Tracking and Operation	4274
<i>Xiuxuan Wang, Hongwei Liang</i>	

PAPERS IN THE SPECIAL ISSUE ON SYNERGIES OF NEURAL NETWORKS, NEUROBOTICS, AND BRAIN-COMPUTER INTERFACE TECHNOLOGY: ADVANCEMENTS AND APPLICATIONS :

The Employment of Carbon Nanotubes in Biomedical Applications	4283
<i>Jafaar Fahad A. Rida</i>	

Driver Drowsiness Detection	4301
<i>Ann Zeki Ablahd, Alyaa Qusay Aloraibi, Suhair Abd Dawwod</i>	

Ensemble Transfer Learning for Botnet Detection in the Internet of Things	4312
<i>Ali Aalsaud, Shahab Wahhab Kareem, Raghad Zuhair Yousif, Ahmed Salahuddin Mohammed</i>	
Secure Medical Image Retrieval Using Fast Image Processing Algorithms	4323
<i>Sameer Abdulsttar Lafta, Amaal Ghazi Hamad Rafash, Noaman Ahmed Yaseen AL-Falahi, Hussein Abdulqader Hussein, Mohanad Mahdi Abdulkareem</i>	
On Soft Strongly b^*–Compactness and Soft Strongly b^*–Connectedness in Soft Topological Spaces	4335
<i>Saif Z. Hameed, Abdelaziz E. Radwan, Essam El-Seidy</i>	
PAPERS IN THE SPECIAL ISSUE ON DEEP LEARNING IN HEALTHCARE :	
Sports Event Data Management System and Its Application in Competition Organization	4343
<i>Zhenyu Li</i>	
The Integration of Personalized Training Program Design and Information Technology for Athletes	4351
<i>Penghui Hao, Kun Qian</i>	
Diagnosis and Treatment System based on Artificial Intelligence and Deep Learning	4360
<i>Xiaoxi Zheng, Qili Fan, Geng Wang</i>	
The Integration and Innovation of Sports Social Platforms and Information Technology	4368
<i>Yongjun Chen</i>	
The Application of Information Technology for Athlete Data Analysis and Automatic Generation of Training Plans	4376
<i>Shuli Yuan</i>	
Deep Learning Model Construction of Urban Planning Image Data Processing and Health Intelligence System	4383
<i>Can Xu</i>	
Sports Data Privacy Protection and Information Security Management	4390
<i>Biao Jin</i>	
Data Collection and Analysis based on Sensor Technology in Sports Training	4399
<i>Xianbin Shi, Huagang Zou</i>	

PAPERS IN THE SPECIAL ISSUE ON NEXT GENERATION PERVASIVE RECONFIGURABLE COMPUTING FOR HIGH PERFORMANCE REAL TIME APPLICATIONS :

Introduction to the Special Issue on Next Generation Pervasive Reconfigurable Computing for High Performance Real Time Applications 4407

C. Venkatesan, Yu-Dong Zhang, Chow Chee Onn, Yong Shi

REGULAR PAPERS :

ML-CSFR: A Unified Crop Selection and Fertilizer Recommendation Framework based on Machine Learning 4411

Amit Bholra, Prabhat Kumar

A New Multi-Robots Search and Rescue Strategy based on Penguin Optimization Algorithm 4428

Ouarda Zedadra, Amina Zedadra, Antonio Guerrieri, Hamid Seridi, Douaa Ghelis

REVIEW PAPERS :

Federated Learning for Internet of Medical Healthcare: Issues and Challenges 4442

Nikita Chelani, Shivam Tripathy, Malaram Kumhar, Jitendra Bhatia, Varun Saxena, Sudeep Tanwar, Anand Nayyar



THE TRAJECTORY DATA MINING MODEL FOR COLLEGE STUDENTS IN CAMPUS LIFE AND ACADEMIC MANAGEMENT

WUGANG LIU*

Abstract. The main objective of the study is to address the lack of comprehensive management technology in student campus life in universities. Starting from the life trajectory data of students in campus life, a trajectory mining model combining data mining technology and university information system is designed. In addition, an applied clustering algorithm is designed to classify different trajectory feature types. The research results show that in actual trajectory analysis, the categories of action trajectories from dormitories to canteens, from 0 to 4, are 87.64%, 87.86%, 86.97%, 88.63%, and 88.71%, respectively, which are the most matched effective action trajectories. It can be seen that the trajectory analysis model designed in the study is effective and can provide assistance for the comprehensive academic management of college students.

Key words: Data mining, Clustering, Academic management, Trajectory features

1. Introduction. With the increasing attention of the state to social talent cultivation in recent years, college students, an important source of national talent reserve, have gradually received extensive attention from all sectors of society. The traditional talent cultivation system lacks adaptability to personalized and practical talent cultivation, and it is difficult to meet the talent demand of today's society, and it has become a new way of talent cultivation to update the talent cultivation system by using the current wave of social informationization and data development [1-3]. The new practical talent cultivation system not only involves the application of information technology in college talent cultivation, but also involves the definition of modern talent cultivation in college. In the traditional education concept, academic achievement is the most important evaluation index for college students and the ultimate value of students' learning career. However, with the gradual convergence of university talent education and social talent demand, academic performance can no longer form a more comprehensive assessment of students. Students will also develop various campus learning activities such as school-enterprise joint practice, campus activities, part-time entrepreneurship, campus exchanges, etc. Meanwhile, students' learning habits in the process of efficient learning have also become one of the important factors to assess students' comprehensive quality [4-6]. It is difficult to assess the daily learning life of students because they are influenced by social networks and learning life landscape, and they are characterized by both group and diversity. Data mining technology provides a technical grip for this problem. By integrating data mining technology with the information record system of college students, it can track and manage students' campus life and study in a trajectory way, and then achieve adaptive management and efficient management [7-9]. Therefore, this study designs a trajectory mining model combining data mining technology and college information system from the perspective of students' academic trajectory, and achieves academic tracking and analysis by analyzing students' trajectories.

The innovation of this study is to extend data-driven student management from learning management to comprehensive management of campus life, and apply trajectory mining models to information systems. By comprehensively analyzing the life trajectory of students, targeted management is implemented.

2. Related Works. Lee S M's team conducted a follow-up study on the adjustment of dental hygiene students to campus life and proposed appropriate management strategies. The study analyzed students' adjustment to campus activities in terms of their club participation, personality, professional adaptation, and interpersonal relationships. The results of the study showed that the campus life management strategy developed by the study can effectively improve the students' adaptability to campus life [10]. Li W's team developed

*School of Arts and Science, Nanning College of Technology, Nanning, 530100, China (wuganglw@163.com)

an intelligent campus management system based on IoT technology from the perspective of smart campus, which uses IoT face recognition technology as long as the data collector and realizes the tracking of students' campus life trajectory and campus life through standardized data analysis. management. The results of the study showed that the system is practical [11]. Kim Y's team analyzed the satisfaction of college students with campus life and analyzed the mediating role between students' social network consistency and satisfaction in campus life based on the survey data. The results of the study showed that the satisfaction of college students with campus life must be reflected in their self-efficacy through participation in campus social life [12]. Purnama S's team proposed a support system for college students' digital weaknesses in the learning process, which is based on the perspective of students' learning life and combines electronic devices and student-centered blockchain to enhance the digital capabilities of cooperative education while meeting the needs of university students' learning lives [13]. Way conducted a study on the important factors influencing students' learning behaviors in their daily learning lives in higher education, which is a combination of qualitative and quantitative analysis from the perspective of students' daily learning lives, emotional content, and temporal dimensions. The results of the study showed that online teaching and learning can fully complement offline teaching and learning and contribute to student outcomes [14].

In the development of data mining technology, its specific application in various fields is its main development situation. Haoxiang team applied data mining technology to online privacy data protection and used a perturbation algorithm to solve similar problems. There is also a significant improvement in the efficiency of the model, compared to other privacy-preserving algorithms [15]. Kuma applied data mining to finance and marketing and designed a data mining-based decision system for financial market information. This system analyzed organizational performance from a practical point of view and determined how the decision solution can be used to help companies balance competitive pressures under external environmental factors such as tax pressure and industrial costs. The results of the study showed the feasibility of this solution [16]. Edastama P's team proposed a data mining tool-based student data analysis system, which is a comprehensive data warehouse in the form of web information reports, and mined the characteristics and patterns of student data through basic data to finally achieve the effect of assessing the status of students' academic and campus life [1,7]. Ageed addressed the issue of combining data mining technology with cloud computing notation. The results of the study showed that the technique designed in the study effectively solves the cloud compatibility problems that arise when data mining is applied in parallel with cloud computing [18]. Mengash designed a data mining model for predicting the performance of college applicants in colleges and universities. Data mining model, which is combined with a reliable standardized admissions system, enables the prediction of possible post-admission learning outcomes of cohort students before they are admitted to colleges and universities. Over two thousand students were selected as the dataset to validate the model proposed in the study, and the predictive accuracy of the model was analyzed by tracking the actual academic performance of students after admission. The results of the study showed that the model designed in the study is able to predict the academic performance of students after enrollment, and such performance prediction can be used as a basis for student admission judgment [1,9].

Garg et al. proposed a decentralized evaluation system to address the issue of tampering in online education evaluation systems, to ensure the integrity of online education evaluation and further achieve the review of online education content. The research results showed that the system is practical [20]. Dutt et al. applied fuzzy set technology to learning neural network classification technology and proposed a digital learning assistance system for people with learning disabilities, achieving intelligent and personalized learning guidance. The research results showed that this method has a more efficient guidance function [21]. Choudhary et al. applied deep learning algorithms to personalized learning recommendation systems based on the extraction and analysis of user preference information, thereby improving the accuracy and personalization of recommendations for different users. The results showed that the system can effectively improve recommendation accuracy and personalization level [22].

It can be seen that data mining technology has unique advantages in the analysis of group data, and it can be better integrated with other systems. There have been some research examples of data mining technology applied in the university system so far. However, it can be found that the current data analysis of students in universities mainly focuses on the analysis of students' academic performance and learning status, but not

from the perspective of students' comprehensive campus life, which is relatively one-sided. Therefore, this study starts from students' life trajectory data in campus life, combines data mining technology with university information system, and designs a trajectory mining model to provide new ideas for students' data analysis.

This study designs a trajectory mining model that can manage students' comprehensive campus life by combining campus life trajectory data and big data mining strategies. The model collects information using different information endpoints of integrated information systems and processes trajectory characteristics based on this. This is a novel perspective and effective information tool for managing students' campus life. This study further designs an applied clustering algorithm and uses it to classify different trajectory feature types. The application of this method greatly enhances the accuracy and effectiveness of research and implementation of management strategies. This study can predict students whose behavior patterns may change through models, and predict the direction of changes, which has important practical significance for early warning and prevention of student behavior problems. Meanwhile, the applied clustering algorithm in this study is better at clustering data into groups with features. Compared to other clustering algorithms, it processes and classifies data more meticulously and accurately. Overall, this study provides a new perspective and effective means to assist universities in comprehensively managing students' academic lives by designing and implementing a model that excavates their daily academic life trajectories, combined with big data mining technology and existing university information systems.

3. Data Mining Model Design for College Students' Trajectories. These feature points include campus consumption records, campus network usage records, campus access control records, scholarship information, and student information. Through these data, it can roughly depict the campus image of students. When analyzing the characteristics of students' campus life trajectory, it needs to label the campus and its internal functional areas, as universities may have multiple campuses and multiple functional areas with similar functions within the same campus. Therefore, it labels the campus and differentiates its functions into twelve different categories. On this basis, the semantic trajectories of students will be further segmented, and the segmentation operation can appropriately divide the long-term trajectories of students. The main trajectory segmentation method used is time-threshold trajectory segmentation. After extracting the required student activity trajectory feature information, an academic management model based on clustering algorithm is proposed. The model is mainly divided into three modules, namely data clustering analysis, trajectory frequent pattern analysis, and trajectory deviation analysis. The model first uses the k-means algorithm for clustering analysis, and then uses the PrefixSpan strategy to perform frequent pattern analysis on the student trajectories within the cluster based on the clustering results. Finally, based on the output results of frequent trajectory patterns, it calculates the degree of deviation between the trajectories of individual students within the cluster and the trajectories of the cluster center, and provides academic warnings based on the degree of deviation to achieve previous academic management. Overall, this model extracts valuable information by analyzing students' behavioral trajectories, and then identifies common and abnormal patterns of student behavior through clustering and pattern analysis, thereby achieving early warning and management of students' academic performance.

3.1. Design of Campus Life Information Collection Model for College Students. In the campus life of college students, the factors that affect students' academic achievement are mainly divided into two categories, which are personal and impersonal factors. The personal factors include students' cognitive ability, creative ability and other inherent abilities, while the impersonal factors are based on the view of academic life and social network in which students live [23]. In the student trajectory feature analysis model part, the trajectory feature analysis model construction is shown in Fig 3.1.

From Fig 3.1, the trajectory feature analysis model converts and analyzes information feature points of students in campus life and academic management information feature points, respectively. The campus life information feature points include campus consumption records, campus network usage records, and campus access control records. The academic management information feature points include scholarship information and student information, etc. The dimension of information features is shown in Figure 3.2.

From this, three types of data collection contacts are derived. One is the campus information system, i.e., the system containing dormitory access control information, student campus network information, and student consumption information. The second is the basic student information and dormitory assignment information. The third is student academic performance and scholarship data. Starting from these three data

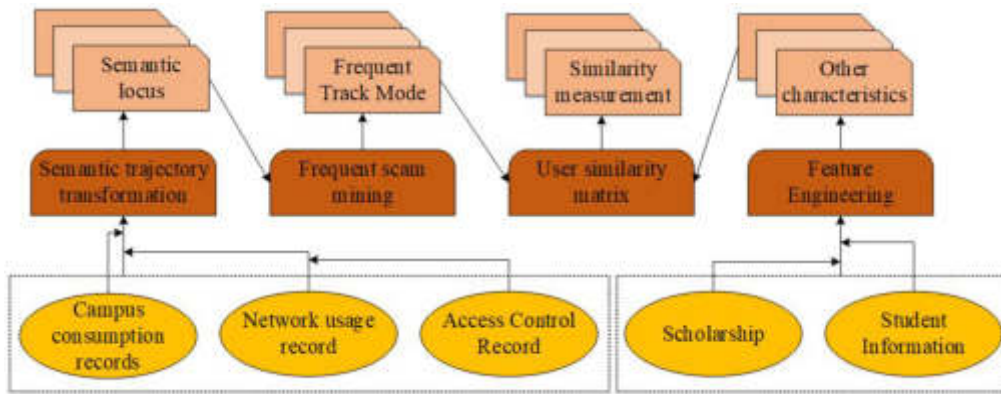


Fig. 3.1: Trajectory feature analysis model

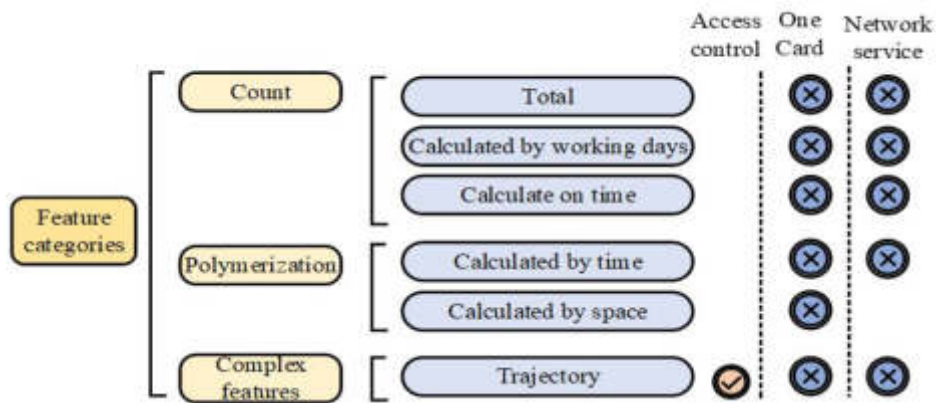


Fig. 3.2: Information feature dimension

dimensions, the model can basically outline the campus image of students when conducting data collection. The data collection process is based on quantitative behavior statistics, time statistics and frequency statistics as the main quantitative measures, and different quantitative statistics are used depending on the nature of the behavior. Quantitative behavior statistics refers to the quantitative count of a behavior or behavior results. Time statistics is used to measure the duration of a student’s behavior, while the frequency statistics is used to measure the number of times a student performs a particular behavior. The model quantitation values are collected in the manner shown in Fig 3.3.

In analyzing the trajectory characteristics of students’ campus life, it is necessary to mark the campus and the functional areas within the campus repeatedly, because the university may have more than one campus, and there may be several functional areas with similar functions within the same campus for diverting student traffic. Therefore, the campus is labeled as C_γ and the functional areas are classified into twelve different categories according to the functions they perform: classroom, dormitory, cafeteria, library, courtyard building, bathroom, office, hospital, supermarket, water room, multimedia area, and other areas, denoted by F_η . For the first λ location within the campus, it can be expressed by $p_{(\gamma,\lambda)}$, which can be defined by the location and spatial position in the form of equation (3.1).

$$P = (p.loc, p.fun) \tag{3.1}$$

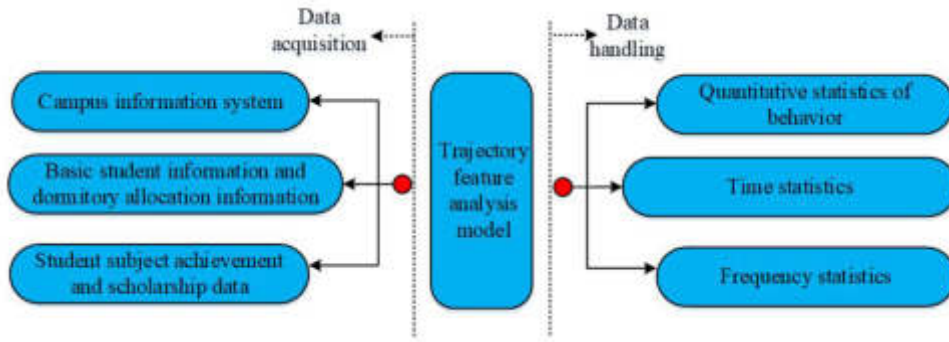


Fig. 3.3: Model quantization value acquisition method

In equation (??), $p.loc$ denotes the actual spatial location and $p.fun$ denotes the functional area. Based on this, the study defines student activity as a vector dependent on activity behavior, represented by $a.attr$. The student activity sequence is also defined as a spatio-temporal sequence around the individual student, as shown in equation (3.2).

$$Aseq = \{(t_1, p_1), \dots, (t_k, p_k)\} \quad (3.2)$$

In equation (3.2), t denotes the timestamp, $t_i < t_j$ ($i < j$), assuming the existence of a given active sequence with sequence parameters. p_1 and p_k are different spatio-temporal point locations. When both spatio-temporal point locations satisfy the constraints of Equation (3.3) at the same time, the two point locations can be judged as the same point location.

$$\begin{cases} p_i = p_{i+1} \\ |t_i - t_{i+1}| < \xi \end{cases} \quad (3.3)$$

In equation (3.3), ξ denotes the sequence parameters, t_i denotes the time, and p_i denotes the location. The model performs dwell point detection on the activity sequence, and then uses the dwell point data as the basis for trajectory compression, and finally outputs the semantic trajectory. In the semantic trajectory representation, the activity sequence of individual student and individual is fixed, and the relationship between trajectory Tra and activity sequence is shown in equation (3.4).

$$Tra \subseteq Aseq \quad (3.4)$$

3.2. Analysis Model of Campus Life Information Trajectory for College Students. . Due to the different daily routines of different students, there are periodic differences in their life trajectory information. Therefore, the system needs to effectively distinguish this differential information [24]. On this basis, the model will further segment the semantic trajectories of the students, and the segmentation operation can divide the trajectories of the students for a long time appropriately. The main trajectory segmentation methods can be divided into three types: time-threshold trajectory segmentation, set topology trajectory segmentation, and trajectory semantic trajectory segmentation. Since the student trajectories are based on the campus teaching and activity time as the main axis, the study adopts the time-threshold trajectory segmentation method. According to the time-threshold trajectory segmentation method, students' action trajectories are divided into day-based daily trajectories, and each segment of daily trajectories represents a day's travel of students. When dividing the daily trajectory, a day is not a day divided by a specific time point in physical time, but a complete activity of students in a basic time unit of a day is used as the basis for dividing the trajectory. If a student's activity trajectory exceeds the physical time boundary of a day, but is still within the complete activity of the day, then this part of the trajectory is still slid into the daily trajectory. An example of this classification is shown in Fig 3.4.

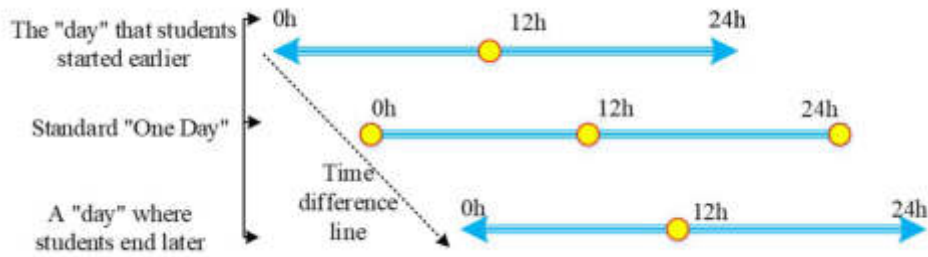


Fig. 3.4: Division example

The daily trajectory can be expressed in the form of equation (3.5).

$$DTra = \{(t_1, p_1), \dots, (t_s, p_s)\}, DTra \in Tra \tag{3.5}$$

In Tra , there exists one and only (t_j, p_j) , such that $(t_1, p_1) = (t_j, p_j), (t_{1+n}, p_{1+n}) = (t_{j+n}, p_{j+n})$. The imputation pattern is then shown in equation (3.6).

$$TraP = P'_1 \xrightarrow{\Delta t'_1} P'_2 \xrightarrow{\Delta t'_2} \dots \xrightarrow{\Delta t'_{u-1}} P'_u \tag{3.6}$$

Where $p'_i \in \{p_j\}$. If p'_i corresponds to (t_i, p_i) , and p'_{i+1} corresponds to (t_j, p_j) , then $\Delta t'_i = t_j - t_i$.

3.3.3 Academic Management Model Design. . After extracting the required student activity trajectory feature information, the study proposes an academic management model based on clustering algorithm. The model is divided into three main modules, which are data clustering analysis, trajectory frequent pattern analysis and trajectory deviation analysis. K-means can classify student trajectory data information based on data features [25]. The model first uses the k-means algorithm for clustering analysis, based on which the frequent pattern analysis of student trajectories within the clustered clusters is performed using the PrefixSpan strategy based on the clustering results. The PrefixSpan method flow is shown in Figure 3.5.

Finally, based on the output results of frequent trajectory patterns, it calculates the degree of deviation between the trajectories of individual students in the clusters and the trajectories of the cluster centers, and carries out academic warning according to the degree of deviation to achieve prior academic management. The specific structure is shown in Fig 3.6.

When the model performs the clustering operation, it assumes that there exists a base data set X and each data has a M dimensional feature vector X_n , then x_{nm} represents the feature value of the m feature of the n data. The clustering approach is shown in Fig 3.7.

The model divides the data instances into clusters as shown in equation (3.7).

$$C = \{C_1, C_2, \dots, C_k\} \tag{3.7}$$

There is no intersection between clusters, while each cluster has a cluster center, and the similarity between different clusters is relatively low, but the data instances inside the clusters are more similar. The sum of the distance between the data inside the cluster and the cluster center is the objective function, as shown in equation (3.8).

$$P(U, c) = \sum_{k=1}^K \sum_{n=1}^N u_{nk} \sum_{m=1}^M d(x_{nm}, c_{km}) \tag{3.8}$$

In equation (3.8), U denotes the matrix describing the affiliation status of the clusters, and u_{nk} denotes the affiliation status of the data instance with the ordinal number n for the ordinal number k . $d(x_{nm}, c_{km})$ denotes the distance between the center of the clustering cluster and the data instance. Since this distance

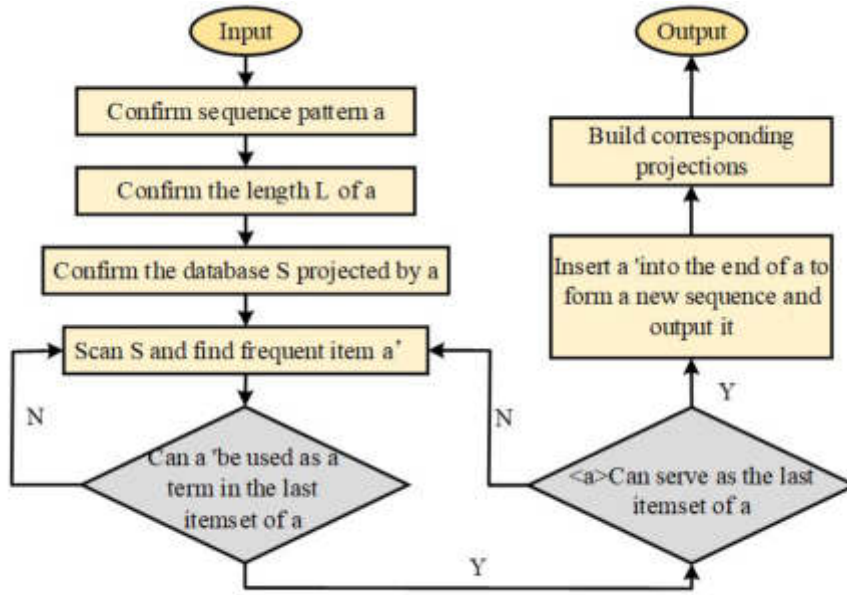


Fig. 3.5: The PrefixSpan method flow

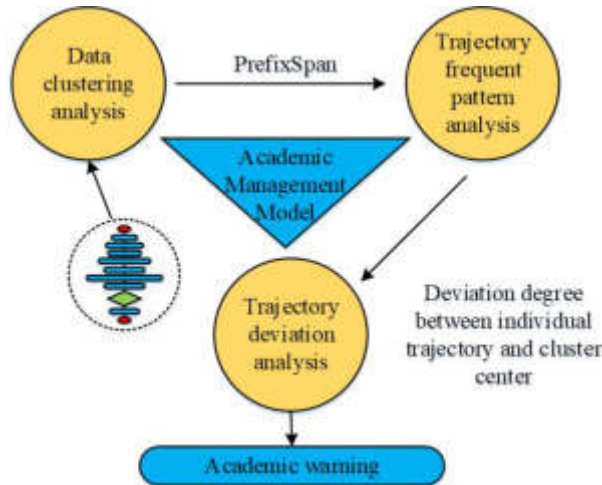


Fig. 3.6: Academic management model structure

yields different results with different metrics, the study requires a choice of metric for the model. The model mainly uses the Euclidean metric as the main metric, as shown in equation (3.9).

$$P(U, c) = \sum_{k=1}^K \sum_{n=1}^N u_{nk} \sum_{m=1}^M d(x_{nm} - c_{km})^2 \tag{3.9}$$

As can be seen from equation (3.9), the metric treats all data features equally, and differences in data feature differences lead to different clustering results, so a weighting mechanism needs to be added to the

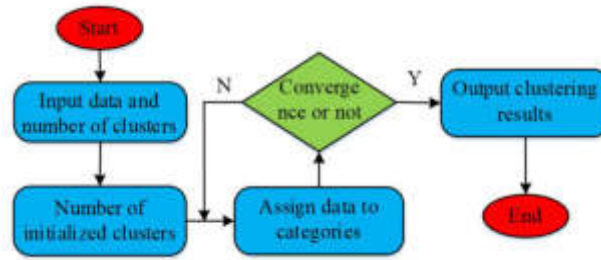


Fig. 3.7: Clustering method

model, as shown in equation (3.10).

$$P(U, c) = \sum_{k=1}^K \sum_{n=1}^N u_{nk} \sum_{m=1}^M w_m^\beta (x_{nm} - c_{km})^2 \quad (3.10)$$

In equation (3.9), w_m denotes the data feature weights with the number m and β denotes the custom parameter. Equation (3.10) can be transformed into equation (3.11) since the weights are subject to the naturalness condition of data sum equal to 1.

$$w_m = \frac{1}{\sum_{t \in F} [D_m / D_t]^{1/(\beta-1)}} \quad (3.11)$$

In equation (3.11), D_m denotes the sum of variances of all features within the clustered clusters. After designing the weighting mechanism, the study introduces the objective and subjective combining weighted k-means (Wosk-means) algorithm to assign values to each data feature, and the assignments are made in two ways: subjective weight assignment and objective weight assignment, and the integrated weights are shown in equation (3.12).

$$a_m = \frac{w_m v_m}{\sum_{i=1}^M w_m v_m} \quad (3.12)$$

In equation (3.12), w denotes the subjective weights, v denotes the objective weights, and m denotes the data feature numbers within the data clusters. The flow of the Wosk-means algorithm is shown in Fig 3.8.

In Fig 3.8, the model first standardizes the initial data, after which the data feature weights are initialized to make all data feature weights consistent. After processing the weights, the cluster centers need to be confirmed, and in the face of the given cluster centers and weights, the distance metric of the weights needs to be used to update the division of clusters. Based on this, the mean values of all features within the clusters are divided according to the existing weights and clusters, and the clustering centers are calculated and updated. Finally, the feature weights are updated according to the new clustering centers and clusters, and whether the algorithm converges or not is observed. At present, there are two main algorithms in the field of trajectory frequent pattern analysis, namely, Apriori algorithm and tree algorithm. The study uses the PrefixSpan algorithm which integrates the two algorithms for analysis, and the method can effectively reduce the cost of data mining. In the PrefixSpan algorithm, all sequences are arranged in an ordered manner, while all sequences are composed of item sets, which can be further split into different items. First, the database is input to the model and the minimum support minSup is defined. The length of the sequence pattern α is set to L and the projected database is $S|_\alpha$. A word scan is performed on $S|_\alpha$ and frequent items are found that satisfy the qualification.

On top of the trajectory pattern, the study transforms the trajectory data in a certain way and thus forms the distance between the trajectory features. The distance is expressed in the form of similarity. Similarity is essentially a comparison of the percentage of similar nodes with similar matches. Since there may be reading trajectory matches with different lengths between two trajectories, it is necessary to first find the frequent match

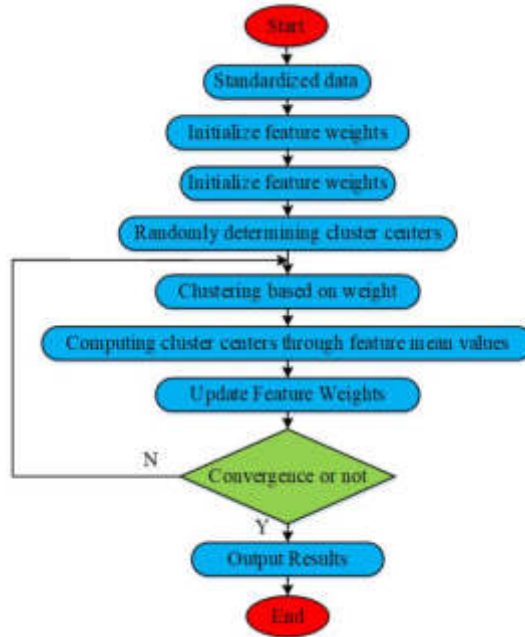


Fig. 3.8: Wosk-means algorithm flow

pattern, and then calculate the different match lengths to finally obtain the combined similarity of different length trajectory matches. It supposes that there exists a trajectory pattern $TraP_1$, as shown in equation (3.13).

$$TraP_1 = p'_{11} \xrightarrow{\Delta t'_{11}} p'_{12} \xrightarrow{\Delta t'_{12}} \dots \xrightarrow{\Delta t'_{1[u-1]}} p'_{1u} \tag{3.13}$$

In equation (3.14), t indicates the time and p indicates the location. A trajectory pattern $TraP_2$ is also presented.

$$TraP_2 = p'_{21} \xrightarrow{\Delta t'_{21}} p'_{22} \xrightarrow{\Delta t'_{22}} \dots \xrightarrow{\Delta t'_{2[u-1]}} p'_{2v} \tag{3.14}$$

Then the similarity of the two trajectories is shown in equation (3.15).

$$S(TraP_1, TraP_2) = \sum_{k=1}^K f_w(k) Sl(FT_1^k, FT_2^k) \tag{3.15}$$

In equation (3.15), k represents the trajectory matching length, $f_w()$ is the weight ratio representation, l is the trajectory matching pattern representation, and FT_1^k and FT_2^k represent the pattern matching subset, respectively.

4. College Student Trajectory Data Mining Model Trajectory Analysis Results.

4.1. Elbow Method Test . In the study of trajectory analysis of college students' trajectory data mining model, student information was first collected from various information segments within the university. The main information collection ends were five types of campus student information system, all-in-one card consumption record, network service record, access control record and action trajectory record. The dataset settings used in the experiment are shown in Table 4.1.

The experimental setup is shown in Table 4.2.

Table 4.1: Data type and data scale

Data source	Data properties	Data size
Campus student information system	Number of students	6714
	Location type and quantity	23
	Time interval	2021.12.01-2022.12.01
All-in-one card consumption record	Number of data records	185298
Network service record	Number of data records	187592
Access control record	Number of data records	167817
Action trajectory record	Maximum action trajectory length	181

Table 4.2: The experimental setup

Simulation settings	Detailed description
Experimental software	Use Python programming language for algorithm development and data processing.
	The applied clustering algorithm was implemented using the Scikit learn library.
	Use Jupyter Notebook for interactive calculations.
Experimental hardware	Desktop computer with Intel Core i7 processor and 16GB of memory.
	Using solid-state drives as storage devices.
Experimental condition	Experimental data collection period: December 1, 2021 to December 1, 2022.
	Use internal data sources within universities for analysis.
	Use elbow analysis and homogeneity analysis to evaluate performance.

Based on Table 4.1 and 4.2, the study first tested the performance of the applied clustering algorithm designed for the study, in which the elbow analysis method and the homogeneity analysis method were used to analyze the data, as shown in Fig 4.1.

In Fig 4.1, in the elbow method test, the overall error sum of squares of the applied clustering algorithm designed in the study showed a significant decreasing trend when the number of clusters rose. The decrease shrank significantly after the number of clusters was greater than 5, while the decrease almost disappeared when the number of clusters reached about 8, which shows that the number of clusters in the interval of 5 to 8 is the optimal setting range. The homogeneity test showed that the homogeneity of the algorithm was significantly improved at the number of clusters 5 and 8. In the comparison test, the clustering of data features in the traditional k-mean algorithm was more evenly distributed and did not provide effective information. In contrast, the applied clustering algorithm designed in the study first clustered the data into two large clusters of 0 and 1, where the feature distribution was still balanced. Then the algorithm further divided the two large clusters into five small clusters, where the feature differences between the clusters were already obvious.

4.2. Trajectory Feature Analysis. The results of trajectory features for different student types are shown in Table 4.3.

From Table 4.3, the model can classify trajectories in more detail for different disciplines and genders, from which the trajectory classification can reveal the characteristic patterns of action trajectories of different types of students in campus actions. The matching statistics of action trajectories between two two locations are shown in Fig 4.2.

In Fig 4.2, the action trajectory category from dormitory to dormitory received the highest number of matches within each cluster, with 98.92%, 98.98%, 96.15%, 98.37%, and 97.61% from category 1 to category 4, respectively. This was followed by the category from dormitory to canteen, with 87.64%, 87.86%, 86.97%, 88.63%, and 88.71% from category 1 to category 4, respectively. However, the category from dormitory to dormitory was somewhat invalid, so a trajectory similarity analysis was also needed, as shown in Fig 4.3.

Figure 4.3 illustrates the discrepancy in similarity between students' behavioral trajectories and the centroid trajectories of the cluster category to which they belong. A lower similarity indicates a greater divergence in students' behavioral trajectories from those of other students within the same cluster category. Moreover, a

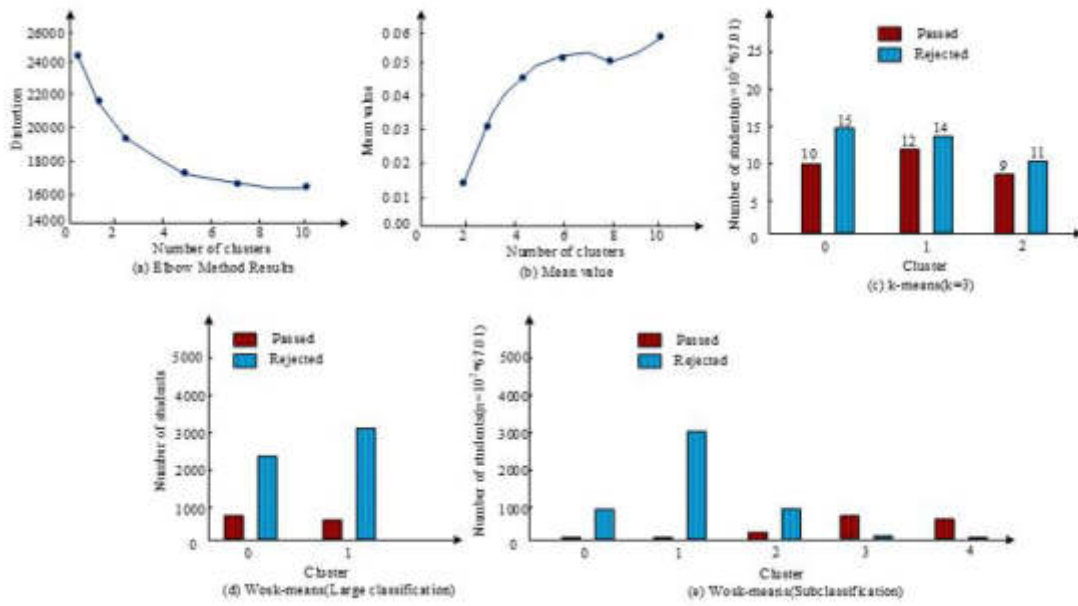


Fig. 4.1: Elbow analysis and homogeneity analysis

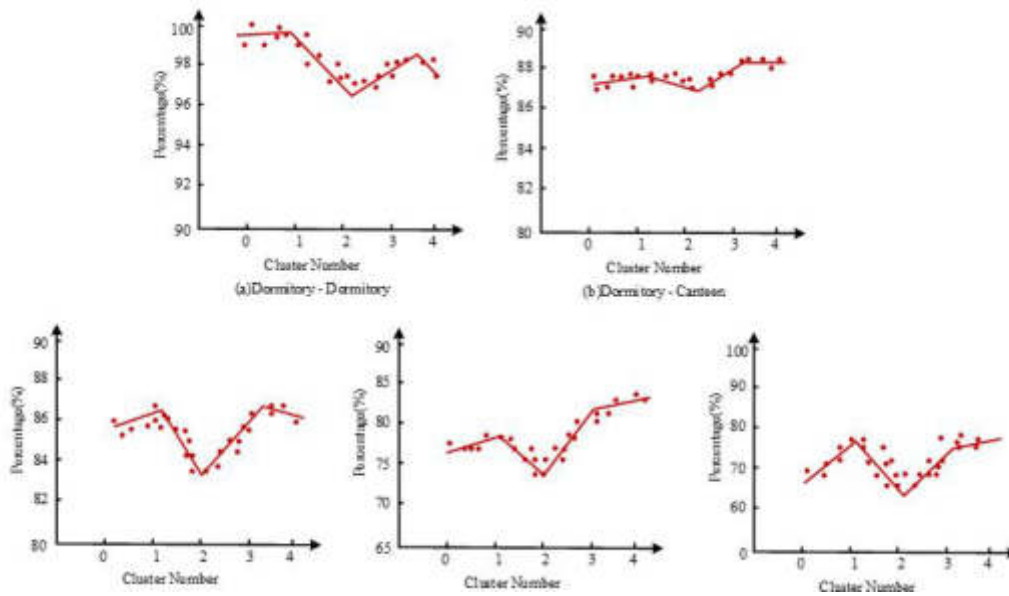


Fig. 4.2: Matching of action trajectories between two locations

higher percentage of this group of students indicates a greater number of students within that cluster category whose action patterns have changed. In this study, the similarity threshold was set at 10%, where the percentages of students with less than 10% similarity in categories 0, 1, 2, 3, and 4 were 1.18%, 1.17%, 0.66%, 0.42%, and 1.45%, respectively. The behavior patterns of this group of students were likely to change dramatically. The deviation directions of students whose trajectories deviated from the centroid cluster category are shown

Table 4.3: Trajectory features results for different student types

Cluster number	Subdivision	Campus card consumption data						Network usage		
		Breakfast shop	Print shop	Dining room	Bath	Playing field	School hospital	Rate of flow	Duration	Number of connections
0	The male sex	-34.61	39.71	21.06	-1.82	8.96	-42.65	0.76	3.76	6.71
	Femininity	-10.43	51.94	21.05	39.72	-12.52	8.19	3.06	14.77	13.96
	Science	-23.58	48.41	-5.14	16.22	9.77	-27.39	2.48	4.75	15.19
	Liberal arts	-7.55	35.72	10.69	31.87	-17.35	4.33	2.02	15.88	6.78
1	The male sex	-15.31	-49.11	-1.05	-22.07	28.86	-15.92	-7.45	-5.97	-3.77
	Femininity	6.13	-19.05	1.84	29.86	-0.58	28.83	-12.17	7.48	6.95
	Science	2.07	-5.48	-16.18	-12.75	43.48	-5.68	-6.38	-3.84	-2.13
	Liberal arts	-12.55	-49.05	-1.98	13.51	-38.44	8.66	-15.37	4.01	4.28
2	The male sex	2.02	-29.03	-9.02	-57.82	41.57	-11.33	-31.53	-45.43	-48.46
	Femininity	20.66	24.51	7.56	-29.95	-39.98	-16.28	-34.52	-42.65	-45.52
	Science	-1.46	-40.53	-13.46	-42.1	1.32	-28.37	-27.75	-43.92	-45.62
	Liberal arts	29.11	-12.27	2.21	-39.05	-20.44	-0.81	-38.94	-43.46	-47.63
3	The male sex	-10.68	1.92	-13.01	7.25	-5.15	-48.17	10.42	7.82	10.78
	Femininity	-21.81	43.91	14.05	37.65	-40.78	-37.78	4.32	7.75	9.38
	Science	-13.98	-23.05	-2.97	9.52	-29.08	-66.24	11.86	8.43	10.79
	Liberal arts	-17.99	16.41	8.63	40.08	-25.54	-33.18	-0.28	6.46	9.01
4	The male sex	40.43	15.36	3.97	-4.68	77.19	29.07	84.97	54.08	46.39
	Femininity	40.44	35.54	28.12	32.18	-20.62	46.96	47.12	39.27	31.67
	Science	56.66	-0.12	-13.48	10.31	39.43	10.11	63.61	34.31	28.31
	Liberal arts	19.72	71.58	2.33	34.77	-29.93	81.43	52.07	55.72	46.03

Table 4.4: Deviation direction

Percentage (%)	0	1	2	3	4
0	\	41.64	33.34	8.37	16.65
1	25.12	\	58.35	13.82	2.72
2	12.61	37.53	\	50.00	0.00
3	0.00	33.34	66.68	\	0.00
4	10.01	20.00	40.00	30.00	\

in Table 4.4.

In Table 4.4, the vertical direction represents the cluster in which the student is located, and the horizontal direction represents the cluster that the student is deviating towards. The bias of category 0 toward category 1 was higher, with a bias value of 25.12%. The bias of category 1 toward category 0 was higher, with a bias value of 41.64%. The bias of category 2 toward category 3 was higher, with a bias value of 66.68%. The bias of category 3 toward category 2 was higher, with a bias value of 50.00%. The bias of category 4 toward category 1 was higher, with a bias value of 16.65%. This showed that the model designed in the study can not only characterize the trajectory of students' campus actions, but also predict the possible changes of students' action patterns, providing new ideas for the management of students' campus life and academics.

In the benchmark comparison, the total sum of squared errors of the trajectory analysis clustering algorithm designed in the study was slightly lower than that of the benchmark K-means clustering algorithm. From the

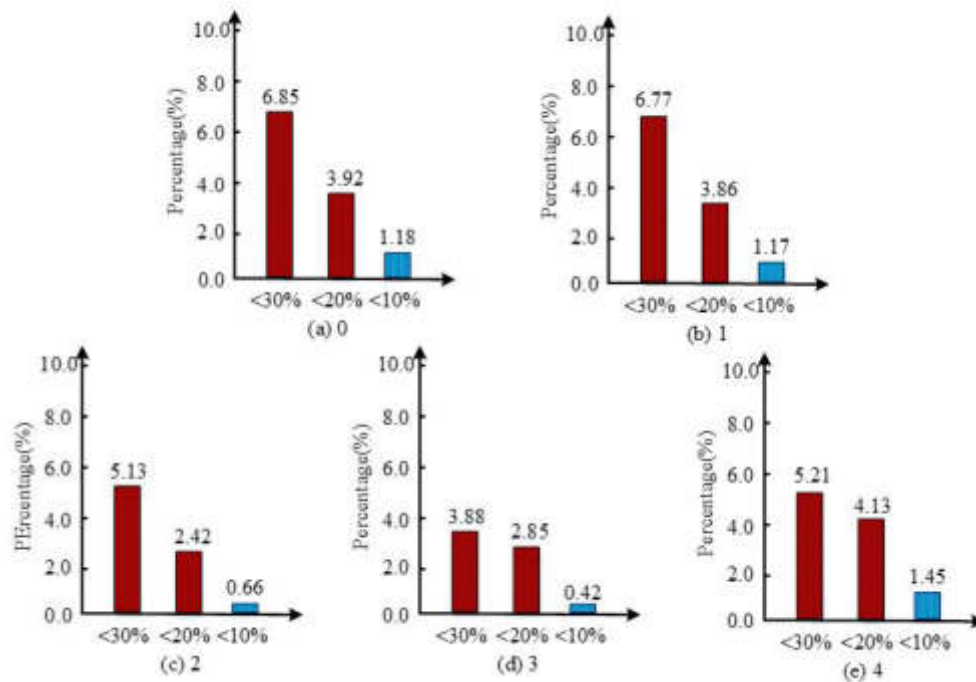


Fig. 4.3: Trajectory similarity analysis

Table 4.5: Benchmark comparison of different dataset sizes

Dataset size	Algorithm	Number of clusters	Total sum of squared errors	Cluster homogeneity
Sample size: 100 cases	Benchmark K-means	5	10251	0.83
		8	7562	0.88
	Applied clustering algorithm	5	8674	0.83
		8	7346	0.89
Sample size: 1000 cases	Benchmark K-means	5	102302	0.77
		8	75525	0.83
	Applied clustering algorithm	5	86007	0.82
		8	73220	0.89

perspective of clustering homogeneity, the total sum of squared errors of the trajectory analysis clustering algorithm was higher. Therefore, the trajectory analysis clustering algorithm designed in the study had more advantages in data feature extraction and analysis.

By increasing the size of the dataset by 10 times, the trajectory analysis clustering algorithm designed in the study still had advantages in the total sum of squared errors, indicating that the designed algorithm had stronger processing power when facing large-scale datasets. In the comparison of clustering homogeneity, the trajectory analysis clustering algorithm had higher clustering homogeneity, indicating that as the dataset size increased, the designed algorithm had more performance advantages.

With the increasing emphasis on talent cultivation in society, college students, as an important source of national talent reserves, are gradually receiving widespread attention from all sectors of society. The traditional talent cultivation system lacks adaptability to personalized and practical talent cultivation, making it difficult to meet the talent needs of today's society. Therefore, utilizing the current wave of social informatization and data development to update the talent cultivation system has become a new way of talent cultivation [26-27]. However, how to effectively utilize data mining technology to track and manage students' campus

life and learning, in order to achieve adaptive and efficient management, is an urgent problem to be solved [28]. In the research results, it can be seen that the model first collected student information from various sub sources within the university. After clustering analysis of these data, the applied clustering algorithms demonstrated significant performance in effectively distinguishing students' behavioral characteristics, while traditional K-means clustering algorithms could not provide this effective information. In addition, the applied clustering algorithm designed in the study had the optimal setting range within the range of 5 to 8 clusters. After analyzing the trajectory characteristics of different types of students, it can be found that the model could classify trajectories in more detail for students of different disciplines and genders. This indicated that through trajectory classification, characteristic patterns of different types of students' school behavior can be discovered. At the same time, it can be observed that the matching rate of behavior trajectories from dormitories to canteens was relatively high, but further trajectory similarity analysis is needed to confirm. In trajectory similarity analysis, there were significant differences between the behavior trajectories of most students and the centroid trajectories of their clustering categories. This difference may indicate a change in students' behavioral patterns.

By analyzing the similarity of trajectories, it was found that the similarity between students' action trajectories and the centroid trajectories of their cluster category was low, indicating that their behavior trajectories were different from those of other students within the cluster category. The more students in this situation, the greater the likelihood of changes in student behavior patterns within the cluster category. Finally, the study also found that students' action trajectories deviate from the direction of centroid clustering categories to a certain extent. For example, the deviation degree from category 0 to category 1 was relatively high, with a deviation value of 25.12%. The deviation degree from category 1 to category 0 was relatively high, with a deviation value of 41.64%. The deviation degree from category 2 to category 3 was relatively high, with a deviation value of 66.68%. The deviation degree from category 3 to category 2 was relatively high, with a deviation value of 50.00%. The deviation from category 4 to category 1 was relatively high, with a deviation value of 16.65%. These results indicated that the model designed in this study can not only describe students' campus action trajectories, but also predict possible changes in student behavior patterns, providing new ideas for students' campus life and academic management. Meanwhile, in the research, it is also possible to predict the possible changes in student behavior patterns, providing new ideas for students' campus life and academic management. Overall, through data mining technology, it is possible to gain a deeper understanding and analysis of the learning behavior and life trajectory of college students, thereby providing more effective support and methods for personalized education of students and talent cultivation in universities. So far, data mining technology has provided people with a novel and efficient method for student management and educational reform.

5. Conclusion. The research addressed the problem that the academic management of students in universities lacks daily campus life management, and proposed a model for mining students' daily academic life trajectories that combines the existing university information system and data mining technology. The model extracted and analyzed the information of students' daily activity, and designed an applied clustering algorithm to classify different trajectory types on this basis. The research results showed that in the elbow test and the homogeneity test, the applied clustering algorithm had obvious variation characteristics at the cluster number 5-8, and this interval was the best cluster number interval. The applied clustering algorithm in the comparison test was better at clustering the data into clusters with features than the ordinary clustering algorithm. In the trajectory analysis, the model could classify trajectories in more detail for different disciplines and genders, in which the categories of action trajectories from dormitory to canteen were 87.64%, 87.86%, 86.97%, 88.63%, 88.71% from category 0 to 4 respectively, which were the most effective action trajectories with the highest number of matches. The percentage of people with similarity less than 10% in categories 0 to 4 were 1.18%, 1.17%, 0.66%, 0.42%, and 1.45%. The behavior patterns of this group of students were likely to change dramatically. In addition, the model was able to predict the direction of trajectory change, with the main directions being category 0 to category 1, category 1 to category 0, category 2 to category 3, category 3 to category 2, and category 4 to category 1. This showed that the model designed in the study can effectively analyze student estimation and provide assistance for comprehensive academic life management in universities.

However, the drawback is that this study mainly relied on data from university information systems, but students' daily life trajectories may be influenced by more elements, such as social media activities, health,

and psychological conditions. At the same time, this study mainly focused on the analysis of student behavior trajectories, but did not involve how to effectively intervene based on these analysis results. Therefore, in future research, it is possible to explore how to use these analysis results to design and implement effective student life management strategies. At the same time, future research can consider integrating more types of data sources to provide a more comprehensive trajectory of student life.

Fundings. The research is supported by: 2023 Guangxi University Young and Middle-aged Teachers Scientific Research Basic Ability Improvement Project. The Exploration and Practice of Rural Modernization in the Ethnic Areas of Northern Guangxi Promoted by the Digital intelligence, (No.2023KY1725).

REFERENCES

- [1] P. Radanliev, D. De Roure, and R. Walton, "Diabetes & Metabolic Syndrome: Clinical Research & Reviews," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 14, no. 5, pp. 1121-1132, 2020.
- [2] A. Ishaq, S. Sadiq, M. Umer, S. Ullah, S. Mirjalili, V. Rupapara, and M. Nappi, "Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques," *IEEE Access*, vol. 9, pp. 39707-39716, 2021.
- [3] H. Lou, "Design of college English process evaluation system based on data mining technology and internet of things," *International Journal of Data Warehousing and Mining (IJDWM)*, vol. 16, no. 2, pp. 18-33, 2020.
- [4] Y. Zeng, "Evaluation of physical education teaching quality in colleges based on the hybrid technology of data mining and hidden Markov model," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 15, no. 1, pp. 4-15, 2020.
- [5] K. Yu, T. Yuan, and Y. Li, "Application of Data Mining Technology in Sports Data Analysis in Colleges and Universities," in *2021 International Conference on Information Technology and Contemporary Sports (TCS)*, IEEE, 2021, pp. 329-332.
- [6] I. S. P. James, P. Ramasubramanian, and D. M. D. Angeline, "Student absenteeism in engineering college using rough set and data mining approach," *International Journal of Advanced Intelligence Paradigms*, vol. 23, no. 3-4, pp. 423-433, 2022.
- [7] Z. Zhang, Z. Chen, and C. Xu, "Recognition method of diversified teaching mode of college physical training based on data mining," *Ann. For. Res.*, vol. 65, no. 1, pp. 9420-9432, 2022.
- [8] L. D. Yulianto, A. Triayudi, and I. D. Sholihati, "Implementation Educational Data Mining for Analysis of Student Performance Prediction with Comparison of K-Nearest Neighbor Data Mining Method and Decision Tree C4.5: Implementation Educational Data Mining for Analysis of Student Performance Prediction with Comparison of K-Nearest Neighbor Data Mining Method and Decision Tree C4.5," *Jurnal Mantik*, vol. 4, no. 1, pp. 441-451, 2020.
- [9] H. Zhang and M. Fang, "Research on the integration of heterogeneous information resources in university management informatization based on data mining algorithms," *Computational Intelligence*, vol. 37, no. 3, pp. 1254-1267, 2021.
- [10] S. M. Lee and J. H. Lee, "The influence of adaptation of dental hygiene students to campus life on satisfaction in major," *Journal of Korean Society of Dental Hygiene*, vol. 21, no. 3, pp. 281-290, 2021.
- [11] W. Li, "Design of smart campus management system based on internet of things technology," *Journal of Intelligent & Fuzzy Systems*, vol. 40, no. 2, pp. 3159-3168, 2021.
- [12] Y. Kim, B. Kim, H. S. Hwang, and D. Lee, "Social media and life satisfaction among college students: A moderated mediation model of SNS communication network heterogeneity and social self-efficacy on satisfaction with campus life," *The Social Science Journal*, vol. 57, no. 1, pp. 85-100, 2020.
- [13] S. Purnama, Q. Aini, U. Rahardja, N. P. L. Santoso, and S. Millah, "Design of Educational Learning Management Cloud Process with Blockchain 4.0 based E-Portfolio," *Journal of Education Technology*, vol. 5, no. 4, pp. 628-635, 2021.
- [14] K. A. Way, L. Burrell, L. D'Allura, and K. Ashford-Rowe, "Empirical investigation of authentic assessment theory: An application in online courses using mimetic simulation created in university learning," *Assessment & Evaluation in Higher Education*, vol. 46, no. 1, pp. 17-35, 2021.
- [15] W. Haoxiang and S. Smys, "Big data analysis and perturbation using data mining algorithm," *Journal of Soft Computing Paradigm (JSCP)*, vol. 3, no. 01, pp. 19-28, 2021.
- [16] T. S. Kumar, "Data mining based marketing decision support system using hybrid machine learning algorithm," *Journal of Artificial Intelligence*, vol. 2, no. 03, pp. 185-193, 2020.
- [17] P. Edastama, A. Dudhat, and G. Maulani, "Use of Data Warehouse and Data Mining for Academic Data: a Case Study at a National University," *International Journal of Cyber and IT Service Management*, vol. 1, no. 2, pp. 206-215, 2021.
- [18] Z. S. Ageed, S. R. M. Zeebaree, M. M. Sadeeq, F. F. Kak, H. S. Yahia, M. R. Mahmood, and I. M. Ibrahim, "Comprehensive survey of big data mining approaches in cloud systems," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 29-38, 2021.
- [19] H. A. Mengash, "Using data mining techniques to predict student performance to support decision making in university admission systems," *IEEE Access*, vol. 8, pp. 55462-55470, 2020.
- [20] A. Garg, P. Kumar, M. Madhukar, O. Loyola-González, and M. Kumar, "Blockchain-based online education content ranking," *Education and Information Technologies*, vol. 27, no. 4, pp. 4793-4815, 2022.
- [21] S. Dutt, N. J. Ahuja, and M. Kumar, "An intelligent tutoring system architecture based on fuzzy neural network (FNN) for special education of learning disabled learners," *Education and Information Technologies*, vol. 27, no. 2, pp. 2613-2633, 2022.
- [22] C. Choudhary, I. Singh, and M. Kumar, "SARWAS: Deep ensemble learning techniques for sentiment based recommendation system," *Expert Systems with Applications*, vol. 216, pp. 119420, 2023.

- [23] K. Y. Siu, "The Effect of Working Memory on Bilingual Learning Ability," *Journal of Education, Humanities and Social Sciences*, vol. 8, pp. 2118-2123, 2023.
- [24] R. Darmayanti, "Gema Cow-Pu: Development of Mathematical Crossword Puzzle Learning Media on Geometry Material on Middle School Students' Critical Thinking Ability," *Assyfa Learning Journal*, vol. 1, no. 1, pp. 37-48, 2023.
- [25] Y. Du, "Study on Cultivating College Students' English Autonomous Learning Ability under the Flipped Classroom Model," *English Language Teaching*, vol. 13, no. 6, pp. 13-19, 2020.
- [26] G. HR and P. S. Aithal, "Sales Personnel Training—An Integrated Framework for Indian Brick-and-Mortar Retailers," *International Journal of Case Studies in Business, IT, and Education (IJCSBE)*, vol. 4, no. 1, pp. 172-187, 2020.
- [27] E. F. Zeer, V. S. Tretyakova, and V. I. Miroshnichenko, "Strategic directions of pedagogical personnel training for the system of continuing vocational education," *The Education and Science Journal*, vol. 21, no. 6, pp. 93-121, 2019.
- [28] H. Zhang, X. He, and H. Mitri, "Fuzzy comprehensive evaluation of virtual reality mine safety training system," *Safety Science*, vol. 120, pp. 341-351, 2019.

Edited by: Achyut Shankar

Special issue on: Machine Learning for Smart Systems: Smart Building, Smart Campus, and Smart City

Received: Jun 16, 2024

Accepted: Jun 13, 2024



MULTI CHANNEL ELECTRONIC COMMUNICATION SIGNAL PARAMETERS BASED ON NONLINEAR PHASE PRINCIPLE MODULATION AND DEEP LEARNING

XIAOQING YAN*

Abstract. In order to solve the problem of high sampling rate and large number of sampling points required by current phase modulation signal parameter estimation methods, a parameter modulation method for multi-channel electronic communication signals based on nonlinear phase principle and deep learning is proposed. Firstly, classify and introduce the modulation methods, and propose a new algorithm for identifying instantaneous feature parameters. The author conducted nonlinear phase principle modulation recognition on seven typical digital signals: 2ASK, 4ASK, 2FSK, 4FSK, 2PSK, 4PSK, and 16QAM. Using the author's algorithm, experiments were conducted on the recognition of seven digital nonlinear phase modulation signals under different signal-to-noise ratios. As can be seen from the results, when the signal-to-noise ratio is greater than or equal to 10dB, the recognition accuracy of the seven digital nonlinear phase modulation signals can reach 100%, verifying that the new algorithm proposed by the author improves the recognition accuracy.

Key words: Nonlinearity, Phase principle modulation, Communication signal, characteristic parameter

1. Introduction. Automatic modulation recognition technology is a very important topic in the field of non cooperative communication signal processing research. The task of modulation recognition for communication signals is to identify signals without sufficient or complete prior knowledge, by performing various processing on the received signal, the modulation method and related modulation parameters used in the signal can be accurately determined[1]. For the signal receiving end, determining the modulation method of the received signal and correctly demodulating the signal is a necessary prerequisite for restoring the original signal. The study of automatic modulation recognition technology for signals has significant practical value in both military and civilian fields. The practical value of modulation recognition technology is mainly reflected in: In the military field, successfully determining the modulation mode of the signal is a prerequisite for achieving reconnaissance and interference of enemy communication. Knowing the modulation method of enemy signals can estimate some useful parameters, in order to conduct targeted reconnaissance and electronic interference on enemy communication; In the civilian field, the task of radio management work in the communication management department is to monitor whether legitimate radio stations comply with the working parameters assigned by the management department during the communication process, while listening for interference from illegal radio stations to ensure the normal communication of legitimate radio stations. The most crucial technology to achieve these non cooperative communication tasks is modulation recognition technology. There are two methods for modulation recognition of wireless communication signals: One is manual judgment, and the other is machine automatic recognition. Early modulation recognition methods used a set of demodulators with different modulation methods, the received signal is downconverted and input into each demodulator to obtain an observable signal, which is then judged by the operator based on information such as time-domain waveform, signal spectrum, instantaneous amplitude, instantaneous frequency, and instantaneous phase [2,3]. The recognition method of manual judgment requires experienced operators. Due to the subjective factors involved in the judgment process, the judgment results will vary from person to person, and the modulation types that can be recognized by manual judgment will be very limited. And automatic modulation recognition technology can solve the above problems.

The ultimate goal of automatic modulation recognition technology is to develop a machine that can recognize as many modulation modes as possible without any prior knowledge and low signal-to-noise ratio. We hope that the less prior knowledge there is in modulation recognition, the better, or the more "blind" the

*College of Education, Jiangxi University of Engineering, Xinyu Jiangxi, 338000, China (XiaoqingYan78@126.com)

modulation recognition algorithm is [4,5]. However, in the actual research process of modulation recognition technology, researchers will more or less add some prior knowledge, such as only studying digital modulation recognition, which means that they already know that the received signal is a digital signal, not an analog signal.

2. Literature Review. Radar signal recognition is an important aspect of electronic reconnaissance, which refers to the process of matching the features of the received signal emitted by the radar signal source with the pre accumulated signal features to confirm the signal modulation method. Radar signal recognition usually includes: Intentional modulation recognition and unintentional modulation recognition of radar signals, target recognition of radar signal source platforms, and estimation of recognition credibility [6]. At present, Western countries led by the United States are in a leading position in radar signal recognition technology, but due to their military confidentiality, they have limited access to information. As far as we know, the main algorithms for radar signal recognition include time-frequency analysis, spectral correlation, time-domain autocorrelation, wavelet transform, digital intermediate frequency, and time-domain cepstrum. The time-domain cepstrum method extracts modulation features and related modulation parameters by calculating the cepstrum of the signal. This method requires various transformations, requires a large amount of computation, is difficult to implement in hardware, and has low accuracy, so its practical application value is not significant. The digital intermediate frequency method can comprehensively recognize radar signals, with the increasing processing speed of DSP chips, it is a promising technology, however, the relevant technology is not yet very mature and requires a lot of research. The advantage of spectral correlation method is that it has good resolution, but the actual environment is complex and the received signal length is limited, resulting in low recognition accuracy. The time-frequency analysis method and wavelet transform method are newly developed and highly effective tools for processing non-stationary signals in recent years [7]. The time-frequency analysis method is a two-dimensional joint analysis of the time-domain and frequency-domain characteristics of a signal, the real-time frequency analysis method can simultaneously describe the energy density of a signal at different times and frequencies, and can effectively describe the local characteristics of the signal, in recent years, it has received increasing attention. The wavelet transform method is also a time-frequency analysis method, which has the characteristics of multi resolution analysis and can characterize the local characteristics of the signal. The signal has high frequency resolution and low time resolution in the low frequency range, while it has low frequency resolution and high time resolution in the high frequency range, therefore, applying wavelet transform to the signal can obtain different details. And different radar signals have different detailed features, which can be used to identify radar signals [8]. Researchers have been striving to find fast and efficient automatic recognition technologies, and have achieved considerable success. However, the research on automatic modulation recognition technology has not yet matured and finalized, due to: One reason is that new modulation methods are constantly emerging, and the modulation types of communication signals are becoming more diverse, while previous modulation recognition algorithms only worked on specific types of modulation signals. Secondly, the complexity of wireless communication environments poses challenges to non cooperative communication. Compared to wired communication, wireless communication has its own characteristics: Firstly, the wireless channel of wireless communication is open and susceptible to interference from other signals and various noises; Second, radio propagation has a variety of ways, including diffraction, reflection and refraction. The signal received by the receiver will cause signal fading due to multi-path effects; Thirdly, there is also the Doppler effect in mobile communication, which can cause signal items to change at times. The multipath and Doppler effects seriously affect the reception quality of signals. In the process of non cooperative communication, the receiver cannot obtain the signal parameters of the sender like in cooperative communication. The diversified wireless communication technology requires non cooperative communication receiving systems to have characteristics such as wide coverage, strong adaptability, and anti fading. Thirdly, the signal environment is becoming increasingly dense, and at the same time, multiple signals with different modulation methods will enter the receiver. This puts forward new requirements for signal modulation recognition, that is, how to achieve recognition of multiple modulation signals at the same time. These situations all determine that there are many new research works to be carried out in the field of automatic modulation and recognition of communication signals [9].

This article briefly introduces a digital nonlinear phase modulation recognition algorithm proposed by E.E. Azzouz and A.K. Nandi to address these issues, because the features extracted by the nonlinear phase modula-

tion recognition algorithm based on instantaneous features are all derived from the operation of instantaneous amplitude, instantaneous phase, and instantaneous frequency, the algorithm proposed by the author is used to identify, simulate, and analyze seven types of digital nonlinear phase modulation signals, and the decision process and selected decision threshold are provided.

3. Methods.

3.1. Classification of modulation methods .

From the perspective of modulation recognition, communication signals can be classified using various methods. The first classification is based on the information content contained in the signal, and any communication signal can be classified into one of the following four categories:

1. If a signal only contains amplitude information but not phase information, it is called an amplitude signal [10,11]. The so-called amplitude information here refers to the instantaneous amplitude of the signal not being constant; Phase information refers to the instantaneous phase of a signal that is not constant. Correspondingly, without amplitude information, the instantaneous amplitude of the signal is constant; No phase information refers to the instantaneous phase of a signal being constant. Amplitude signals such as MASK (M-scale amplitude keying) signals.
2. If a signal only contains phase information but not amplitude information, it is called a phase signal. For example, MFSK (M-ary Frequency Shift Keying) signal and MPSK (M-ary Phase Shift Keying) signal.
3. If a signal has both amplitude and phase information, it is called a composite signal. For example, MQAM (M-ary Orthogonal Amplitude Modulation) signal [12].
4. If a signal has neither amplitude nor phase information, it is called a carrier wave (CW) signal. Such as sine and cosine signals.

The second classification is based on the symmetry of the signal spectrum with respect to the carrier frequency. Usually, the spectrum of a signal consists of one carrier component and two sideband components, but in some modulation methods, the carrier component and two sideband components may not be all preserved. According to the presence of sidebands, communication signals can be divided into two categories: symmetric signals and asymmetric signals.

The third classification is divided into analog modulation signals and digital modulation signals based on the properties of modulation signals.

The fourth classification is divided into two categories based on the types of carriers: sine wave modulation and pulse modulation [13].

This project studies the sine wave modulation methods of digital signals, including the following modulation methods: 2ASK (binary amplitude keying), 4ASK (quaternary amplitude keying), 2FSK (binary frequency shift keying), 4FSK (quaternary frequency shift keying), 2PSK (binary phase shift keying), 4PSK (quaternary phase shift keying), and 16QAM (hexadecimal orthogonal amplitude modulation). Other modulation methods are not discussed here.

3.2. Recognition algorithm based on new instantaneous feature parameters . Parameter extraction and threshold selection: The author conducted nonlinear phase principle modulation recognition on 7 typical digital signals, including 2ASK, 4ASK, 2FSK, 4FSK, 2PSK, 4PSK, and 16QAM. After comprehensive consideration of various aspects, the following 5 instantaneous feature parameters were extracted for signal classification.

(1) *The mean M_a^2 of the normalized instantaneous amplitude square at zero center.* The mean M_a^2 of the normalized instantaneous amplitude square at zero center is obtained by the following equation:

$$M_a^2 = \frac{1}{N} \sum_{i=1}^{N_s} |a_{cn}(i)|^2 \quad (3.1)$$

In Equation 3.1, N_s is the total number of sampling points; $a_{cn}(i)$ is the zero center normalized instantaneous amplitude, and $a_{cn}(i)$ is calculated from Equation 3.2:

$$a_{cn}(i) = a_n(i) - 1 \quad (3.2)$$

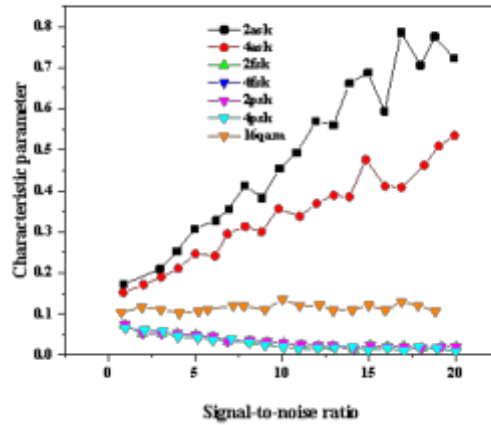


Fig. 3.1: The variation of parameter M_a^2 of modulated signals with different phase principles with signal-to-noise ratio

In Equation 3.2, the normalized instantaneous amplitude $a_n(i) = \frac{a(i)}{m_s}$, while $m_s = \frac{1}{N_s} \sum_{i=1}^{N_s} a(i)$ is the average of the instantaneous amplitude $a(i)$, and the characteristic parameter M_a^2 . Seven types of digital nonlinear phase modulation signals can be divided into three categories: MASK signals are classified into one category, 16QAM signals are classified into one category, and MFSK and MPSK signals are classified into another category. The instantaneous amplitude of MASK and 16QAM signals varies [14]; The instantaneous amplitude of the MPSK signal only undergoes a sudden change in amplitude at the moment of phase change, so its characteristic parameters are relatively small; The instantaneous amplitude of the MFSK signal is constant, the envelope is constant, and its characteristic parameter is zero. The actual simulation results are shown in Figure 1. From the figure, we can observe that at low signal-to-noise ratios, the characteristic parameters of signals modulated by different nonlinear phase principles are not significantly different due to the influence of noise[15]. However, as the signal-to-noise ratio increases, the characteristic parameters of signals modulated by different nonlinear phase principles begin to approach the theoretical calculated values, therefore, by selecting appropriate thresholds, MASK, 16QAM, and MFSK, MPSK signals can be separated. Based on multiple simulation attempts and weighing the impact on global decisions, the threshold $t1(M_a^2)$ of the mean M_a^2 of the normalized instantaneous amplitude squared at the zero center was selected as 0.12, and the threshold $t2(M_a^2)$ was selected as 0.08. When the threshold is $t2(M_a^2) < t(M_a^2) < t1(M_a^2)$, it is determined as a 16QAM signal; When the threshold is $t(M_a^2) > t1(M_a^2)$, it is judged as a MASK signal; When the threshold is $t(M_a^2) < t2(M_a^2)$, it is determined as an MFSK signal or an MPSK signal [16,17].

(2) Recursive Zero Center Normalized Instantaneous Amplitude Square Mean RM_a^2 .

$$RM_a^2 = \frac{1}{N} \sum_{i=1}^{N_s} |ra_{cn}(i)|^2 \tag{3.3}$$

In Equation 3.3, N_s is the total number of sampling points, $ra_{cn}(i)$ is the recursive zero center normalization instantaneous amplitude, that is, after normalizing the zero center, the instantaneous amplitude $a_{cn}(i)$ is calculated, and then the zero center normalization instantaneous amplitude $ra_{cn}(i)$ is calculated by the following equation:

$$ra_{cn}(i) = \frac{a_{cn}(i)}{\frac{1}{N} \sum_{i=1}^{N_s} a_{cn}(i)} - 1 \tag{3.4}$$

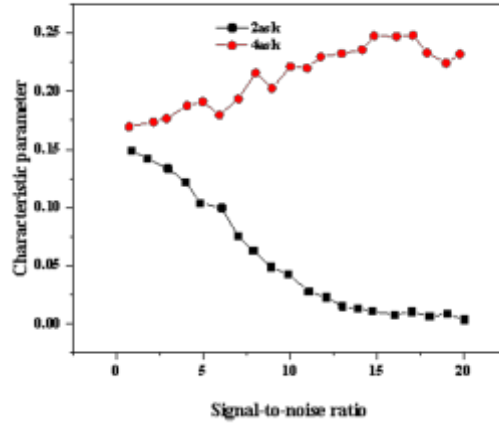


Fig. 3.2: Changes in parameter AA of modulated signals with different phase principles as a function of signal-to-noise ratio

The feature parameter RM_a^2 is used to distinguish between 2ASK signals and 4ASK signals. According to the time-domain characteristics of these two types of signals, their instantaneous amplitudes are 2 and 4, respectively, indicating that the RM_a^2 corresponding to the 4ASK signal is greater than the RM_a^2 corresponding to the 2ASK signal. Therefore, by setting an appropriate threshold value $t(RM_a^2)$, 2ASK and 4ASK signals can be identified. Based on multiple simulation attempts and weighing the impact on global decisions, finally, the threshold $t(RM_a^2)$ of the mean RM_a^2 of the normalized instantaneous amplitude square of the zero center is selected as 0.17. The variation of the mean RM_a^2 of the recursive zero center normalized instantaneous amplitude square of modulated signals with different digital nonlinear phase principles with signal-to-noise ratio is shown in Figure 3.2.

(3) The mean M_f^2 of the square of the normalized instantaneous frequency at zero center.

$$M_f^2 = \frac{1}{N_s} \sum_{i=1}^{N_s} |f_{cn}(i)|^2 \quad (3.5)$$

In Equation 3.5, N_s is the total number of sampling points: $f_{cn}(i)$ is the zero center normalized instantaneous frequency. According to the time-domain characteristics of the signal, the MFSK signal has at least 2 instantaneous frequency values, while the MPSK signal only has 1 instantaneous frequency value, meaning that the M_f^2 corresponding to the MFSK signal is greater than the M_f^2 corresponding to the MPSK signal. Therefore, this feature parameter can be used to distinguish between MFSK signals and MPSK signals. Based on multiple simulation attempts and weighing the impact on global decisions, finally, the threshold $t(M_f^2)$ of the mean M_f^2 of the zero center normalized instantaneous frequency squared is selected as 0.075. The variation of the mean M_f^2 of the zero center normalized instantaneous frequency square of modulated signals with different digital nonlinear phase principles with signal-to-noise ratio is shown in Figure 3.3 [18].

(4) Recursive Zero Center Normalized Instantaneous Frequency Square Mean. RM_f^2

$$RM_f^2 = \frac{1}{N_s} |rf_{cn}(i)|^2 \quad (3.6)$$

In Equation 3.6, N_s is the number of sampling points, and $rf_{cn}(i)$ is the recursive zero center normalized instantaneous frequency, namely, normalize the instantaneous frequency of the zero center and then calculate

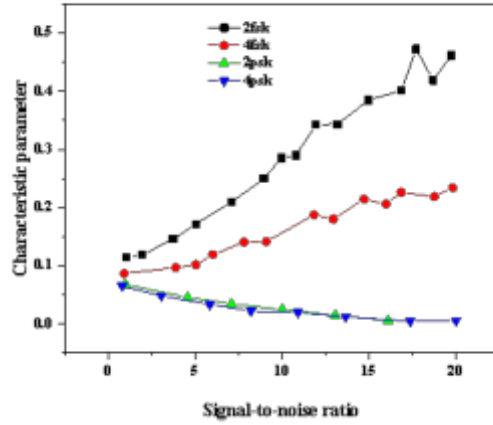


Fig. 3.3: The variation of parameter M_f^2 of modulated signals with different phase principles with signal-to-noise ratio

the normalized instantaneous frequency $f_{cn}(i)$ of the zero center using the following formula:

$$f_{cn}(i) = \frac{f_{cn}(i)}{\frac{1}{N_s} \sum_{i=1}^{N_s} f_{cn}(i)} - 1 \tag{3.7}$$

In Equation 3.7, $f_{cn}(i)$ is the zero center normalized instantaneous frequency. According to the time-domain characteristics, the number of instantaneous frequency values of the 2FSK signal is 2, which is significantly smaller than the number of instantaneous frequency values of the 4FSK signal, therefore, the RM_f^2 value corresponding to 2FSK is smaller than the RM_f^2 value of 4FSK, so this feature parameter RM_f^2 can distinguish between 2FSK and 4FSK signals. Based on multiple simulation attempts and weighing the impact on global decisions, the threshold $t(RM_f^2)$ of the mean RM_f^2 of the normalized instantaneous frequency square of the recursive zero center was ultimately selected as 0.225 [19]. The variation of the mean RM_f^2 of the recursive zero center normalized instantaneous frequency square of modulated signals with different digital nonlinear phase principles with signal-to-noise ratio is shown in Figure 3.4.

(5) Mean M_p^2 of normalized instantaneous phase squared at zero center.

$$M_p^2 = \frac{1}{N} \sum_{i=1}^{N_s} |p_{cn}(i)|^2 \tag{3.8}$$

In Equation 3.8, N_s is the number of sampling points, $p_{cn}(i)$ is the zero center normalized instantaneous phase, calculated by the following equation:

$$p_{cn}(i) = p_n(i) - 1 \tag{3.9}$$

In Equation 3.9, $p_n(i) = \frac{p(i)}{m_s}$, while $m_s = \frac{1}{N_s} \sum_{i=1}^{N_s} p(i)$ is the average of the instantaneous phase $p(i)$. The instantaneous phase number of 4PSK is greater than that of 2PSK, and the characteristic parameter M_p^2 can distinguish between 4PSK and 2PSK signals. Based on multiple simulation attempts and weighing the impact on global decisions, the threshold $t(M_p^2)$ of the mean M_p^2 of the normalized instantaneous amplitude squared at the zero center was ultimately selected as 0.2. The variation of the mean M_p^2 of the zero center normalized instantaneous phase square of modulated signals with different digital nonlinear phase principles with signal-to-noise ratio is shown in Figure 3.5.

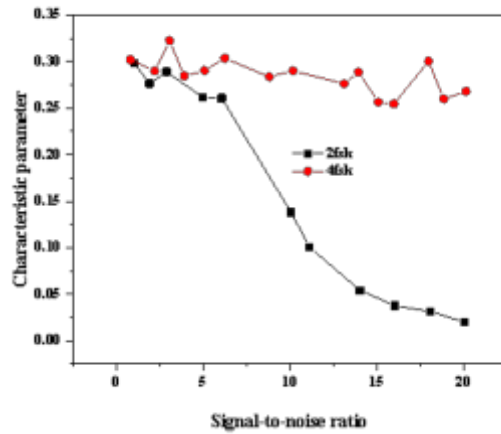


Fig. 3.4: The variation of parameter RM_f^2 of modulated signals with different phase principles with signal-to-noise ratio

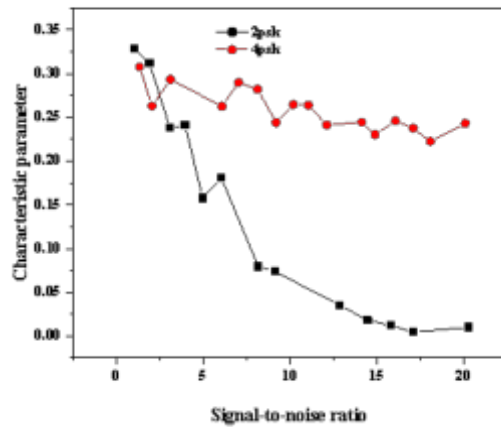


Fig. 3.5: The variation of parameter M_p^2 of modulated signals with different phase principles as a function of signal-to-noise ratio

4. Results and Analysis. Figure 4.1 is the non-linear phase principle modulation recognition flowchart of the algorithm in this paper. In the recognition algorithm proposed by the author, only 5 feature parameters can identify 7 types of digital nonlinear phase principle modulation signals. However, the five features proposed by scholars E.E. Azzouz and A.K. Nandi can only recognize six types of digital nonlinear phase modulation signals [20,21].

Firstly, in order to ensure that when the signal sender uses symbol 0 to modulate the MASK signal using the nonlinear phase principle, the MASK can be recognized, we can only use the feature M_a^2 related to instantaneous amplitude to distinguish MASK signals from other signals, and RM_a^2 to distinguish 2ASK signals from 4ASK signals.

Secondly, from the instantaneous characteristic maps of MFSK and MPSK, it can be seen that the instantaneous frequency of MFSK has only a finite number of discrete values, while the instantaneous frequency of

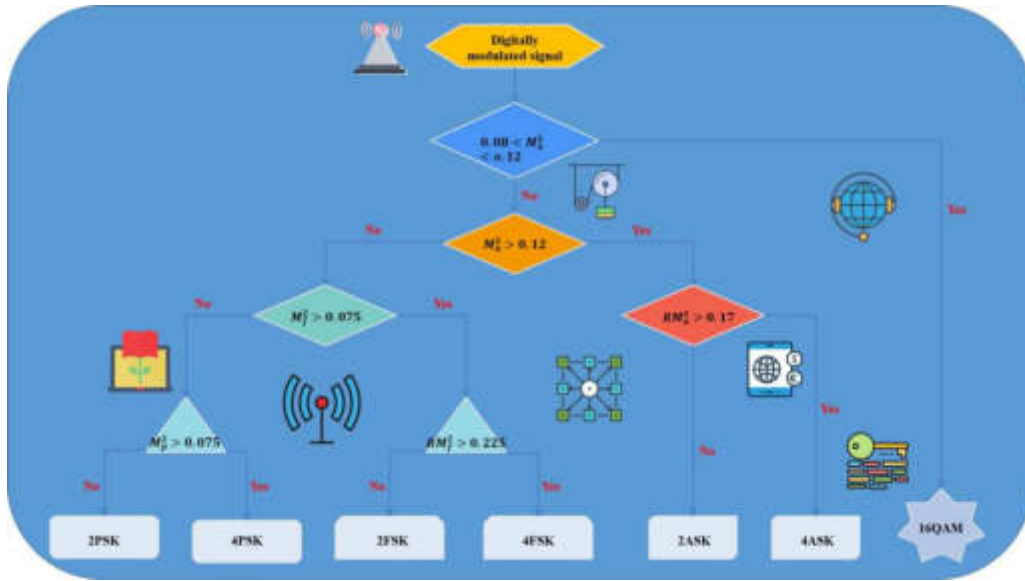


Fig. 4.1: Nonlinear Phase Principle Modulation Recognition Flowchart of the Author's Algorithm

MPSK is constant, the use of feature M_f^2 can effectively distinguish between MFSK signals and MPSK signals. It is not appropriate to use feature M_p^2 to distinguish between MFSK signals and MPSK signals, because the instantaneous phase of the MFSK signal is not constant, but time-varying.

Finally, feature RM_f^2 is used to distinguish between 2FSK and 4FSK signals, and feature M_p^2 is used to distinguish between 2PSK and 4PSK signals. At this point, all seven types of digital nonlinear phase modulation signals have been distinguished [22].

Figure 4.2 shows the recognition results of seven digital nonlinear phase modulation signals using the author's algorithm under different signal-to-noise ratios. As can be seen from Figure 4.2, when the signal-to-noise ratio is greater than or equal to 10dB, the recognition accuracy of all seven digital nonlinear phase modulation signals can reach 100%.

E. Azzouz and A.K. Nandi, two scholars, did not provide a simulation diagram similar to Figure 4.2 showing the variation of digital nonlinear phase principle modulation signal recognition results with signal-to-noise ratio. Instead, they only provided the recognition accuracy under three conditions of signal-to-noise ratio: 10dB, 15dB, and 20dB[23]. Table 1 is a comparison table of the correct recognition rates of the author's algorithm and classical algorithm under three different signal-to-noise ratios of 10dB, 15dB, and 20dB, respectively. By comparison, it can be seen that, compared with the classic algorithms of E.E. Azzouz and A.K. Nandi, the new algorithm proposed by the author achieves better recognition results at low signal-to-noise ratios by adding a 16QAM nonlinear phase principle modulation method .

5. Conclusion. The features extracted by the nonlinear phase modulation recognition algorithm based on instantaneous information are all derived from the operation of instantaneous amplitude, instantaneous phase, and instantaneous frequency, the author analyzed how to extract these three instantaneous feature parameters. Simulation and analysis were conducted on the recognition of seven types of digital nonlinear phase modulation signals using the feature parameters proposed by the author, the decision process and selected decision threshold were provided, and the results showed that the author's algorithm improved the recognition success rate. In recent years, the research methods and directions of automatic nonlinear phase modulation recognition algorithms have been continuously expanded, and progress has been made to some extent, however, there are still many key issues that have not been well resolved. The author's research on nonlinear phase principle modulation recognition algorithms still has many shortcomings. All research on nonlinear phase principle modulation recognition focuses on certain types of modulation signals, the author only studied seven

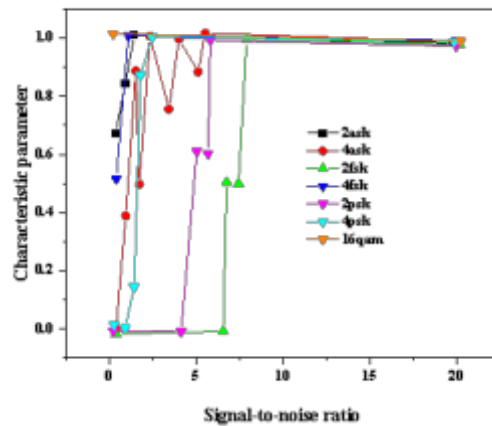


Fig. 4.2: Recognition results of seven types of digital nonlinear phase modulation signals under different signal-to-noise ratios

Table 4.1: Comparison of the correct recognition rates of the author's algorithm and classical algorithm under different signal-to-noise ratios

modulation	Author's Algorithm			Assical algorithm		
	SNR=10	SNR=15	SNR=20	SNR=10	SNR=15	SNR=20
2ASK	100%	100%	100%	98.39%	98.3%	100%
4ASK	100%	100%	100%	100%	99.8%	100%
2FSK	100%	100%	100%	99.5%	99.5%	100%
4FSK	100%	100%	100%	98.3%	98.5%	100%
2PSK	100%	100%	100%	99.3%	99.3%	99.3%
4PSK	100%	100%	100%	98.8%	98.8%	99.8%
16QAM	100%	100%	100%	-	-	-

commonly used digital modulation signals and did not involve other types of digital modulation signals or analog modulation signals. With the continuous emergence of new modulation methods, it is necessary to study automatic recognition algorithms suitable for a wider range of modulation signals.

REFERENCES

- [1] Inoue, T., Matsumoto, R., & Namiki, S. (2022). Learning-based digital back propagation to compensate for fiber nonlinearity considering self-phase and cross-phase modulation for wavelength-division multiplexed systems. *Optics Express*, 30(9), 14851-14872.
- [2] Yang, H., Niu, Z., Zhao, H., Xiao, S., Hu, W., & Yi, L. (2022). Fast and accurate waveform modeling of long-haul multi-channel optical fiber transmission using a hybrid model-data driven scheme. *Journal of Lightwave Technology*, 40(14), 4571-4580.
- [3] Mao, H. (2022). Information Processing Methods of Electronic Warfare Events Based on Communication Technology. *Security and Communication Networks*, 2022, 1-11.
- [4] Argyris, A. (2022). Photonic neuromorphic technologies in optical communications. *Nanophotonics*, 11(5), 897-916.
- [5] Kumar, C. (2022). Performance Analysis of a DPSK Modulated Ultra-Dense WDM System at Different Bit Rates. *Optical Memory and Neural Networks*, 31(2), 191-205.
- [6] Zeng, H., Gong, S., Wang, L., Zhou, T., Zhang, Y., Lan, F., ... & Mittleman, D. M. (2022). A review of terahertz phase modulation from free space to guided wave integrated devices. *Nanophotonics*, 11(3), 415-437.
- [7] Cho, J., & Tkach, R. (2022). On the kurtosis of modulation formats for characterizing the nonlinear fiber propagation. *Journal of Lightwave Technology*, 40(12), 3739-3748.

- [8] Zeng, H., Gong, S., Wang, L., Zhou, T., Zhang, Y., Lan, F., ... & Mittleman, D. M. (2022). A review of terahertz phase modulation from free space to guided wave integrated devices. *Nanophotonics*, 11(3), 415-437.
- [9] Dat, P. T., Umezawa, T., Kanno, A., Yamamoto, N., & Kawanishi, T. (2022). High-speed fiber-wireless-fiber system in the 100-GHz band using a photonics-enabled receiver and optical phase modulator. *Optics Letters*, 47(5), 1149-1152.
- [10] Bai, W., Zou, X., Li, P., Ye, J., Yang, Y., Yan, L., ... & Yan, L. (2022). Photonic millimeter-wave joint radar communication system using spectrum-spreading phase-coding. *IEEE Transactions on Microwave Theory and Techniques*, 70(3), 1552-1561.
- [11] Ren, S., Lai, W., Wang, G., Li, W., Song, J., Chen, Y., ... & Zhou, P. (2022). Experimental study on the impact of signal bandwidth on the transverse mode instability threshold of fiber amplifiers. *Optics Express*, 30(5), 7845-7853.
- [12] Taravati, S., & Eleftheriades, G. V. (2022). Microwave space-time-modulated metasurfaces. *ACS Photonics*, 9(2), 305-318.
- [13] Lee, D., & Chung, W. (2022). Improving the memory efficiency of RTM using both Nyquist sampling and DCT based on GPU. *Journal of Geophysics and Engineering*, 19(4), 706-723.
- [14] Kato, Y. (2022). Fault Diagnosis of a Propeller Using Sub-Nyquist Sampling and Compressed Sensing. *IEEE Access*, 10, 16969-16976.
- [15] Song, Y., Zhang, J., Jin, S., Li, G., & Bi, H. (2022). Frequency-Scaling-Based Spaceborne Squint SAR Sparse Imaging. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 8064-8073.
- [16] Li, S., Wei, Z., Yuan, W., Yuan, J., Bai, B., Ng, D. W. K., Hanzo, L. (2022). Faster-than-Nyquist asynchronous NOMA outperforms synchronous NOMA. *IEEE Journal on Selected Areas in Communications*, 40(4), 1128-1145.
- [17] Xie, J., Zhang, J., Zhang, Y., & Ji, X. (2022). PUERT: Probabilistic Under-Sampling and Explicable Reconstruction Network for CS-MRI. *IEEE Journal of Selected Topics in Signal Processing*, 16(4), 737-749.
- [18] Yadav, S., Sadique, M. A., Ranjan, P., Khan, R., Sathish, N., & Srivastava, A. K. (2022). Polydopamine decorated MoS₂ nanosheet based electrochemical immunosensor for sensitive detection of SARS-CoV-2 nucleocapsid protein in clinical samples. *Journal of Materials Chemistry B*, 10(41), 8478-8489.
- [19] Liu, Y., & Liu, B. (2022). Residual analysis and parameter estimation of uncertain differential equations. *Fuzzy Optimization and Decision Making*, 21(4), 513-530.
- [20] Gabbard, H., Messenger, C., Heng, I. S., Tonolini, F., & Murray-Smith, R. (2022). Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy. *Nature Physics*, 18(1), 112-117.
- [21] Ahmad, S., & Aslam, M. (2022). Another proposal about the new two-parameter estimator for linear regression model with correlated regressors. *Communications in Statistics-Simulation and Computation*, 51(6), 3054-3072.
- [22] Liu, Y., Liu, B. (2022). Estimating unknown parameters in uncertain differential equation by maximum likelihood estimation. *Soft Computing*, 26(6), 2773-2780.
- [23] Xu, L. (2022). Separable Newton recursive estimation method through system responses based on dynamically discrete measurements with increasing data length. *International Journal of Control, Automation and Systems*, 20(2), 432-443.

Edited by: B. Nagaraj M.E

Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing

Received: Dec 20, 2023

Accepted: Feb 27, 2024



APPLICATION OF ARTIFICIAL INTELLIGENCE TECHNOLOGY AND DEEP LEARNING IN LABORATORY INTELLIGENT MANAGEMENT PLATFORM

XING LU*

Abstract. In order to effectively utilize data for laboratory management, a laboratory management model was studied, the author proposed a data-driven intelligent laboratory management process and logical architecture. For actual management work, there are mainly two types of operations: "Selection" and "action", the author proposes a data-driven laboratory intelligent management process and logical architecture; Based on the idea of big data, label systems are used to classify and store laboratory related data and laboratory evaluation GBDT and other algorithmic models; Building an intelligent laboratory management platform based on the label system has realized laboratory management functions, which are widely used and highly scalable. This data-driven laboratory intelligent management platform plays a role in the entire life cycle of laboratories, including laboratory demonstration construction, construction process management, experimental teaching and open use, operation management and maintenance, and experimental effect evaluation, and can promote the maximum effectiveness of laboratories, provide strong support for later construction project approval.

Key words: Artificial intelligence technology, Intelligent laboratory management, Application, Laboratory Evaluation GBDT Algorithm

1. Introduction. The term "artificial intelligence" was coined by McCartney, Minsky, Rochester and Shenon First proposed by a group of young scientists, it marks the emerging science of "artificial intelligence" Generation of family. Artificial intelligence has many advantages, including the following points: First, artificial intelligence can greatly save human cost. Second, artificial intelligence can greatly improve resource utilization. Third, AI can greatly improve work efficiency. Fourth, artificial intelligence has high commercial value. Fifth, artificial intelligence can free people's hands to focus on a better life. Sixth, artificial intelligence can promote social development and human progress. With the development of global economic integration, information technology has made great progress, Artificial intelligence technology has been widely used, such as car autonomous driving, robots Automatic sweeping, robot automatic cooking, robot waiter, rescue and disaster relief robot, Underwater robots and dance performance robots.

With the development of science, technology, and information technology, artificial intelligence theory has attracted more and more attention in recent years, not only because artificial intelligence technology can improve the efficiency of production and work, but also because of the emergence and application of artificial intelligence technology, it has greatly liberated human hands and is an important symbol of humanity's progress towards a new stage. It is conducive to understanding artificial intelligence technology [1]. Studying the theory of artificial intelligence, especially the transformation results of representative artificial intelligence technologies, is conducive to the widespread application of artificial intelligence technology, thereby reducing production costs, improving work efficiency, and benefiting the economic development of enterprises, contribute to China's economic and social development. It is conducive to the development of relevant industries and disciplines. Studying representative AI technologies is not only conducive to guiding the development of AI technology in the industry, but also conducive to the integration of other disciplines with AI, achieving the effect of one plus one greater than two, and promoting the development of relevant industries and disciplines. Third, it is conducive to stimulating the enthusiasm of the whole society for innovation. Through the popularization of artificial intelligence technology, people can realize that artificial intelligence technology is a discipline closely related to our daily life, through the transformation of theoretical achievements in artificial intelligence, it is possible to cultivate the enthusiasm of the whole society for creation and invention, thereby increasing the

*State-owned Assets Management Office, Jilin Agricultural University, Changchun, Jilin, 130118, China (XingLu167@163.com)

vitality of the development of artificial intelligence technology [2]. With the deepening of educational reform and the increasing demand for innovative, skilled, and talented people in society, investment in the construction scale, equipment, and practical teaching arrangements of university laboratories continues to increase, it has played an important role in cultivating socially applicable talents in colleges and universities. While meeting the needs of teaching practice, university laboratories also undertake heavy scientific research tasks, so the traditional manual management model is well suited to the new management requirements. Factors such as the increase in laboratory operation time and instrument usage frequency, as well as insufficient management personnel, not only reduce management efficiency, but also bring various safety hazards. Establishing innovative, open, and resource sharing central laboratories and adopting artificial intelligence technology for scientific management have become an inevitable trend in the development of university laboratories [3,4]. What is the significance of studying artificial intelligence: First, it is conducive to understanding artificial intelligence technology. Research artificial intelligence theory In particular, the main research on the transformation of representative artificial intelligence technologies, It is conducive to the wide application of artificial intelligence technology, thus reducing production costs and improving production Work efficiency, is conducive to the economic development of enterprises, for China's economic and social development Make a contribution. Second, it is conducive to the development of related industries and disciplines. Research is representative Is not only conducive to guiding the development of artificial intelligence technology in the industry Exhibition, but also conducive to the combination of other disciplines and artificial intelligence, to achieve one plus one greater than The effect of two, promote the development of related industries and disciplines. Artificial intelligence technology has It is widely used in all aspects of society. In the current background of rapid economic development and Under the reforming and opening environment, only artificial intelligence technology that suits our national conditions is long For a long time, it will promote the sustainable development of science and technology, sustainable development, sound and rapid development Exhibition. Based on the current situation of laboratory management, this paper mainly studies the application of artificial intelligence technology In the laboratory management of advantages, the main technology, in order to improve and develop artificial intelligence The application of technology in laboratory management is even popularized in the whole society.

2. Methods.

2.1. Laboratory management model based on management process . With the development of the Internet, the information collected in the Internet is timely Feedback to the lab's network platform, through cloud computing and analytics technology, the information Processing, compare the camera data and cloud computing database, handle Cloud computing results, realize the automation of laboratory management, intelligent. The laboratory serves teaching and research work, and its management objects include people, events, materials, information, funds, etc. It involves all activities of laboratory application, construction, and experimental teaching, mainly including: Laboratory construction planning and setup, laboratory management mode and operating mechanism, configuration and use of laboratory instruments and equipment, management of experimental materials and low value consumables, basic laboratory information management and archive management, construction and training of experimental teaching teams, management of experimental teaching and scientific research, and use and inspection of laboratory funds. It can be seen that laboratory management is relatively complex and involves many aspects, but these tasks can be subdivided into specific tasks, from a specific project perspective, the basic model of management work is shown in Figure 2.1 [5].

As can be seen from Figure 2.1, no matter how much management content there is, for a specific project management work, it can be summarized into two steps: "selection" and "action"; "Select" refers to selecting management content, and "Action" refers to generating management results such as reports and emails after appropriate processing. "The operation of selecting management content is actually to limit the content to meet certain requirements, such as when an experimental teacher conducts experimental course management, according to the selected experiment 1 score of 80 or more, moreover, with a theoretical course score of 80 or above and not being a make-up or re major student, the designated students are selected and selected for elective experimental courses. This process is limited to specific experimental objects and can include multiple conditions, the subdivision model is shown in Figure 2.2 [6].

As can be seen from Figure 2.2, through the logical combination of multiple conditions, you can select a management object and then perform corresponding "actions" on the object, "selection" is the basis, and

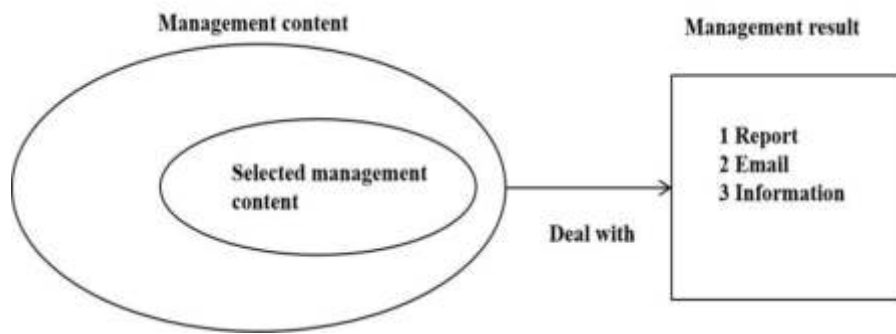


Fig. 2.1: Basic model of laboratory management

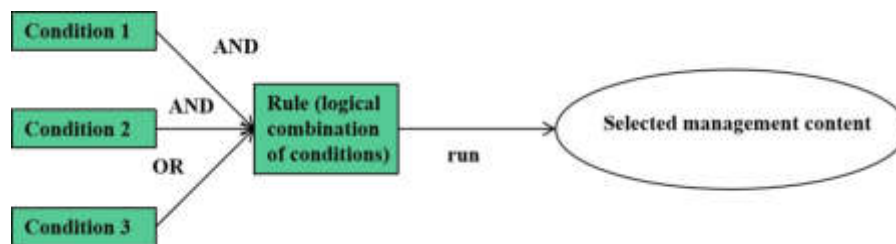


Fig. 2.2: Selection Management Content Segmentation Model

”Action” is the operation based on actual needs. ”Selection” is mainly based on various data sources [7].

2.2. Data-driven laboratory intelligent management platform framework . The project function of laboratory management system based on artificial intelligence technology It mainly includes the following: (1) Facial recognition personnel ID and automatic registration information. (2) Intelligent power distribution function. (3) Remote communication.

Facial recognition personnel ID through the facial features collection and recognition function of electronic eye technology, can identify every A person enters the lab, the electronic eye recognizes facial information and sends a message Information processing and existing database for information comparison, record visitor information, to achieve paperless facial registration.

The power distribution function of intelligent system refers to the use of mature solar energy, wind energy and so on Electric technology, to achieve the continuous circulation of laboratory electricity, will not accidentally cut off the movie The development of laboratory work, when the power supply is insufficient, the intelligent system will automatically lift At the same time, through solar and wind power generation technology to achieve the laboratory electricity reserves,

To achieve continuous uninterrupted laboratory power supply effect. Another way of telecommunication That is, the manager of the laboratory can observe the reality through the intelligent monitoring camera Laboratory conditions, improve the observation times, increase the observation duration, in order to timely understand the laboratory The latest internal dynamics, flexible handling of various laboratory situations.

According to the laboratory management model, in order to implement the ”select” action, it is necessary to make reasonable use of data to formulate rules, but the specific requirements for each management role may vary, rule making is also different, for experimental teachers, experimental technology, and system managers, their management processes are shown in Figure 2.3 [8,9].

The laboratory management platform is divided into a front-end and a back-end, the front-end uses Web pages for user operations, and the back-end uses logical computing for front-end display. Logic and data support. For a data-driven laboratory management platform, the front-end is used by business personnel in business

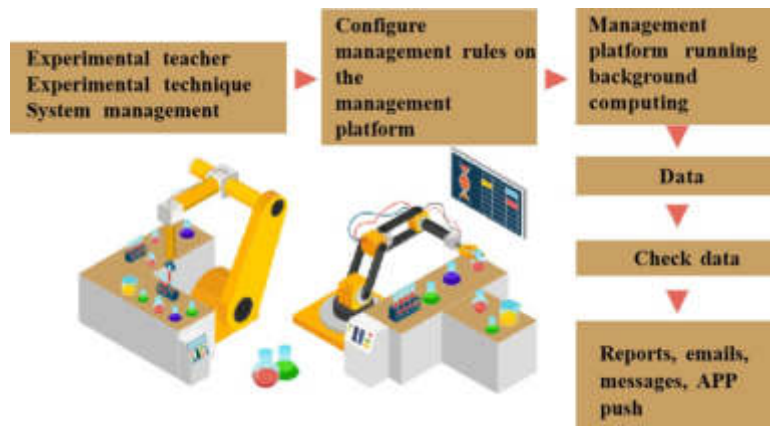


Fig. 2.3: Data-driven laboratory intelligent management process

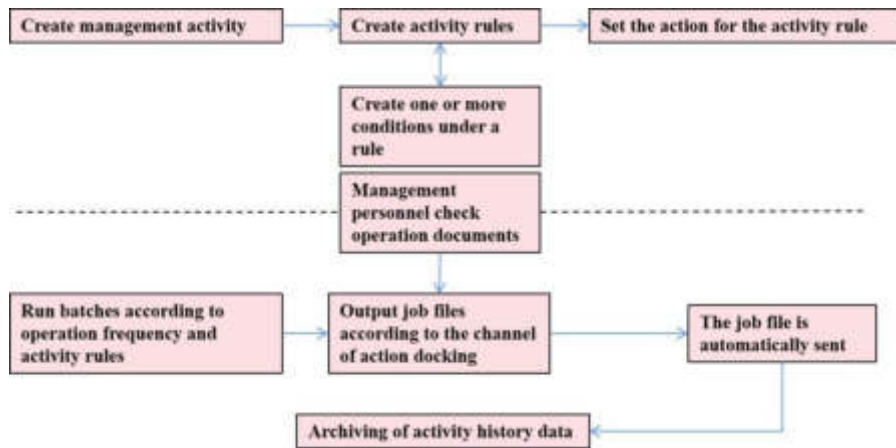


Fig. 2.4: Logical architecture of data-driven laboratory intelligent management platform

departments, such as laboratory teachers and laboratory technicians; The responsible person for the backend is the data engineer and system development engineer in the IT department. According to the laboratory management model, the logical framework of a data-driven laboratory intelligent management platform is shown in Figure 2.4 [10].

Managers log on to the management platform, first create a new laboratory management, set information such as the management name, responsible person, time range, and running frequency, and then create one or more rules under this laboratory management, create different combinations of conditions under each rule, and finally set an action for the created rule. This completes a basic data-driven laboratory management configuration [11].

2.3. Implementation of Lab Intelligent Management Platform Based on Label System . In order to achieve intelligent management on the above management platform, another important step is to set the conditions of activity rules, we use the label system to implement labels, which refer to data labels, which describe entity attributes, the value of the label marks a piece of information about the entity, for example, for a student, "gender" is a label, and "male" is the value of this label, which marks the gender information of the student. The label system is a collection of labels that are calculated and stored according to certain rules, it performs class management on labels according to established logic, and calculates and updates label values

Table 2.1: Various laboratory data labels

Label Theme	Base Label	Behavior label	Titles
Experi- -mental Courses	Name, course, type (course experiment, open experiment, online experiment), experiment content, goal, first opening time, responsible person	Last Opening Time, Last Opening Number, Hours Opened this Month, Hours Opened this Month, Hours Opened this Semester, number of Users this Semester	The number of days since the first opening, cumulative open class hours, cumulative number of open students, and cumulative extracurricular open class hours
Experimental equipment	Name, manufacturer, production date, price, first use time, storage location, responsible person	Last usage time, last usage time, usage time of this month, usage time of this semester, last maintenance time, performance status	Cumulative usage time, cumulative usage times, open usage times, loan times
Experimental site	Site name and size. Number of work stations, first use time, number of storage equipment, responsible person	Last usage time, last usage time, usage time of this month, usage time of this semester, power consumption data, monitoring access control	Cumulative usage time, cumulative usage times, open usage times
student	Name, date of birth, class, mobile phone number, email address, start date of experiment	Last experiment time, last experiment content, last experiment result, number of experiments this month, content of experiments this semester	Cumulative number of experiments, cumulative experimental hours, in class experimental hours, open experimental hours
teacher	Name, unit, personnel type (full-time teacher, part-time technician), mobile phone number, email address, start of experimental course, time of first course opening	Last experiment time, last experiment content, last number of experiments this month, content of experiments this month, number of experiments this semester	Cumulative number of experiments, cumulative experimental hours, in class experimental hours, open experimental hours

according to rules, data access issues are resolved through pre calculated tags, reducing the threshold for data usage [12,13].

(1) *Laboratory Data Label System*. The label system is a collection of a series of labels, which can be divided into experimental courses, experimental equipment, venues, students, teachers, administrators, and other topics, these topics are divided into basic tags, behavior tags, and derived tags by data update method, as shown in Table 2.1 [14].

As can be seen from Table 2.1, label topics are divided into basic labels, behavior labels, and derived class labels. The value of the base tag is generally fixed or has a long update cycle, incremental updates are used

Table 2.2: Mapping Table of Label System Hierarchy and Data Backend

Hierarchical structure	Relational database	HBase
Label Theme	Schema	Table Name
Label Type	Table Name	Clan name
label	Column Name	Column Name

to refresh the tag value, updating only tags with changed values each time or inserting newly added customer base tags; The behavior tag is used to describe historical behavior, which is always in change, and adopts periodic full volume updates or real-time (message queue+stream processing) fixed point updates; Derived class tags are logical combinations between other tags, they do not store tag values themselves, but rather store computational logic between tags, the tag value is calculated in real time only when called, which is a special dynamic tag. In this way, experimental data can be converted into label data, and the label system, like the management system, should also be divided into two parts: The front end and the back end [15].

The label administrator uses the front end (management) page of the label system, configure the mapping relationship between the label and the data background, the configuration information is stored in the label mapping Table, after the label user enters query criteria on the front-end (query) page of the label system, the system first locates the physical location of the label through the label mapping Table, then, the corresponding label value is read from the label data background and returned to the page side for display, the label mapping Table is associated with the front-end and back-end of the label system, it stores all attributes in the label system except for the label value, including all description information about the label hierarchy, the backend of the system stores the values of all labels, which are stored through relational database tables or HBase Tables, the label mapping table is shown in Table 2.2, the values of the corresponding labels can be intelligently located based on the information sequence stored in the label mapping Table [16].

(2) *Labelled Laboratory Evaluation GBDT Model.* In the actual management processes such as laboratory construction demonstration, course effectiveness evaluation, and laboratory benefit evaluation, how to effectively use data pairs for scoring and evaluation is a matter of great concern to all parties, the author adopts GBDT (GradientBoosting Decision Tree) regression algorithm to model the previous experimental data. However, in the GBDT modeling phase, a large amount of computation is required, so the modeling process is completed in the back-end through offline computation, and the established GBDT model is converted into tags for online use. Taking the benefit evaluation of a new laboratory as an example, in the GBDT modeling stage, based on the previously stored student experimental data, performance, equipment purchase prices and updates, equipment usage data, laboratory electricity, access control, and other data, models such as in-class experiment scores, open experiment scores, equipment usage benefit scores, equipment sharing scores, equipment depreciation rate scores, and site operation efficiency scores are established on the server backend, store these models as label data. When evaluating the benefits of a new laboratory, call these models on the Web side and input the application data for the new laboratory, the corresponding score can be quickly obtained for reference by the evaluation experts. The laboratory evaluation is implemented using the GBDT algorithm, with historical experimental data as training data, after training using the GBDT algorithm, the model functions are stored as tags for online invocation [17].

3. Management process example based on label system. The laboratory intelligent management platform implements laboratory management through configuration management rules. The following is an example of pushing selected experiments to outstanding students to illustrate the management process based on the label system, the corresponding table of conditions and labels for this management activity is shown in Table 3.1.

In order to configure this rule, managers need to add the three tags in Table 3.1: iEX_Score iTH_Score BRE_EXA sets the corresponding conditions and connects with AND, you can manually edit logical relationships to make adjustments, Table 4 and Table 5 show the key page for rule settings. It can be seen that the label system needs to provide as many public labels as possible to meet as many rule (condition) setting requirements as possible. Once a label user discovers that a label required by the condition does not exist, they

Table 3.1: Condition and Label Correspondence Table

Condition	Corresponding label name	Conditions after converting to labels
Experiment 1 with a score of 80 or above	Experiment 1 Score (EX_Score)	iEX_Score>=80
Theoretical course score above 80 points	Theoretical Course Score (iTH_Score)	iTH_Score>=80
Not a retake student	Makeup Exam (bRE_EXA)	bRE_EXA("TRUE")

Table 3.2: Label Settings

Label Name	Tag ID	Label Value Type	Subject	Label Type
Experiment 1 Score	35	Integer	student	Behavior label
Theoretical Course Score	36	Integer	student	Base label
A make-up exam	37	BOOL	student	Base label

Table 3.3: Selected Conditions for Management Platform Rules

Condition 1	Experiment 1 Score	>=	80	Delete this condition
Condition 2	Theoretica Course Score	>=	80	Delete this condition
Condition 3	A make-up exam	==	TRUE	Delete this condition

need to submit a new label request to the label manager, after the label manager adds the new label to the label system, the label user can see and use the label on the above page of the management platform [18].

When the rule setting page is submitted, the logical relationship corresponding to the rule is saved to the background database, the tag is stored in the form of a tag ID, which allows you to further find the fact Table where the tag is located and obtain the corresponding tag value. If the final logical relationship value is true, it indicates that the student meets the rule; If the result is false, it indicates that the customer does not meet the rule, so the optional experiment is not recommended. The Run Frequency option in the interface is "Every day", which refers to the logical combination of the rules converted into conditions through the page, parse the conditions into SQL statements, perform batch processing in the background, and store the results in a result Table, the subsequent management action stage will produce different formats of job files, or production reports or email job files based on the result Table. In addition to batch processing, you can also choose scenario based management, and management activities can be processed in real-time based on scenarios, for example, after a student has completed Experiment 1 and received a system evaluation score of 80 or more, they can be directly recommended for the experiment. This requires changing batch processing to "real-time processing", the management process is actually consistent, and only technically requires the introduction of "message queues" to complete management based on these messages [19,20].

4. Conclusion. The development of intelligent technology has for the development of our society It plays an important role in applying artificial intelligence technology to laboratory management Now the trend of intelligent and automatic laboratory management. Artificial intelligence technology should Used in laboratory management work, such as facial recognition personnel ID, automatic registration letter The intelligent power distribution function ensures sufficient power supply to the laboratory and prevents accidental breaks The uncontrollable loss caused by electrical accidents can be realized by the remote communication technology of artificial intelligence Laboratory management personnel remotely monitor the internal conditions of the laboratory to find laboratory differences in time Often, take timely measures to nip in the bud. Applying artificial intelligence technology to the real world The development process of laboratory management is not smooth sailing, nor can it be accomplished overnight Need us to give full play to their own subjective initiative, actively contribute to daily life From quantity to quality, the innovation in our science and technology development

and social progress Make contributions and strive to realize the great Chinese Dream. Structured data brings value, and data brings new ideas, with the application of new devices and the Internet of Things in laboratories, there are more and more sources of experimental related data, and more and more management bases are available. Based on the reality of laboratory management, the author studied a laboratory management model. Using a label system to classify and store laboratory data and laboratory evaluation GBDT and other algorithmic models, an intelligent laboratory management platform based on the label system is constructed, realizing batch processing and scenario based laboratory management, the intelligent management platform has strong scalability and can play a role in the entire life cycle of the laboratory.

REFERENCES

- [1] Lv X , Li M . Application and Research of the Intelligent Management System Based on Internet of Things Technology in the Era of Big Data[J]. Mobile Information Systems, 2021, 2021(16):1-6.
- [2] Shan T , Tay F R , Gu L . Application of Artificial Intelligence in Dentistry[J]. Journal of Dental Research, 2021, 100(3):232-244.
- [3] Galvan P , Fusillo J , Portillo J , et al. PP119 Innovative Screening System For COVID-19 Using Application Of Artificial Intelligence For Telemedicine[J]. International Journal of Technology Assessment in Health Care, 2021, 37(S1):20-20.
- [4] Dong Z , Sheng Y , Huang Z , et al. Living Cell Nanoporation and Exosomal RNA Analysis Platform for Real-Time Assessment of Cellular Therapies[J]. Journal of the American Chemical Society, 2022, 144(21):9443-9450.
- [5] Chen S , Huang J , Gao Z . Development of Intelligent Management System for High Value Consumable Material in Operating Room[J]. Zhongguo yi liao qi xie za zhi = Chinese journal of medical instrumentation, 2021, 45(1):42-45.
- [6] Peng J . Oil Painting Material Collection System Based on Artificial Intelligence[J]. Journal of Physics: Conference Series, 2021, 1852(2):022029-.
- [7] Liu Y , Wang Z , Pan Y , et al. Research on Intelligent Monitoring and Early Warning of Electric Power Safety Based on Artificial Intelligence Technology[J]. Journal of Physics: Conference Series, 2021, 1748(5):052046 (5pp).
- [8] Kuang L , Liu H , Ren Y , et al. Application and development trend of artificial intelligence in petroleum exploration and development[J]. Petroleum Exploration and Development, 2021, 48(1):1-14.
- [9] Ji J , He Y . Application of Artificial Intelligence in Computer Network Technology[J]. Journal of Physics: Conference Series, 2021, 1881(3):032073 (5pp).
- [10] Xu K , Wang Z , Zhou Z , et al. Design of industrial internet of things system based on machine learning and artificial intelligence technology[J]. Journal of Intelligent and Fuzzy Systems, 2021, 40(2):2601-2611.
- [11] Yang X , Li H , Ni L , et al. Application of Artificial Intelligence in Precision Marketing[J]. Journal of Organizational and End User Computing, 2021, 33(4):209-219.
- [12] Lu W . Research on Marketing Development under the Background of Artificial Intelligence Technology[J]. Journal of Physics Conference Series, 2021, 1769(1):012071.
- [13] Mase C , Maillard J F , Paupy B , et al. Speciation and Semiquantification of Nitrogen-Containing Species in Complex Mixtures: Application to Plastic Pyrolysis Oil[J]. ACS Omega, 2022, 7(23):19428-19436.
- [14] Song Y , Zhao Z , Zheng Y , et al. Investigation and Application of High-Efficiency Network Fracturing Technology for Deep Shale Gas in the Southern Sichuan Basin[J]. ACS Omega, 2022, 7(16):14276-14282.
- [15] Gao J , Dong L I , Dan S U , et al. Research Progress on the Extraction Technology of Seabuckthorn Fruit Oil and the Application of Nutritional Factors[J]. Science and Technology of Food Industry, 2022, 43(13):400-407.
- [16] Xu X , Li X , Gao X , et al. Application of Large-Scale Molecular Prediction for Creating the Preferred Precursor Ions List to Enhance the Identification of Ginsenosides from the Flower Buds of Panax ginseng[J]. Journal of Agricultural and Food Chemistry, 2022, 70(19):5932-5944.
- [17] Zhang C , Du B , Li K , et al. Selection of the Effective Characteristic Spectra Based on the Chemical Structure and Its Application in Rapid Analysis of Ethanol Content in Gasoline[J]. ACS Omega, 2022, 7(23):20291-20297.
- [18] Hung H M , Duc L M , Dat N D , et al. Synthesis and Characterization of Polypyrrole Film Doped with Both Molybdate and Salicylate and Its Application in the Corrosion Protection for Low Carbon Steel[J]. ACS Omega, 2022, 7(23):19842-19852.
- [19] Cao D , Chen M , Liu Y , et al. Ion Migration in the All-Inorganic Perovskite CsPbBr₃ and Its Impacts on Photodetection[J]. The Journal of Physical Chemistry C, 2022, 126(23):10007-10013.
- [20] Miskin C K , Pradhan A A , Deshmukh S D , et al. Solution Processed Fabrication of Se-Te Alloy Thin Films for Application in PV Devices[J]. ACS Applied Energy Materials, 2022, 5(3):3275-3281.

Edited by: B. Nagaraj M.E

Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing

Received: Dec 20, 2023

Accepted: Feb 27, 2024



A MACHINE INTELLIGENCE EVALUATION SYSTEM BASED ON INTERNET AUTOMATION TECHNOLOGY AND DEEP LEARNING

HONGCHUAN LIU*

Abstract. To realize the machine intelligence evaluation system, a method based on Internet automation technology is proposed. Firstly, then extracted and optimized, and finally combined with each other. A BP neural machine evaluation system is designed to compare the results of machine evaluation with the average value of teachers' independent evaluation by selecting 20 students' test paper translation samples from a class randomly. The test results show that by selecting a random class of 20 students, the comparison of machine evaluation results and teacher independent evaluation shows that the error range of the evaluation results of 20 samples is -5.6% -6.7%, which is within the allowable range of translation evaluation and meets the requirements of teaching evaluation. It is proved that the Chinese-English machine translation evaluation system based on Internet automation technology has excellent performance, which can improve the reliability and accuracy of the evaluation and reduce the degree of human intervention and misjudgment rate of the Chinese translation evaluation.

Key words: Internet, automation, evaluation system, translation, misjudgment rate, Intelligent evaluation system

1. Introduction. Intelligent characteristics are one of the important characteristics of intelligent systems. Qualitative evaluation of the intelligent characteristics of intelligent systems is a challenge and a new issue. Because, firstly, people's definition of the concept of intelligent characteristics itself is not clear enough. Secondly, the evaluation of intelligent characteristics is related to the environment, era, and conditions. Therefore, this evaluation has relativity, correlation, and time effects. In addition, there is very little specialized research and communication on this evaluation. And the theory and application of intelligent control continue to develop. It is necessary to conduct specialized research on the evaluation of intelligent characteristics in a timely manner, as various products and systems with the term "intelligent" are constantly entering the market. This is not only an academic need, but also an application need. Intelligent system is a system with anthropomorphic intelligence. anthropomorphic intelligence is the intelligent characteristic of simulating, extending and expanding human. Such as: self-learning, self-adaptation, self-organization, self-optimization, self-stabilization, self-identification, white planning, self-coordination, self-repair, self-reproduction, etc. Because the intelligence of human body control system is multi-level and multi-faceted, the intelligence of anthropomorphic system is also divided into different levels and different aspects such as high level, middle level and basic level. In order to realize the intelligent characteristics in the system, the commonly used intelligent methods include expert system, artificial neural network, fuzzy control and so on. With the development of the Internet and the advent of the era of economic globalization, the need to overcome language barriers and realize free communication across languages has become increasingly prominent [1]. The language barrier severely restricts the breadth, depth and speed that most users can obtain information from the Internet [2]. However, the development of advanced machine translation technology and the realization of large-scale application of machine translation products pose new challenges to the machine translation technology.

With the increasing progress and development of modern science and technology such as "Internet +" and artificial intelligence, people's work, study and life have been closely related to the modern information technology, and people rely on intelligent technology and tools increasingly [3]. Computer-aided translation means that translators can improve translation efficiency and control translation costs effectively by scientifically selecting language translation tools based on Internet, artificial intelligence and big data technologies. The computer translation technology emerged at the end of the 20th century, providing a technical support for

*School of cultural communication, Institute of disaster prevention, Langfang, Hebei, 065201, China (HongchuanLiu5@163.com)

people's scientific research activities, work and life, and promoting the cross-language communication [4,5].

In foreign countries, the research on this technique can only provide some reference for translators. Due to the limitation of the algorithm level, early computer translation is difficult to form a smooth and reasonable translation text. In the context of the rapid development of the Internet, big data and artificial intelligence provide strong support for the development of this technology and promote CAT technology to achieve great progress. Many translation websites and platforms based on CAT technology provide convenience for translators and relieve the pressure of translation work. CAT technology has been further developed with the help of translation memory libraries. Based on a brief review of the history of machine translation, the research discusses the existing methods of machine translation, and then discusses the challenges and technical routes of Internet machine translation.

2. Literature Review. Broadly speaking, machine translation involves all aspects of natural language processing technology, and almost all the research results of natural language processing can be directly or indirectly applied to machine translation. In a narrow sense, machine translation methods can generally be divided into three categories: rule-based machine translation, case-based machine translation and statistical machine translation, of which the latter two methods can be collectively referred to as corp-based methods [6]. The rule-based translation approach, which holds that the process of translation needs to analyze the source language and express the meaning of the source language, and then regenerate into an equivalent target language, has been dominant in the field of machine translation from the mid-1970s to the late 1980s. A large-scale rule-based commercial machine translation system should not only solve the problem of machine translation methodology, but also organize the system from the perspective of knowledge engineering and software engineering, in which the rules are often multi-level and fine-grained. The refinement of rule level and knowledge granularity can control the interaction and conflict between rules effectively, and make the rule system have good expansibility.

The essence of case-based machine translation is "machine translation based on translation instance and similarity principle". Translation instances can be stored in their natural form without any processing, or they can be represented in a completely structured form. Recent researches show that semi-structured translation instance representation approaches strike a good balance between the difficulty of preprocessing translation instances, the temporal and spatial efficiency of translation and the quality of translation. Another on the principle is very similar with case-based machine translation technology is translation memory, which is a computer-aided. It is a kind of auxiliary translation in essence. It retrieves similar translation instances from the instance library and submits them to users in a friendly form, thus achieving the purpose of assisting users in translation [7]. In recent years, translation memory technology is increasingly integrating various automatic translation technologies [8]. Statistical machine translation is also based on bilingual corpus, but unlike the case-based method, which directly uses translation examples in the translation process. Statistical method abstracts the translation knowledge implied in bilingual corpus into statistical model through prior training process. The translation process is usually a decoding process based on these statistical models. Statistical models used in statistical machine translation usually include translation model and language model [9]. Compared with language model and decoding, translation model is currently the most involved content in statistical machine translation research [10]. Generally, translation models can be divided into three types: word-based model, phrase-based model and grammar-based model. At present, phrase-based and grammar-based models have significantly better performance than word-based models. Although statistical methods are valued for their good mathematical model, unguided learning ability and robustness, rule methods are also valuable for their good generalization and description of language rules and instance methods for the accurate translation of similar sentences. In fact, the combination of multiple methods is becoming an important direction in the development of machine translation, such as the combination of rules and statistical methods, case-based methods and statistical methods, phrase-based and syntactic statistical translation methods.

On the basis of the current research, a machine translation evaluation system for TCSL based on Internet automation technology is proposed in the research. Firstly, then extracted and optimized, and finally combined with each other. A BP neural machine evaluation system is designed to compare the results of machine evaluation with the average value of teachers' independent evaluation by selecting 20 students' test paper translation samples from a class randomly. The test results show that by selecting a random class of 20 students,

Table 3.1: Description of statement set information

The statement text	Number of translation	Number of statement text	Scores range	Number of sentence patterns
1	1650	10	0-3	8
2	1825	10	0-3	8
3	1756	6	0-3	4
4	1622	6	0-3	4

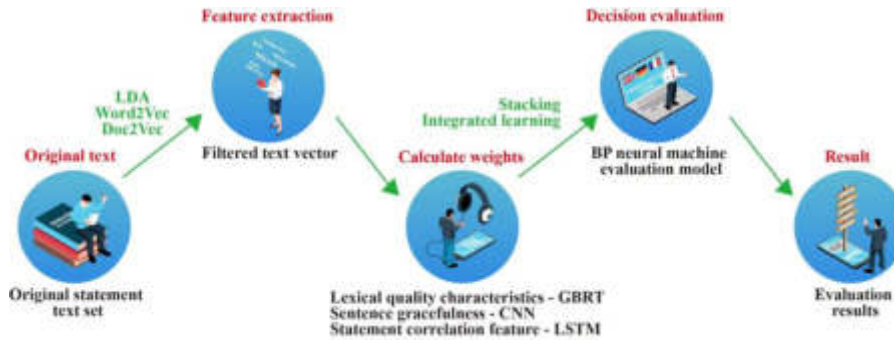


Fig. 3.1: Basic framework of ETSS system (high-end)

the comparison of machine evaluation results and teacher independent evaluation shows that the error range of the evaluation results of 20 samples is -5.6% -6.7%, which is within the allowable range of translation evaluation and meets the requirements of teaching evaluation.

3. Research Methods.

3.1. The development and design of ETSS system. In the research, the Chinese translation texts of the university Chinese (1-4) courses of a university in 2019 are selected. The sentence translation texts include the sub-texts of the four courses, with a corresponding number of translation sentences [10, 10, 6, 6]. More than one thousand students answer the translation sentences under each sub-text, and the score of each sentence ranges from 0 to 3. The details are shown in Table 3.1.

3.1.1. Basic system framework. The basic framework of ETSS(English Translation Scoring System) is shown in Figure 3.1. The main functions of the system are divided into three core modules: text feature extraction, weight calculation and decision evaluation [11]. System uses the annotated corpus, i.e., nearly 8 years of text translation for libraries of a university, as well as Chinese corpus, structures, vocabulary quality evaluation model, the beautiful statement and statement relevance evaluation model, evaluation model. Then the input text sets are evaluated separately, and then integrated with BP neural machine model for comprehensive evaluation [12].

3.1.2. Extraction of ETSS feature vectors. Figure 3.1 shows the ETSS feature text library [13]. As can be seen from the system structure diagram in Figure 3.1, the text library W of ETSS system needs to establish the text library h [T] for vocabulary quality evaluation, the text library g [T] for sentence elegance evaluation and the text library l [T] for sentence relevance evaluation. According to the comprehensive summary, the basic characteristics of sentence text judgment include 11 items, such as word accuracy, average word length, number of high-frequency words, number of advanced words, proportion of nouns, adjectives or verbs, number of connectives, number of word blocks, advanced sentence patterns, specialty of key words, word granularity and sentence granularity, etc [14]. There is a progressive relationship between different levels of the system, so the system database design first needs to analyze and integrate different types of text data, and obtain the text feature vector. In order to extract feature vectors quickly, the extraction standards and numbers are

Table 3.2: The feature text database W [i] of ETSS system

The text library	Number	Features	Feature description	Specificity	Weight
h [T] for vocabulary quality evaluation	T_1	Word accuracy	The proportion of words that are spelled correctly	0% ~ 100%	X_1
	T_2	Average word length	Obtained by median or standard deviation	2-20	X_2
	T_3	Number of high-frequency words	Commonly used unmarked words	1 ~ 8	X_3
	T_4	Number of advanced words	Commonly used marker words	1 ~ 5	X_4
	T_5	The proportion of nouns, adjectives and verbs	Average the proportion of the number of words in different parts of speech	0 ~ 100%	X_5
g [T] for sentence elegance evaluation	T_6	Number of connectives cause and effect	Including phrases such as turning point, juxtaposition, choice,	1 ~ 3	X_6
	T_7	Number of word blocks	Including phrasal verbs, prepositional phrases, adverb phrases, adjective phrases and so on	1 ~ 5	X_7
	T_8	Advanced sentence patterns	Including emphasis sentences, clauses, inversion sentences, hypothetical sentences and so on	Y/N	X_8
	T_9	Specialty of key words	The rank of key words in the reference answer	1 ~ 5	X_9
l [T] for sentence relevance evaluation	T_{10}	Word granularity	Describe the relatedness characteristics of words	0 ~ 100%	X_{10}
	T_{11}	Specialty Sentence granularity	Describe the semantic dispersion characteristics of statements	0 ~ 100%	X_{11}

formulated specially, as shown in Table 3.2. In order to ensure that the system data query, modification and update can be saved in advance, it is necessary to adjust the text library system, which is not only conducive to data management, but also convenient for the system to modify the stored procedure of the required data according to the actual demand, and improve the portability of the system source code [15].

After the establishment of ETSS system feature text database W [i], the filtering extraction method of

selected text features is studied. Word2Vec tool model and K-means clustering method are used to filter and extract the word quality features (T1-T5). According to the text library h [T], the text is mapped from the statement text to the feature vector of fixed dimensions, and the text words are encoded. The feature degree and feature base weight are obtained according to the feature description of the text base [16,17]. The feature base weights obtained through experiments (X1~ X5) can be adjusted by modifying the feature degree in practical application to actively adapt to Chinese translated texts of different stages and difficulties. The research is also applicable to the calculation of other weights or weight coefficients later [18,19].

Extraction of feature vectors of sentence elegance is as follows [20]. Elegant Chinese sentences need to integrate advanced lexical blocks, sentence patterns and ingenious rhetorical devices. The identification of elegant Chinese sentences can also be treated as a text classification problem. Doc2Vec tool is used to filter and obtain the text feature information, and then the feature of sentence elegance (T6-T7) is filtered and extracted. According to the text library g [T], the feature base weight (x6-x8) is obtained from the given text label classification training. Due to the influence of language features, cultural background, oral expression and other factors, the feature and degree of sentence elegance are difficult to grasp. Convolutional neural network CNN can be used for modeling feature extraction, and its advantage lies in automatic feature selection and combination, which is used in the subsequent text elegance evaluation in the research [21].

Extraction of sentence relevance feature vectors is as follows. The judgment of sentence relevance is subjective and difficult to be judged by machine, which requires the integration of meaning, sentence meaning and vocabulary. Therefore, the judgment of Chinese sentence relevance needs to combine the above extraction features to comprehensively analyze word granularity, sentence granularity and sentence theme. Here, LDA model is used to read text feature information, and then sentence relevance features (T9-T11) are extracted. According to the text library L [T], subject probability distribution of text is obtained through Bayesian network learning and training, and then feature base weight (x9-x11) is obtained [22].

3.1.3. Feature extraction algorithm. The symbol used to identify or distinguish text is feature. In the research, VSM method of vector space model is used to filter and extract feature information in text. A feature vector is used to represent a Chinese sentence text, which consists of feature terms and weights. The feature vector extracted from the text directly represents the original text, and the extracted optimized feature vector is one of the key factors affecting the results of system evaluation [23]. In VSM model, Chinese text uses space vector $(T_1, X_1; T_2, X_2; \dots; T_j, X_j)$. T_j is the feature item, and X_j is the corresponding basic weight of the feature item, which is used to define the importance of the feature item in the description statement text. In order to improve the accuracy and speed of feature item acquisition, Doc2Vec method, NLTK and StanfordParser toolkit are used for text filtering and extraction processing (including counting, part of speech tagging, average, local maximum and minimum, word frequency weighting, position weighting, syntax analysis, etc.). The text encoding and text feature degree are obtained [24].

Feature vector filtering and extraction methods: Text vector features are obtained through Doc2Vec text parsing. Feature details are obtained through NLTK and Stanford Parser tool package sampling. The second level decomposition is the same, with more detailed space division. It can not only rely on one filtering to extract degree of text feature. And filtering should be continuously repeated several times in order to avoid misoperation accident conditions. By using the wavelet transform and short time Fourier analysis mathematical tools, sentence text features are processed again. Equation 3.1 is used to obtain the square mean root value of the feature degree in the i th time window at the j node, which can achieve better feature discrimination effect [25].

$$X_{j,i} = \left(\frac{1}{N} \sum_{n=1}^N K_{j,n}^2 \right)^{\frac{1}{2}} \quad (3.1)$$

In Equation 3.1, $X_{j,i}$ is the square mean root value of the feature degree in the i th time window at the j node. $K_{j,n}$ is the n th coefficient at the j node. N is the total number of coefficients of j node.

In order to judge the weight of feature degree, the basic assignment table corresponding to feature degree of Chinese sentences is first established, as shown in Table 3.3. Then, based on Chinese sentence rules and translation characteristics, a simulation model of feature weight assignment is established, as shown in Equation

Table 3.3: Chinese sentence feature assignment

Feature	Assignment x				
	1	1.5	2	2.5	3
T_j	1	1.5	2	2.5	3
T_1	10%~20%	21%~40%	41%~60%	61%~80%	81%~100%
T_2	2~5	17~20	6~8	14~16	9~13
T_3	1	2	3~4	7~8	5~6
T_4	1	2	3	5	4
T_5	1%~10% or 91%~100%	11%~30%	31%~50%	51%~70%	71%~90%
T_6	1	-	2	-	3
T_7	1	2	3	4	5
T_8	N	-	-	-	Y
T_9	5	4	3	2	1
T_{10}	10%~20%	21%~40%	41%~60%	61%~80%	81%~100%
T_{11}	10%~20%	21%~40%	41%~60%	61%~80%	81%~100%

3.2. The maximum and minimum values of feature weights are obtained by using reference answers and random answers of translation sentences, and the coefficient changes after wavelet transform are used as the basis of the model. Taking the 5-layer wavelet packet decomposition method as an example, the node importance ratio λ is used, and the ratio of importance between 2 to 5 nodes and the importance of 1 node in Equation 3.3 is taken as the effective weight of feature quantity. The threshold value is set as 0.025, and 11 feature degrees are calculated continuously. In a preset period, the weight of feature quantity is obtained through this simulation model.

$$\frac{1}{g} \frac{dg}{dt} = \frac{1}{\tau} \left(\frac{x_j^2}{x_c^2} - 1 \right) \quad (3.2)$$

$$\lambda_{j,i} = \frac{E}{E_1} = \frac{\sum_{r=2}^{r=j} \sum |u_r(n)|^2}{\sum |u_1(n)|^2} \quad (3.3)$$

In Equation 3.2 and 3.3, g is the derivative value of the characteristic quantity. τ is a time constant; x_j is base weight assignment. x_c is weight assignment coefficient. $\lambda_{j,i}$ is the weight of feature quantity. E is the sum of importance between node 2 and node 5. E_1 is the importance of 1 node. $u_1(n)$ is the reconstruction coefficient of 1 node. $u_r(n)$ is reconstruction coefficient of r node.

3.2. Application of ETSS system.

3.2.1. Comprehensive evaluation of sentence text. The previously extracted sentence text feature vectors are then input into the trained BP neural machine evaluation model for interactive fusion promotion regression verification after stacking learning by sentence vocabulary quality evaluation model GBRT, sentence elegance evaluation model CNN and sentence relevance evaluation model LSTM, respectively. The final score of the sentence text is obtained, and the process of comprehensive evaluation is shown in Figure 3.2.

3.2.2. BP neural machine evaluation method. The proposed ETSS system is based on BP neural network algorithm and machine learning to automatically judge the results of Chinese sentence translation. This evaluation method can consider the language factors of Chinese sentences and judge the correctness of sentence translation relatively accurately. ETSS system introduces the translation result evaluation into BP neural machine algorithm to ensure that input vector and output vector meet nonlinear mapping, which can greatly improve the accuracy of system evaluation results. Based on the above simulation model of feature weight assignment, the weight of feature quantity and the weight coefficient, which affect the translation quality, are regarded as the input layer of BP neural machine, and the output layer of BP neural machine is the judgment value of the system after the decomposition of 5-layer wavelet packet and the calculation of 312 subdivision consecutively within one period.

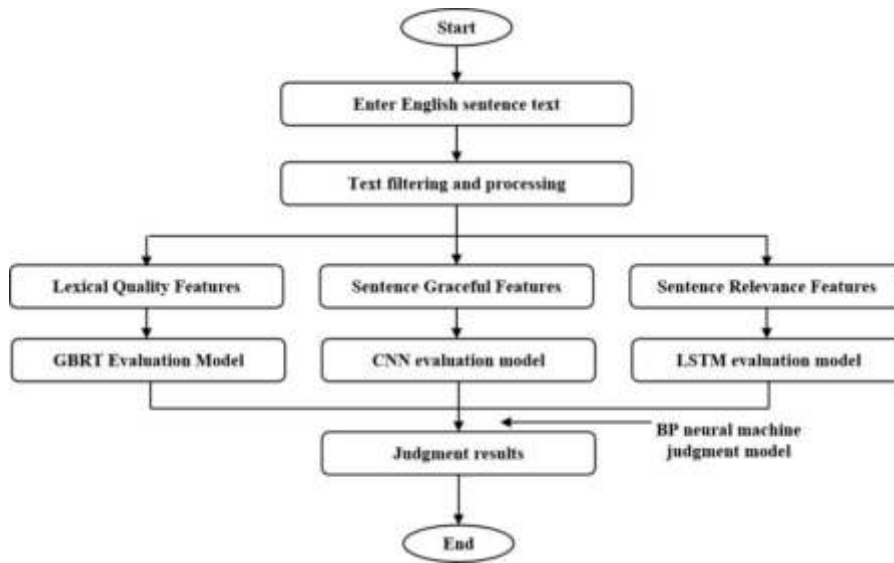


Fig. 3.2: Flow chart of comprehensive evaluation

4. Result Analysis. In the research, the English-Chinese translation questions of the final examination of College Chinese courses in a certain university are selected as the evaluation sample, and the score features, feature degree and feature assignment of English-Chinese translation questions are integrated and provided to the examination evaluation team members for understanding and familiarity. Experts in the field of Chinese translation and linguistics are invited to form an evaluation team, and the feature base assignment x is given according to the text features and feature degree, as shown in Table 4.1. And the feature assignment coefficient xc is given according to the features and difficulty of the actual translation topics. Through MAT-LAB software input Equation 3.1-3.3 and text feature weight, machine learning is carried out in a preset period to obtain the evaluation score ix . BP neural machine algorithm can carry out self-diagnosis and detection, and finally modify machine learning according to the evaluation results and manual correction, so as to meet the needs of teaching effect evaluation. If the error of the output result is less than the set , or the number of training learning exceeds the set maximum, the algorithm ends and starts the next text automatically. If the criteria are not met, the retraining is required from Equation 3.1. The software operation rules are as follows:

Output001: IF(results fit well);
 THEN(go to the next text to learn);
 Output002: IF(result coincidence is general);
 THEN(adjust the weight assignment coefficient);
 Output003: IF(the result is relatively poor);
 THEN(returns to the previous stage to adjust the feature degree and base weight assignment).

The operation of the BP neural machine learning rules mentioned above should also be based on the classification of Chinese course translation. With the help of this system, translation evaluation teachers set evaluation criteria of different levels according to the difficulty of Chinese translation at different stages of university, so as to meet the ultimate goal of examination scoring. Figure 4.1 shows the results of a specific example. By selecting test paper translation samples of 20 students in a class randomly, the comparison between the machine evaluation results and the average value of teachers' independent evaluation shows that the error range of the evaluation results of the 20 samples is -5.6%-6.7%, which is within the allowed range of translation evaluation and meets the requirements of teaching evaluation.

5. Conclusion. In the research, an automatic judgment algorithm for Chinese sentence text was proposed. The algorithm first splits and filters sentences, then extracts and optimizes them, and finally combines them

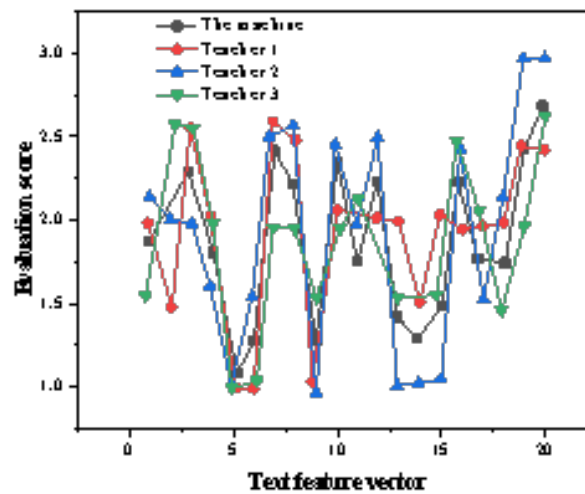


Fig. 4.1: Comparison of evaluation results

with each other. The recognition and extraction of text feature vectors, as well as the fusion and interaction of feature weights, play a crucial role in determining the correctness of the results. A basic framework for evaluating Chinese sentence translation has been designed. By comparing the results of machine validation and manual evaluation, the ETSS system has excellent performance and high evaluation reliability and accuracy. Developed a BP neural machine evaluation model, completed automatic processing of natural language and evaluation of Chinese translated sentences. By selecting test paper translation samples of 20 students in a class randomly, the comparison between the machine evaluation results and the average value of teachers' independent evaluation shows that the error range of the evaluation results of the 20 samples is -5.6%-6.7%, which is within the allowed range of translation evaluation and meets the requirements of teaching evaluation. Artificial intelligence and computer technology have promoted the development of intelligent assisted teaching methods, reducing human intervention in Chinese translation evaluation.

This article has achieved certain results in the above aspects, but there are still many shortcomings. The main problem is the quantity and quality of user feedback. It is difficult to collect large-scale user feedback in laboratory environments. At the same time, there is a large amount of noise in user feedback. This article uses corresponding quality control strategies to remove some of the noise. However, overall, larger scale user feedback experiments are needed to further confirm the experimental conclusions of this article. At the same time, there are still shortcomings in the noise processing work of this article.

In future research work, we should explore using large-scale user feedback data to confirm the conclusions of this article. The work of this article reveals the contribution of several user behavior features to automatic translation evaluation, and the processing and utilization of overall user feedback information is still quite limited. Therefore, future work should explore more diverse user behavior features and other feature selection and fusion methods on this basis for automatic translation evaluation.

REFERENCES

- [1] Chu, X., & Leng, Z. (2022). Multiuser computing offload algorithm based on mobile edge computing in the internet of things environment. *Wireless Communications and Mobile Computing*, 2022(1), 1-9.
- [2] Ren, H., Wang, J., Pang, J., Wu, L., & Shi, J. (2020). Review on machine translation post-editing of science and technology texts in china. *Open Journal of Modern Linguistics*, 10(1), 1-10.
- [3] Su, A., Jueng, J., Dupuis, L., Brooks, I., Sinha, R., & Maner, B., et al. (2021). Artificial intelligence (ai) comparison of social

- media-based patient-reported outcomes of pd-1, braf, and ctla-4 inhibitors for melanoma treatment. *Journal of Clinical Oncology*, 39(15_suppl), e21572-e21572.
- [4] Xu, Y., & Zhang, S. (2021). Research on the application of language transfer theory based on computer-aided translation software in Russian teaching. *Journal of Physics: Conference Series*, 1992(2), 022001-.
- [5] Duek, R. (2021). Project-based learning approach to marketing competencies development. *SHS Web of Conferences*, 91(2), 01004.
- [6] Bayatli, S., Kurnaz, S., Ali, A., Washington, J. N., & Tyers, F. M. (2020). Unsupervised weighting of transfer rules in rule-based machine translation using maximum-entropy approach. *Journal of Information Science and Engineering*, 36(2), 309-322.
- [7] Sapaa, B., & Turska, M. (2022). The recovered past?: deliberations on translation in the context of historical knowledge and collective memory. *Babel*, 68(1), 114-138.
- [8] Yan, Y. (2021). A bibliometric analysis visualization of translation education from 2011 to 2020. *Open Access Library Journal*, 8(8), 13.
- [9] Abidin, Z., Permata, Ahmad, I., & Rusliyawati. (2021). Effect of mono corpus quantity on statistical machine translation Indonesian – Lampung dialect of Nyo. *Journal of Physics: Conference Series*, 1751(1), 012036 (11pp).
- [10] Wu, B., He, X., Sun, Z., Chen, L., & Ye, Y. (2020). Atm: an attentive translation model for next-item recommendation. *IEEE Transactions on Industrial Informatics*, 16(3), 1448-1459.
- [11] Tian, X. (2021). Research on English translation of Chinese college students based on computer scoring system. *Journal of Physics: Conference Series*, 1992(3), 032021 (5pp).
- [12] Sun, X., & Lei, Y. (2021). Research on financial early warning of mining listed companies based on BP neural network model. *Resources Policy*, 73(2), 102223.
- [13] Zhao, G., Ding, J., Li, H., & Xu, F. (2021). Research on efficient retrieval technology of optical characteristics of image resources in digital library. *Journal of Physics Conference Series*, 1769(1), 012047.
- [14] Ruddle, R. A., Bernard, J., Lucke-Tieke, H., May, T., & Kohlhammer, J. (2021). The effect of alignment on people's ability to judge event sequence similarity. *IEEE Transactions on Visualization and Computer Graphics*, PP(99), 1-1.
- [15] Yang, H., Zhou, B., Wang, L., Wei, Q., & Zhang, R. (2021). Design and implementation of an open-source MATLAB code for GNSS/mems-INS deep integrated navigation. *Optik - International Journal for Light and Electron Optics*, 242(6), 166987.
- [16] Fan, Y., Li, Y., & Zhu, A. (2021). A few-shot learning algorithm based on attention adaptive mechanism. *Journal of Physics: Conference Series*, 1966(1), 012011-.
- [17] Greenberg, J. N., & Tan, X. (2020). Dynamic optical localization of a mobile robot using Kalman filtering-based position prediction. *IEEE/ASME Transactions on Mechatronics*, PP(99), 1-1.
- [18] Andayani, U., Efendi, S., Siregar, N., & Syahputra, M. F. (2021). Determination system for house improvement recipients in Serdang Bedagai by using clustering k-means method and *viekriterijumsko kompromisno rangiranje (vikor)*. *Journal of Physics: Conference Series*, 1830(1), 012023 (12pp).
- [19] Wang, J., Teng, F., Li, J., Zang, L., & Wang, X. (2021). Intelligent vehicle lane change trajectory control algorithm based on weight coefficient adaptive adjustment. *Advances in Mechanical Engineering*, 13(3), 168781402110033.
- [20] Wang, L., Zhang, Y., Yao, Y., Xiao, Z., & Wang, J. (2021). Gbrt-based estimation of terrestrial latent heat flux in the Haihe river basin from satellite and reanalysis datasets. *Remote Sensing*, 13(6), 1054.
- [21] Li, G., Liu, F., Sharma, A., Khalaf, O. I., Alotaibi, Y., & Alsufyani, A., et al. Research on the natural language recognition method based on cluster analysis using neural network. *Mathematical Problems in Engineering*.
- [22] Selva, Deepaa & Pelusi, Danil & Rajendran, Arunkumar & Nair, Ajay. (2021). Intelligent Network Intrusion Prevention Feature Collection and Classification Algorithms. *Algorithms*. 14. 224.
- [23] Chen, J., Liu, J., X Liu, X Xu, & Zhong, F. (2020). Decomposition of toluene with a combined plasma photolysis (CPP) reactor: influence of UV irradiation and byproduct analysis. *Plasma Chemistry and Plasma Processing*.
- [24] Huang, R., Zhang, S., Zhang, W., Yang, X. Progress of zinc oxide-based nanocomposites in the textile industry, *IET Collaborative Intelligent Manufacturing*, 2021, 3(3), pp. 281-289.
- [25] Kaabar, M., Kalvandi, V., Eghbali, N., Samei, M., Siri, Z. & Martínez, F. (2021). A Generalized ML-Hyers-Ulam Stability of Quadratic Fractional Integral Equation. *Nonlinear Engineering*, 10(1), 414-427.

Edited by: B. Nagaraj M.E

Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing

Received: Dec 20, 2023

Accepted: Feb 27, 2024



RESEARCH ON INTELLIGENT BUILDING INTEGRATED CABLING SYSTEM BASED ON INTERNET OF THINGS AND DEEP LEARNING

RONG ZHOU*

Abstract. In order to solve the problem that the traditional integrated wiring system has many inconveniences in management, which affects the stability and efficient operation of the entire system, an intelligent building energy management control system in the Internet of Things era is proposed. In terms of hardware, design intelligent building switches, and in terms of software, plan the IP address of intelligent building wiring based on the Internet of Things, formulate comprehensive backbone wiring layout structure of intelligent building, set up auxiliary power distribution horizontal lines, and connect intelligent building wiring data, so as to realize intelligent building integrated wiring. On this basis, the architecture design of the wiring assistant design system is carried out, and the artificial intelligence building integrated wiring assistant system is realized with the help of the Net platform, and the system test is carried out. The experimental results show that the PM10 concentration parameters and the energy consumption simulation curve are highly consistent with the actual value curve. According to the calculation of the data volume points, the coincidence degrees can reach 96.4% and 98.9%, respectively. The experimental results show that the designed wiring system is superior to the traditional wiring system in terms of energy saving, and has certain reference value.

Key words: IoT technology, intelligent building, system integration, artificial intelligence, general wiring

1. Introduction. Cabling system refers to the network transmission, design time Attention should be paid to the connection of various devices, including voice, data processing equipment. Adopt An integrated wiring system, which connects equipment to the interior and exterior of the same building, is The general idea of the current design. The Internet of Things and digital technologies are developing rapidly in the current environment. In the good state of the momentum of the Internet of things, the modern construction industry is also gradually inclined to intelligent development. Simply put, intelligent building is a combination of information technology and modern building technology, with a new way of display to provide a service platform, so that people become more comfortable in the use of the process, instead of the traditional method, so that the display is more comprehensive. Following the emergence of computer, Internet and mobile communication network information control technology, the rise and development of Internet of Things technology has promoted the third scientific and technological revolution, which is also an important part of information technology under the new situation [1]. One after another, all parts of the world have taken the lead in launching research plans for the strategic development of the Internet of Things. Under the influence of the global Internet of Things trend, the Internet of Things was written into the Chinese government work report for the first time in 2010, and it was officially listed as one of the five emerging development strategies of the country, and was given a high degree of attention and policy support [2,3]. In recent years, the Internet of Things technology has gradually become a new hot spot for development, and it also presents a huge development prospect, penetrating into all aspects of people's lives. Traditional wiring techniques often lack reliability and energy saving in terms of transmission Also not satisfactory, can not meet the current needs of smart buildings, can not replace the end equipment, and has high maintenance and renewal costs, the overall is also very ugly.

In the field of construction, usually shallow geothermal energy needs to be collected and exchanged by ground source heat pump system before it can be utilized. After years of research and practice, ground source heat pump technology has proved that this technology is suitable for the requirements of sustainable development, and has the advantages of high efficiency, energy saving and environmental protection.

Traditional intelligent buildings are based on integrated wiring and use computer networks as bridges. Most of them use an extensive three-layer structure, which is the field control layer, the automatic control layer and

*Shaanxi vocational and Technical College, Xi'an, Shaanxi 710038, China (RongZhou56@163.com)

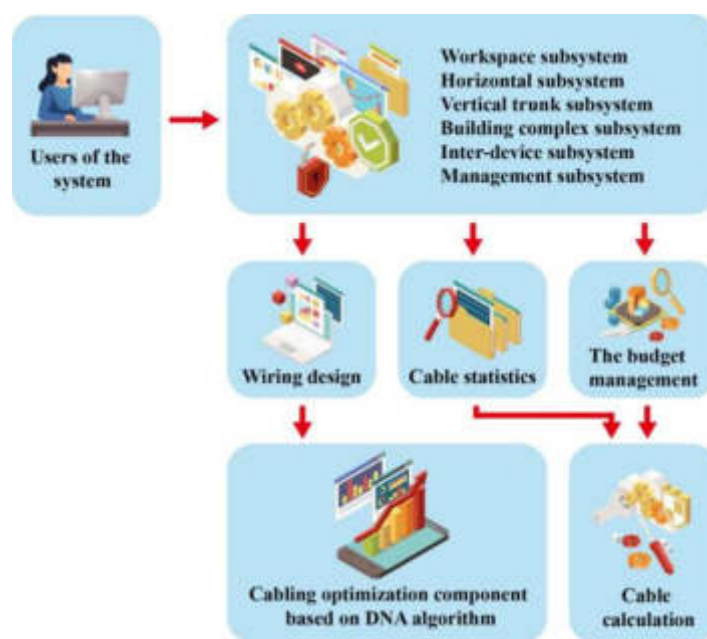


Fig. 1.1: System integration architecture diagram of integrated wiring of intelligent building

the top management layer. Various subsystems in the building (elevator, water supply and drainage, HVAC, power distribution, intelligent lighting, security, etc.) are configured through various communication protocols, and the system integration method of a unified protocol is often used to realize the integration of various systems in the building. Comprehensive management and centralized monitoring of various equipment and subsystems. The traditional integration method is easy to cause problems such as difficult coordination and operation between subsystems, heavy configuration workload, poor openness, and poor flexibility [4-5], so it is not conducive to the development of building intelligence. The object in the Internet of Things is the basic unit. If this concept is applied to the building, the electrical equipment can be regarded as a basic information unit and is endowed with "wisdom", then the network is connected to the Internet of Things system according to the unified communication protocol. All information perceived by the underlying device can be received. Therefore, the Internet of Things technology solves the existing drawbacks from the bottom device side, and the information exchange and information communication based on the Internet of Things technology become much easier. Therefore, to realize system integration of intelligent buildings, the overall structure will change, and the emergence of a new integrated system architecture based on the Internet of Things is an inevitable trend. Figure 1.1 is a system integration architecture diagram of integrated wiring of intelligent buildings. The traditional cabling method makes the reliability of each line is poor, and the cost is high Design of intelligent building cabling system based on Internet of Things. In terms of hardware, I designed intelligent building switches; in terms of software, I planned intelligent building wiring IP address based on the Internet of Things. Formulate the intelligent building comprehensive trunk wiring layout structure, set the auxiliary distribution horizontal line, connect the intelligent building wiring data, so as to realize the intelligent building comprehensive wiring.

2. Literature review. Qian, H. et al. proposed that in recent years, the application field of the Internet of Things has become more and more extensive, and it has been widely used in fields such as smart home, smart transportation, agriculture, environmental protection, industry, medical and health, etc., and has achieved good results. Demonstration effect [6]. According to the research report by Zhu, Z. M. et al., during the 12th and 13th Five-Year Plan period, the Internet of Things technology has developed rapidly, has received strong support from national policies, and has become the driving force for the economic development of China's key industries. Under the influence of the global trend of smart earth and smart city, China has also put forward

the slogans of "perceive China" and "smart city" based on the Internet of Things [7]. Xiao, B. et al. proposed that, as the most basic unit of a smart city, smart buildings can effectively promote the development of smart cities with the help of the Internet of Things [8]. Liu, Z. and others believe that the Internet of Things makes the various subsystems of intelligent buildings "smart", and each system can freely increase or decrease the corresponding functions and services according to the needs of users, turning "intelligent buildings" into "Smart building" realizes the integration of "management, control and operation" to build a smart city [9].

X, He. et al. pointed out that a large number of sensors are installed in various subsystems in the building, such as lighting, HVAC, and security systems, and the data measured by the sensors constitute the information basis. The Internet of Things technology can realize the collection and transmission of data., computing processing, and the structure of the Internet of Things system is shown in Figure 2.1 [10]. According to the research of Qin, N. et al., on the one hand, the comprehensive perception and intelligent analysis of Internet of Things technology provide technical support for intelligent buildings. New intelligent buildings should be open, equipment can self-organize, and the system should be flat to meet the needs of different manufacturers[11]. On the other hand, with the continuous maturity of the Internet of Things, big data, and artificial intelligence technologies, the intelligent era of the Internet of Everything and the integration of the human-machine-object ternary world have become inevitable, "connection + big data intelligence + personalization" "Service" will become the basic paradigm of building intelligence. Tong, Y. U. et al. proposed that the application of the Internet of Things in the field of intelligent buildings is very limited, and generally there are only four aspects: intelligent monitoring, intelligent security, smart home and energy saving and emission reduction. For example, various sensors are installed in home equipment, information is transmitted through the network, and users are monitored through B/S access mode [12]. Therefore, Ren, L. et al. pointed out that through the perception layer device, the building can fully perceive the data information generated by people/environment, and store, analyze and learn, and can think independently [13]. Abdurraheem, AS et al. believe that people-oriented, in addition to the standards of building equipment itself, pay more attention to people's needs; at the same time, it solves the problems existing in traditional intelligent buildings, and a new type of intelligent building architecture that adapts to the Internet of Things era emerges as the times require [14]. Guo, L. et al. proposed that IoT buildings apply IoT technology to realize comprehensive perception of various physical parameters in buildings. Through heterogeneous network fusion, information aggregation, decision-making diagnosis, online control, big data analysis and other means, Form an integrated service management system from the bottom equipment to the upper application, and realize the comprehensive optimization management of energy saving, comfort, safety, health and other objectives in the whole life process of the building [15].

3. Research method.

3.1. Design and implementation of artificial intelligence DNA algorithm based on Internet of Things. The design idea of integrated wiring system is mainly structure and modularization, and adopt Hierarchical star topology is used for integrated wiring of the whole building. From the machine room to The structure of each floor adopts the star topology, the wiring cabinet of each floor and each work Regional information points are no exception. The data communication and signal transmission of the 3A system of a smart building requires a complex wiring system to provide transmission support. The signal cables are routed from the trenches. Interleaving can negatively affect network performance and signaling [16]. The solution to cable crossover is to perform layered wiring: if crossover cables occur, route them to different layers of the trunking, while ensuring that cables on the same layer do not cross. In order to reduce the amount of wiring construction, it is required to control the number of wiring layers to a minimum. The hierarchical problem can be mapped to the vertex coloring problem of the graph. Vertices of the same color can be routed to the same layer, and vertices of different colors need to be routed to different layers. Compared with the traditional computing mode, the artificial intelligence DNA sorting algorithm based on the Internet of Things has the advantages of fast computing speed, low power consumption, high storage capacity and high degree of parallelism. At present, the fastest supercomputer operates at an order of magnitude of 10¹² operations per second, while the speed of DNA computers can reach 10¹⁴ operations; the biggest problem of traditional computers is power consumption, and the power consumption of DNA computers is only one billionth of that of traditional computers; the information stored in one gram of DNA is equivalent to 2.5 million optical discs; traditional computers have the characteristics of seriality, while DNA has native support for parallel

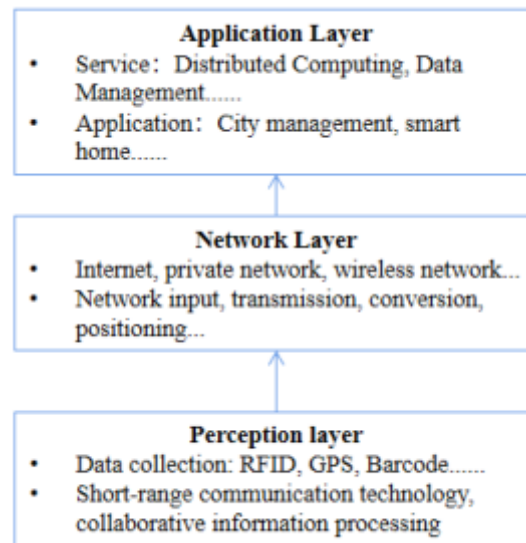


Fig. 2.1: IoT Architecture

computing [17].

Firstly, the basic model composition of artificial intelligence DNA algorithm is described. The first problem that needs to be solved is the coding problem. The coding in the artificial intelligence DNA model is realized by single-stranded or double-stranded DNA molecules. DNA molecular strand is a storage complex, which consists of storage strand and sticking strand. A storage chain can be formed by concatenating n heterogeneous sub-chains, each sub-chain contains m bases, and each pasting chain also contains m bases. Firstly, the basic model composition of artificial intelligence DNA algorithm is described. According to the principle of base complementarity, it can be determined that there is a complementary relationship between the pasted chain and a certain sub-chain in the storage chain. Therefore, it can be agreed that when a single chain exists in the storage complex, it means 0, and if it is a double chain, then Represents 1[18]. On this basis, four basic operations of the storage chain are defined:

- 1) Set: Set the non-zero storage location to "1", represented by $\text{Set}(T, i)$;
- 2) Clear: change the storage location from non-zero to "0", represented by $\text{Clear}(T, i)$;
- 3) Merge: Use the ligation reaction to merge the two DNA single-stranded or double-stranded DNA, that is, merge the two storages into one, expressed as $T=T_1 \cup T_2$;
- 4) Decomposition: Decompose the storage T represented by the single-stranded or double-stranded DNA molecule into two sets $+(T, i)$ and $-(T, i)$ as needed, where $+(T, i)$ represents the storage bit A combination of bit strings of 1, similarly $-(T, i)$ represents the set of bit strings of 0 [19].

The vertex coloring algorithm for solving the graph can be expressed as the algorithm of deleting uncolored points and deleting adjacent same-color points, which are described in Table 3.1 and Table 3.2 respectively:

In the above code, r and t are the subscripts of the two vertices that make up the edge e_i , respectively, and Δ represents the deletion operation. After the above iterative operations, the DNA strand in the test tube T_0 is a feasible coloring scheme for the graph G , and the desired result is obtained after decoding[20]. The core of this computational model is to use a library of magnetic bead probes with biomarkers to implement continuous separation of non-solutions in the initial solution space, and finally find the target solution, where the initial solution space is composed of library chains (that is, DNA representing all possible coloring schemes) sequence) and probe library strands representing the structural information of the graph.

Table 3.1: Remove unshaded points algorithm

Remove unshaded vertices pseudocode

```

GranhColoring(T0,n,m,k)
  For t←1 to n do
    Separate+(T0,(i-1)*k+1) and-(T0,(i-1)*k+1)
    T0←+(T0,(i-1)*k+1)
    T1←+(T0,(i-1)*k+1)
    For j←2 to k do
      Separate+(T1,(i-1)*k+j) and-(T1,(i-1)*k+j)
      T0←Merge(T0+(T1,(i-1)*k+j))
      T1←+(T1,(i-1)*k+j)
    End
  abando T1
End

```

Table 3.2: Delete adjacent same color point algorithm

Pseudocode for deleting adjacent points of the same color

```

For i← 1 to m do
  For j←1 to k do
    Separate+(T0,(r-1)*k+j)and-(T1,(r-1)*k+j)
    T0←-(T0,(r-1)*k+j)
    T1←-(T1,(r-1)*k+j)
    Separate+(T1,(t-1)*k+j)and-(T1,(t-1)*k+j)
    T0←Merge(T0+(T1,(t-1)*k+j))
    abando+(T1,(t-1)*k+j)
  End
End

```

3.2. Design of intelligent building integrated wiring system based on IoT artificial intelligence DNA algorithm. When all kinds of new energy are introduced into intelligent buildings to be used, the automatic operation and management of all kinds of application systems will naturally be included in the intelligent building equipment management system. This not only increases the content of building equipment monitoring system, but also expands the scope of energy management services. The intelligent building work area system consists of information sockets, adapters, and cables connecting the user terminal equipment to the sockets. The terminal office environment of the intelligent building is provided directly to the end user, and its design is relatively simple. When the user's network usage requirements are not accurate, an independent work area can be estimated based on the area of 5-10 square meters. The second core problem of workspace design is to count the number of information points. A work area in an ordinary office area is usually equipped with 2-3 information points. For work points with special needs (such as setting up services such as server, fax, video, network printing, etc.), 3-5 dedicated information points can be added. In terms of transmission cable requirements, conventional office areas can lay 10-100M twisted pair cables, while for business or technology development office areas with high bandwidth requirements, fiber optic information points that support more than 100M can be laid. In terms of socket design, the information points are mainly composed of standard RJ45 sockets, and the selection of other sockets must comply with EIA/TIA standards. In scenarios with special requirements, various types of adapters can be selected for connection according to needs [21]. The workflow of the workspace subsystem design is as follows:

- 1) Determination of design level and work area information points: Determine the number of information

points in the work area according to the design level selected by the user. The design of the number of information points N must consider the future development needs, and the number of information sockets corresponding to different design levels can refer to the national standards for integrated wiring and related manuals [22];

2) Calculation of working area: Calculate the area of each working area according to the building plan, and count the total area S of the working area of the building [23];

3) Calibration of the number of sockets and their positions: First determine the area P of the work area. If there is no special requirement, it is usually calculated according to $P = 5 \sim 10m^2$, then the number of sockets is $M = (S \div P) \times N$, where S is the total work area, P is the area of a single workspace, and N is the number of information points in a single workspace [24];

4) Calculation of socket type and the number of associated devices: the type of socket can be surface-mounted or embedded[25]. New buildings usually use the embedded method; the sockets installed on the floor have two types: fixed type and movable type. User needs and costs to choose. Associated equipment includes bottom boxes, covers, panels, etc. The type and quantity of sockets and related connectors can be determined according to user needs and architectural drawings.

4. Result analysis.

4.1. Realization of intelligent building integrated wiring system based on IoT. In order to realize the integrated wiring of intelligent building, it is necessary to distribute the integrated wiring system To the various parts of the smart building. Design cabling for various types of hardware Standard information sockets for equipment need to be provided by capital engineering and open systems. Practical design of integrated wiring system and The installation will be on the building , structure and other industries make many requirements when designing The comprehensiveness of the cabling system must be considered to meet the needs of modern intelligent buildings Demand. Thus, the information transmission between various automatic systems is stable. The system is implemented according to the three-layer structure. The presentation layer, in the form of a local client, provides users with a visual cable routing auxiliary interface. This section focuses on the implementation of the business logic layer. The business logic layer includes core parts such as cable routing optimization design module and database access module, which are encapsulated in the form of reusable components. The database access component provides the encapsulation of the database Create, Retrieve, Update and Delete operations. In terms of coding implementation, the database access operation is divided into three steps: first, create and obtain a business logic object; then create a persistent object related to database access and a business class for storing data through the business object; finally, the business object Call the methods of the persistent object to perform operations such as searching, inserting, deleting, and updating the database. On the .Net platform, database access operations are designed to database-driven loading, connecting to the database, and database operations.

4.2. Test results of integrated wiring system of intelligent building based on Internet of Things. In order to ensure the correctness, reliability and safety of the system, a scientific test strategy must be formulated in the test link. For the test of this system, the project team has formulated the following test links and test strategies:

(1) Functionality and robustness testing: Whether the software system meets user requirements is the primary indicator that needs to be confirmed during software delivery. The functional test is carried out according to the software requirements analysis specification, and the method of black-box testing is used to assess whether the system functional modules meet the requirements of the requirements specification, and whether they can be expected under the given input; Robustness test: robustness test Used to ensure the robustness and stability of the system. During the test, illegal values are entered manually to judge the response of the system. The system should provide feedback for wrong input, and it will not crash; (2) User interface test and performance test: In order to facilitate the use of users and improve work efficiency, the system should have a friendly human-machine interface, which is convenient, intuitive and beautiful and concise; performance test: this system belongs to a typical C/S mode application, the client side involves editing the building structure diagram, and is used to realize the wiring and routing design; the server side mainly involves database operation. This system does not have large server-side load and pressure requirements, and the client

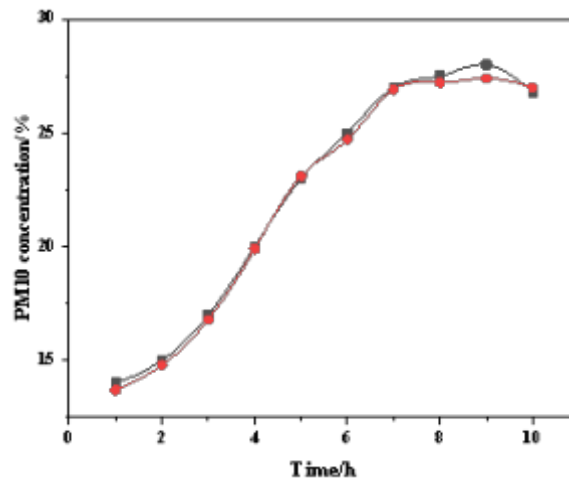


Fig. 4.1: Contrast curve of PM10 concentration in experimental room of passive experimental building

hardware configuration can currently meet the requirements. Therefore, this system can theoretically meet the performance test requirements in most environmental scenarios. The test results for system functionality and robustness are shown in Figure 4.1. In the experiment, the actual PM10 concentration curve was obtained through traditional measurement methods, and the measurement curve was transmitted back to the system in this paper through the artificial intelligence building wiring system based on the Internet of Things. It can be seen from Figure 4.1 that the PM10 concentration parameters obtained by the experimental measurement are highly consistent with the actual value curve, and the degree of agreement can reach 96.4% according to the calculation of the data volume points, indicating that the system has the conditions for measuring various parameters and verifies its functionality; the actual saturation value The difference from the measured saturation value of the system is 6.4%, which verifies that the system has good robustness.

For the user interface and its performance, the energy consumption prediction interface is called to test the prediction accuracy and performance of the system. The obtained prediction model curve and the actual energy consumption curve are compared as shown in Figure 4.2. It can be seen from Figure 4.2 that the energy consumption prediction interface The predicted energy consumption simulation curve coincides with the actual energy consumption curve measured in the experiment at multiple sampling points. The calculated curve fitting degree is as high as 98.9%, which verifies the accuracy and efficiency of the interface of the artificial intelligence building wiring system.

By analyzing the comparison curve between the above prediction model and the measured value, the following conclusions can be drawn: In the module testing process, the project team conducted a black-box test on the remaining functional modules of the integrated wiring auxiliary system. Description of functional requirements. Before functional testing, the system conducts unit testing by means of code walk-through to avoid coding errors. In addition, user interface testing was carried out to ensure that the user interface is friendly, straightforward view. In order to ensure the good compatibility of the client, a platform compatibility test is also carried out, and the test results show that, thanks to .Net The platform naturally supports the Windows platform, and the system can run stably on the Windows series operating system platform. At present, the integrated wiring auxiliary system has been in trial operation in the laboratory, showing good performance and effectively improving the efficiency of design work. A character is designed Integrated environment of intelligent building wiring system, after design, the system experiment, According to the experimental results, the intelligent building wiring system designed in this paper is based on the Internet of Things The system is superior to the traditional wiring method in energy saving and has been popularized to some extent Meaning.

5. Conclusion. Up to now, the concept of the Internet of Things and the idea of "smart energy" have made clear the future development direction of intelligent buildings, especially in the aspect of energy management. It

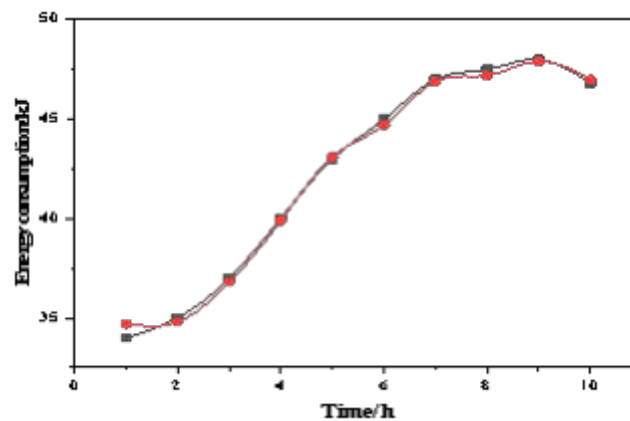


Fig. 4.2: Comparison of the predicted model curve and the actual energy consumption curve of the energy consumption prediction interface

will break through the shackles of single buildings or independent parks, and stride forward to the broader and more integrated field of "smart city" energy management. Aiming at the problems existing in the traditional intelligent building system architecture and the development of the Internet of Things technology in intelligent buildings, this paper improves the system architecture of the traditional intelligent building, and proposes an artificial intelligence building control system based on the Internet of Things technology. Comprehensive perception of various physical parameters in the system, information fusion and aggregation of heterogeneous data, through service decision-making, system diagnosis, online monitoring, big data analysis and other means, to form an integrated service management system from the bottom equipment to the upper application., to achieve multi-objective optimization and comprehensive optimization management of energy saving, comfort, safety and health in the whole life process of the building.

This paper has completed the design and implementation of the IoT artificial intelligence building system, and has achieved certain results by applying it to the passive experimental building, but the research on the energy saving of the IoT in buildings needs to be further deepened. In the following work, the factors affecting building energy consumption, such as the environment, building envelope, building shading, indoor heat gain, etc., should be analyzed first; then the energy consumption correlation model should be established, and the energy consumption data of building monitoring should be analyzed by regression analysis. Or neural network technology can establish energy consumption model, and use big data analysis technology to predict the energy consumption of the entire building. Finally, the system is optimized considering the comfort requirements of green buildings. Under the background of advocating rational use of traditional energy and active exploitation of green renewable energy, intelligent buildings have added new contents: introducing new energy application systems such as solar energy, geothermal energy and wind energy into intelligent buildings to reduce the consumption of traditional energy in buildings; At the same time, it is integrated into the construction equipment management system to make the application of new energy more transparent and reasonable.

REFERENCES

- [1] Kong, L., & Ma, B. (2020). Intelligent manufacturing model of construction industry based on internet of things technology. *The International Journal of Advanced Manufacturing Technology*, 107(1), 35-39.
- [2] Wang, Y., & Ku, J. (2021). Research on collaborative innovation platform of internet of things industry based on data mining technology. *Journal of Physics: Conference Series*, 1881(4), 042072-042078.
- [3] Yuchao, L., Qiu, W., Xiao, L., Yue, Y., & Zhixin, F. (2021). Research on supply and demand forecast of regional integrated energy system based on computer internet of things technology. *Journal of Physics: Conference Series*, 1915(2), 022035-022042.
- [4] Zhao, D., Tai, X., & Ma, Z. (2020). Research on integrated deployment method of digital earth thematic application model. *IOP Conference Series: Earth and Environmental Science*, 502(1), 012003.

- [5] J Ma, Yu, L., Sun, W., Dong, S., & C Yao. (2021). Investigation and evaluation of solid-state marx pulse generator based on 3-d busbar. *IEEE Transactions on Plasma Science*, 49(5), 1597-1604.
- [6] Qian, H. (2021). Optimization of intelligent management and monitoring system of sports training hall based on internet of things. *Wireless Communications and Mobile Computing*, 2021(2), 1-11.
- [7] Zhu, Z. M., Xu, F. Q., & Gao, X. (2020). Research on school intelligent classroom management system based on internet of things. *Procedia Computer Science*, 166(1), 144-149.
- [8] Xiao, B., Yang, K., & Liang, H. (2021). Research on integrated application system of internet of things in oil depot. *Journal of Physics: Conference Series*, 1972(1), 012006-012013.
- [9] Liu, Z., Zheng, X., Xiao, Z., Bao, J., Shuang, L., & Sheng, F., et al. (2020). Research on integrated algorithm based on convolutional neural network for rice disease identification. *Journal of Physics: Conference Series*, 1646(1), 12-18.
- [10] X He. (2021). Design and application of building intelligent integrated wiring system. *Journal of Physics Conference Series*, 1802(3), 032016-032017.
- [11] Qin, N., Tang, Z., & Wang, Y. (2021). The data sharing platform system of electrical main wiring based on the results of 3d digital gim. *Journal of Physics: Conference Series*, 1952(3), 032020 -032025.
- [12] Tong, Y. U., Mei-De, X. U., Zi-Han, Y. U., & Zhang, T. Q. (2021). Design of air quality monitoring system based on light scattering sensor. *IOP Conference Series: Earth and Environmental Science*, 647(1), 012196 -012203.
- [13] Ren, L., Zhai, X., Yang, Y., & Xu, J. (2020). Design of horticultural wireless intelligent maintenance system based on stm32 and android. *IOP Conference Series: Earth and Environmental Science*, 474(3), 032016-032019 .
- [14] Abdurraheem, A. S., Salih, A. A., Abdulla, A. I., Sadeeq, M., & Saeed, R. A. (2020). Home automation system based on iot. *Technology Reports of Kansai University*, 62(5), 2453-2454.
- [15] Guo, L., Chen, F., Xiao, L., & Hu, Y. (2020). The gas detection application of deep well production based on wireless sensor network. *IOP Conference Series: Materials Science and Engineering*, 719(1), 012049-012054.
- [16] Jin, C., Cao, Z., Liu, X., & Jin, H. (2020). Research on key technologies of intelligent energy gateway based on fog computing technology. *IOP Conference Series: Earth and Environmental Science*, 512(1), 123-128.
- [17] Wang, X., & Chen, X. (2020). Research on new intelligent building electrical energy saving technology based on internet of things technology. *IOP Conference Series: Materials Science and Engineering*, 782(3), 032023.
- [18] Chen, G., Zhongan, D., Shian, Z., Wuxiao, C., & Kunrong, Y. (2020). Research on the construction of intelligent meter reading system based on energy metering integrated acquisition technology. *IOP Conference Series: Earth and Environmental Science*, 440(3), 032026-032037.
- [19] Zhang, X., & Qu, Y. (2021). Research on energy acquisition, monitoring and analysis platform of integrated energy system based on nb-iot. *IOP Conference Series: Earth and Environmental Science*, 769(4), 042037-042042.
- [20] Yu, J., Fu, J., Lu, Y., Fu, S., Huang, D., & Zhang, X. A. (2020). Research on aircraft/engine integrated method of civil aircraft products development oriented to system engineering. *IOP Conference Series: Earth and Environmental Science*, 587(1), 012027-012036.
- [21] Kaabar, M., Kalvandi, V., Eghbali, N., Samei, M., Siri, Z. & Martínez, F. (2021). A Generalized ML-Hyers-Ulam Stability of Quadratic Fractional Integral Equation. *Nonlinear Engineering*, 10(1), 414-427.
- [22] R. Huang, X. Yang, "The application of TiO2 and noble metal nanomaterials in tele materials," *Journal of Ceramic Processing Research*, vol. 23, no. 2, pp. 213–220, 2022.
- [23] Liu, Xin and Ahmadi, Zahra. 'H 2O and H 2S Adsorption by Assistance of a Heterogeneous Carbon-boron-nitrogen Nanocage: Computational Study'. 1 Jan. 2022 : 185 – 193.
- [24] Selva, Deepaa & Pelusi, Danil & Rajendran, Arunkumar & Nair, Ajay. (2021). Intelligent Network Intrusion Prevention Feature Collection and Classification Algorithms. *Algorithms*. 14. 224.
- [25] Sharma, K., & Chaurasia, B. K. (2015). Trust Based Location Finding Mechanism in VANET Using DST. *Fifth International Conference on Communication Systems & Network Technologies* (pp.763-766). IEEE.

Edited by: B. Nagaraj M.E

Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing

Received: Dec 20, 2023

Accepted: Feb 27, 2024



THE APPLICATION OF INTELLIGENT ROBOTS AND DEEP LEARNING IN THE CONSTRUCTION MANAGEMENT PLATFORM SYSTEM OF CONSTRUCTION ENGINEERING

YANDONG ZHOU*

Abstract. In order to solve the problem of duplicate data entry between construction management platform systems in construction engineering, the author proposes to apply RPA intelligent process robots to replace manual data collection, operation, entry, and verification. The design of the system is divided into overall architecture, instruction program loading process, human-machine interaction system level services, and other levels. An end-to-end procedural instruction transmission control method is adopted, establish a low-level control command output module for the online calibration system of the flight path, using a basic service architecture system, implement human-machine interactive control of the online calibration system for flight paths on the B/S architecture system. Build a record controller module for the inspection trajectory correction of RPA intelligent process robots, and perform feedback control during the trajectory correction process in LOG-CONTROL-BLOCK, using the RPA feedback correction algorithm, achieve adaptive correction and error feedback tracking of the inspection trajectory of RPA intelligent process robots. Implement calibration system development and design in an integrated DSP (Digital Signal Processing) information processing platform. The experimental results show that good economic benefits have been achieved through application, and the problem of duplicate data entry between the employer's IFS system and the ENPOWER document management system has been solved, greatly reducing error rates and personnel costs. It can replace manual data collection, entry, verification, and business operations; It has the characteristics of low error rate, low cost, high accuracy, compliance, and 24/7 standby; In 2021, 108000 yuan was saved, in 2022, 432000 yuan was saved, and in 2023, 432000 yuan was saved, demonstrating the labor hour cost savings achieved by utilizing RPA intelligent process robots.

Key words: Intelligent robots, Construction engineering, Construction management platform

1. Introduction. Traditional project management work is greatly influenced by the professional qualities and abilities of management personnel. If the management knowledge reserve of management personnel is not rich and lacks project management experience, it may be difficult to implement job responsibilities in engineering practice, increase the quality and safety risks of construction projects, and threaten the life safety of construction personnel. In addition, the construction site environment is relatively complex and diverse, and the cultural level of construction personnel is generally low, they usually lack awareness of safe and civilized construction, do not pay attention to the standardization of operations, and fail to use safety protection facilities reasonably, when a sudden unexpected event occurs, one is at a loss and loses the best opportunity to escape, not only will construction be delayed, but once casualties occur, their families will be in a state of disaster[1]. There are numerous high-rise buildings in the city, and the amount of construction work is larger and the complexity is higher. Traditional project management models are no longer applicable, in order to avoid safety accidents, strengthen the construction process and site management, and use intelligent technology is imperative[2].

With the advent of the information age, the application of information technology in the construction industry has become more and more common, with the help of the Internet of Things, BIM, cloud computing, Big data, artificial intelligence and other technologies, smart site systems are built, develop safety management strategies based on the characteristics of construction projects and the specific situation of the construction site, and make timely adjustments based on the relevant information obtained, storing information in the cloud breaks the limitations of outdated and outdated management, and aligns with the full lifecycle management of buildings, the smart construction site system consists of multiple layers, including perception layer, transmission layer, processing layer, etc., it can process and analyze the collected project information at corresponding levels, providing project management personnel with strong data support for decision-making[3]. Relying on the smart

*Henan Technical College Of Construction, Zhengzhou, Henan, 450000, China (YandongZhou9@163.com)



Fig. 1.1: Intelligent construction site personnel management system

construction site system, we implement smart supervision, coordination, and training to ensure the organic coordination of progress, quality, safety, and environmental management. By practicing the concept of safe and civilized construction, the project management system has been comprehensively reformed.

At present, the smart construction site system has become an important auxiliary tool for construction project management. During the construction phase of the project, the smart construction site system has played a significant role. The application process is to use sensors and video monitoring devices to monitor the operation of construction machinery, transfer the attendance information of construction personnel to the smart construction site system, and the personnel located in the management and command center will combine the video and image information to order the quality and safety responsible person to promptly rectify the hidden dangers and reduce the probability of safety accidents[4]. Ensuring the personal safety and vital interests of construction personnel is the fundamental goal of construction project management, given the low professional quality and poor safety protection awareness of some construction personnel, utilize the smart construction site system for training, assessment, salary management, and other aspects, reflecting the concept of smart management. One is to specially build a database to store the basic information of all construction personnel, and issue smart cards to construction personnel, smart cards must be used for entering and exiting the construction site, as well as for construction and consumption activities within the site, it is strictly prohibited to use others' smart cards under false names[5]. The second is for construction personnel to enter the safety education and training section of the smart construction site system to learn safety knowledge, learn about the causes of safety accidents and their own job responsibilities, participate in safety knowledge assessments online, and obtain certificates to participate in construction after passing the assessment. The third is to distribute exclusive safety helmets to construction personnel, which can automatically locate the positions of construction personnel and record the operation time as a basis for attendance, when construction personnel take off their safety helmets for a period of time or suffer severe impacts, the safety helmets will emit an alarm signal. The fourth is that the smart construction site system can calculate the wages payable based on the attendance status and salary standards of construction personnel, ensuring that the interests of construction personnel are not infringed, as shown in Figure 1.1.

With the development of artificial intelligence control technology, the types and complexity of robots have increased, and robots have been applied in various fields. In auditing, artificial intelligence robots are used to process bills and reports during the auditing process, improving the intelligence level of auditing. In the process of auditing robot operations, due to factors such as the irregularity and uncertainty of the robot's job interface, the robot's trajectory tracking and control ability is not good. Therefore, it is necessary to conduct online calibration of the audit robot's inspection trajectory, combined with environmental parameter recognition and obstacle avoidance processing, to improve the robot's inspection and control performance. Studying the optimization design method of online calibration system for audit robot inspection trajectory is of great significance in improving the inspection ability of audit robots.

2. Literature Review. At present, there are multiple separate information systems in the project, such as NC (financial shared software), OA (office automation), ENPOWER (nuclear power multi project management

system), IFS3. o (construction management information system), and other heterogeneous databases. To develop various API interfaces, WCF (Windows communication development platform), Web Service (remote call technology across programming languages and operating systems) to rebuild and achieve automation processing, it is not only costly, but also costly, And the development cycle is long. At this point, RPA technology has a powerful advantage in connecting these system interfaces. When using RPA intelligent process robots to simulate manual operations, they do not need to modify the original system, but directly imitate human behavior for operation, with good confidentiality. Especially in the processes of data extraction, input, filling out forms, and extracting structured and semi-structured data from various systems, RPA robots can be developed to achieve automation operations without making any program changes to the original system. At this time, RPA technology has powerful advantages in connecting these system interfaces. The RPA intelligent process robot does not transform the original system when simulating the manual operation, but directly imitates human behavior, with good confidentiality. In particular, RPA robots can be developed to extract data, input, filling in forms, and extracting structured and semi-structured data from various systems, and automatic operation can be realized without any program changes to the original system.

The construction industry is one of the pillar industries in China, playing a very important role in the development of the national economy and the employment of the people. In 2016, the national construction industry enterprises (referring to qualified general contracting and professional contracting construction enterprises, excluding labor subcontracting construction enterprises) had a total output value of over 19 trillion yuan, an increase of 7.09% compared to last year, the proportion of its increment to the national GDP is 6.66%, and the number of people employed in the entire construction industry exceeds 50 million, accounting for 6.68% of the total number of employed people in society. However, China's construction industry still has outdated technology and extensive management, resulting in serious waste of resources, building an ordinary residential building can result in up to 40 in 2016, the profit margin of China's construction industry is still very low. At present, the industrialization level of China's construction industry is very low, and many construction methods, processes, and skills in the construction process heavily rely on the on-site construction operations of construction workers, which are greatly influenced by human factors and the environment, this is an important factor contributing to the unfavorable situation of low quality and low profits in China's construction industry. In current construction, although a large number of mechanical equipment have been involved, more processes still rely on manual work, which is inefficient and time-consuming. On the other hand, the issue of worker health and safety is also an important obstacle to the development of China's construction industry. Construction workers are always exposed to dangerous and deadly external environments. In 2015, 43 construction workers in the UK died at work, accounting for 30% of the total number of deaths in various industries throughout the year. In the same year, 937 construction workers in the United States were fatally injured during construction, accounting for 19.37% of the total number of fatal work-related injuries in the country. In China, from 1997 to 2014, there were an average of over 2500 fatal accidents occurring at construction sites every year. Based on global statistical data, the average casualty rate of the construction industry is 2-3 times that of other industries. Despite improvements in recent years, the casualty rate of construction workers is still very high. Moreover, the working environment of construction workers is extremely harsh, with dust and loud noise, which seriously affect the physical and mental health of on-site personnel in engineering projects, bringing health hazards to them. The large number of workers and imperfect management systems on the construction site have also brought many safety hazards, seriously restricting the healthy development of the construction industry.

Zhou.L believe that intelligent manufacturing is the theme and main direction of the development strategy of "Made in China 2025", and the application of industrial robots is an important direction of intelligent manufacturing. In the coming years, industrial robots will be widely used in various enterprises, which will inevitably require a large number of high-tech industries. Industrial robots are high-tech products in modern society, playing an important role in the process of economic development, especially in the manufacturing industry. Industrial robot technology is widely used in automated production lines, greatly improving industrial production efficiency. Replacing manual labor for various complex production operations to achieve industrial production automation. Analyze the current application of industrial robots in automated production lines, explore their future development direction, in order to better serve the manufacturing industry [6]. Wei, H. H conducted a bibliometric analysis of publications related to the application of social network analysis in the field

of engineering construction management to describe existing research activities and determine future directions in this research field. These publications were retrieved from the China National Knowledge Infrastructure Database. There has been a significant increase in the knowledge system of using social network analysis in the field of engineering construction management. Out of 513 retrieved literature, 98 relevant literature related to the application of social network analysis in the field of engineering construction management was selected for research and analysis through reading abstracts. Through a comprehensive analysis of keywords and relevant literature, it can be found that the application of social network analysis in the field of engineering construction management is mainly studied from the perspectives of stakeholders, construction projects, and workers [7].

In order to solve the problem of duplicate data entry between construction management platform systems in construction engineering, the author proposes to apply RPA intelligent process robots to replace manual data collection, operation, entry, and verification. The design of the system is divided into overall architecture, instruction program loading process, human-machine interaction system level services, and other levels. An end-to-end procedural instruction transmission control method is adopted, establish a low-level control command output module for the online calibration system of the flight path, using a basic service architecture system, implement human-machine interactive control of the online calibration system for flight paths on the B/S architecture system. Build a record controller module for the inspection trajectory correction of RPA intelligent process robots, and perform feedback control during the trajectory correction process in LOG-CONTROL-BLOCK, using the RPA feedback correction algorithm, achieve adaptive correction and error feedback tracking of the inspection trajectory of RPA intelligent process robots. Implement calibration system development and design in an integrated DSP (Digital Signal Processing) information processing platform.

3. Intelligent Process Robot Based on RPA Technology.

3.1. Research Content and Objectives.

(1) *Research content.* After research and analysis, the nuclear power project takes the construction of "smart nuclear power" as an opportunity to research new generation information technologies such as RPA technology, OCR, artificial intelligence, etc. on the basis of the existing nuclear power multi-project management system, ultimately, a set of intelligent process management platforms with nuclear power project management characteristics will be formed, achieving intelligent control of project document management, budget data, item warehousing, and other processes, replacing personnel automation for process operations.

(2) *Business pain points.* The project of China Nuclear Power Fifth Company has deployed a nuclear power multi-project construction management system, which can basically cover all businesses during the construction phase of nuclear power projects, however, there is no data interface between the nuclear power multi-project management archive management system (EN-POWER) and the employer's construction management system (IFS3.0), there is a large amount of data re recording work, which not only increases labor costs and low work efficiency, but also cannot guarantee accuracy[8]. At the same time, there is a significant amount of data guidance work in the areas of construction budget data, financial reimbursement, and construction task sheet data backfill. Therefore, there is an urgent need to use more intelligent methods to solve this problem.

(3) *Research objectives.* Based on the analysis of the above issues, nuclear power projects take the construction of "smart nuclear power" as an opportunity to investigate and research new generation information technologies such as RPA technology, OCR, artificial intelligence, etc. on the basis of the existing nuclear power multi-project management system, ultimately, a set of intelligent process management platforms with nuclear power project management characteristics will be formed, achieving intelligent control of project archive management, budget data entry, item management, and construction process management processes, this will replace personnel in automatic file merging, naming, authorization, entry, uploading attachments, distribution, and archiving of the entire process automation management, and replace manual automatic operation and data entry business[9].

1. Having the characteristics of low cost, low error rate, high accuracy, compliance, and 24/7 work III;
2. Implement cross platform automated operation between ENPOWER and IFS3.0 systems;
3. Implement automated input of ENPower budget data and item arrival data;
4. Replacing personnel to automatically execute repetitive business processes, saving labor costs;
5. Automated execution of procedural operations to improve work efficiency; Reduce the workload of technical personnel, enable technical personnel to focus more on creative work;

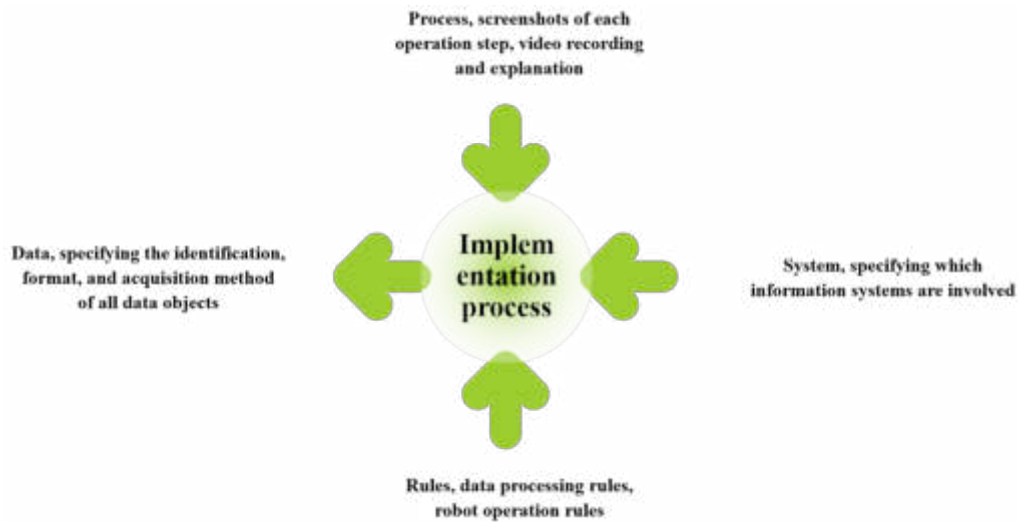


Fig. 3.1: RPA study technical route

6. Seamless connection with ENPOWER for data integration; Assist in Digital transformation of the project [10].

3.2. Technical Proposal and Application.

(1) *Design Platform and Research Methods.* The nuclear power project uses UiBot as the design platform for intelligent process robots; The main research Technology roadmap is shown in Figure 3.2:

1. Sort out and analyze the workflow of various existing information systems, and mine application points; Transform highly repetitive and relatively fixed processes into "intelligent process robots" to achieve automation;
2. By designing a platform, we can develop and implement "intelligent process robots" from a technical perspective, by simulating mouse, keyboard operations, and data interaction during human-computer interaction in specific scenarios, computers can independently operate and complete work tasks[11];
3. Deploying the developed "intelligent process robot" into the actual working environment, computer users can start the intelligent process robot automatically with just one click, and monitor the robot's operation status, if problems occur, they need to handle them in a timely manner.

(2) *Introduction to the functional modules of the R&D platform.* The research and development platform for intelligent process robots based on RPA technology consists of designers, runners, AI integration platforms, and intelligent process robots. The process robot equipment and management platform are shown in Figure 3.3:

1. Designer: Mainly used for developing "intelligent process robots", and can also run and debug RPA robots; Mainly designed to meet the needs of users in developing and designing process robots for different scenarios, helping users easily complete the design work of machine process automation;
2. Runner or controller: After RPA development is completed, users use the runtime platform to run the built robot; When it is necessary to run "intelligent process robots" on multiple computers, these "software robots" can be centrally controlled, such as unified distribution and setting of startup conditions[12];
3. AI integration platform: Providing intelligent process robots with various AI capabilities required for executing process automation; Through OCR character recognition, Natural language processing, image recognition and other AI technologies, the process processing capability and efficiency are further improved.

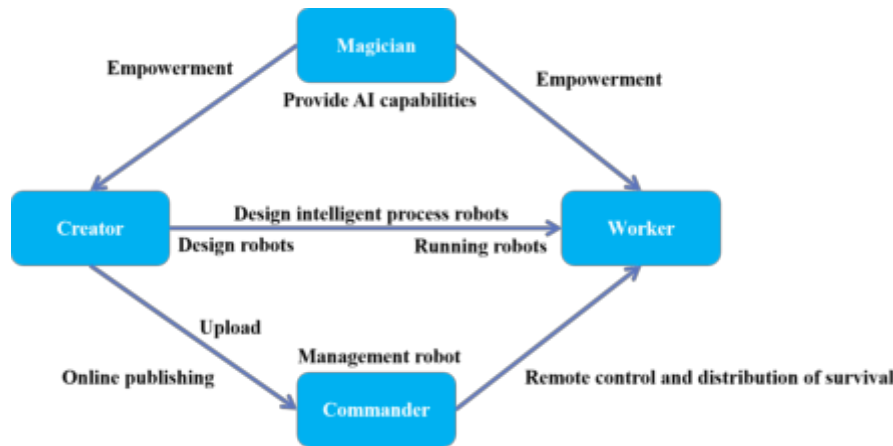


Fig. 3.2: Intelligent process robot design and management platform

3.3. Technological innovation and progressiveness . In order to solve the duplicate data entry work between the IFS and ENPower systems, the Nuclear Power Project Department of CNNC No.5 Company, by researching and developing RPA technology, we greatly reduce the error rate of personnel entering data and reduce project personnel costs. Instead of technical personnel, we automatically perform the entire process of file merging, naming, authorization, input, uploading attachments, distribution, and archiving, saving labor costs[13]; Solved the problem of duplicate entry of design file data in the document management system of the contracting party (IFS3.0) and the nuclear power multi-project management system (ENPOWER). Its main innovation points are as follows:

(1) *Management Innovation.*

1. Implement cross platform automated operation between ENPOWER and IFS3.0 systems;
2. Implement automated input of ENPower budget data and item arrival data;
3. Replacing personnel to automatically execute repetitive business processes, saving labor costs;
4. Automated execution of procedural operations to improve work efficiency;
Reduce the workload of technical personnel and make them more focused on doing creative work;
5. Seamless integration of ENPOWER data; Use AI technologies such as OCR and image recognition to help Digital transformation of nuclear power projects.

(2) *Technological innovation.*

① Visual programming technology

The original visual programming and source code programming can be switched at any time, making it simple and easy to use. Designers do not need advanced programming skills, by visualizing process views and rich source code command library views, the presentation of processes and process blocks can be achieved, effectively improving the efficiency of RPA process development[14].

② Business processing and software process intelligence

In RPA intelligent process design and daily office processes, it is often necessary to automate commonly used software such as Excel, Word, and browser, the use of RPA technology can achieve intelligent operation of these software. Meanwhile, RPA can perform repetitive and mechanical operations based on pre written scripts, replacing manual task processing with automated processing to improve work efficiency.

③ Simulate human-computer interaction

The RPA intelligent process machine mainly simulates the manual operation of users, such as data entry, character copying, pasting, mouse clicking, keyboard input, etc., and automatically processes the data conversion between tables, automatically adjusts the document format and article layout, automatically sends and receives emails, automatically opens the links of inspection web pages, Document retrieval,



Fig. 3.3: The comparison of traditional working mode and RPA robot

collects data and other repeated operations. The comparison between its traditional work and RPA robots is shown in Figure 3.4.

④ Strong scalability and compatibility

Support custom plugins written in multiple programming languages such as Python, C/C++, C #, JAVA, etc., support custom commands, and support a multi-level developer ecosystem. At the same time, it has cross platform advantages, and the engine supports platforms such as Windows/Mac/Android. It is compatible with multiple PC and mobile devices, and supports various UI automation such as browsers, desktops, and SAP[15].

3.4. Design of online calibration algorithm for robot inspection trajectory. In order to achieve the design and research of the online calibration system for the audit robot’s inspection trajectory, combined with the algorithm design and bus transmission design for the online calibration of the audit robot’s inspection trajectory, a bus development design method is adopted, and the integrated DSP control is used to perform the online calibration and trajectory path tracking control of the audit robot’s inspection trajectory from the input end to the output end. The overall structural model of the audit robot’s inspection trajectory online calibration system is constructed, combining the hardware module design of the online calibration system for the audit robot’s inspection trajectory with the MicroChannel expansion bus, the overall control of the audit robot’s inspection trajectory online calibration system is carried out. Based on the audit robot’s inspection control unit, a component functional modular development method is adopted to establish a human-machine interaction control module for the audit robot’s inspection trajectory online calibration system, which is controlled by dynamic units, Implement the output unit conversion and overall structural design of the online calibration system for the audit robot’s inspection trajectory.

In order to achieve the design and research of the online calibration system for the audit robot’s inspection trajectory, combined with the algorithm design and bus transmission design for the online calibration of the audit robot’s inspection trajectory, a bus development design method is adopted, and the integrated DSP control is used to perform the online calibration and trajectory path tracking control of the audit robot’s inspection trajectory from the input end to the output end. The overall structural model of the audit robot’s inspection trajectory online calibration system is constructed, combining the hardware module design of the online calibration system for the audit robot’s inspection trajectory with the MicroChannel expansion bus, the overall control of the audit robot’s inspection trajectory online calibration system is carried out. Based on the audit robot’s inspection control unit, a component functional modular development method is adopted to establish a human-machine interaction control module for the audit robot’s inspection trajectory online calibration system, which is controlled by dynamic units, Implement the output unit conversion and overall structural design of the online calibration system for the audit robot’s inspection trajectory.

$${}^4T_5^{-1}(q_i) = {}^4T_7 \prod_{i=6}^7 {}^{i-1}T_i(q_i) \tag{3.1}$$

Among them, 4T_7 is the 4th order equilibrium moment, and $T_i(q_i)$ is the average degree of freedom of the RPA intelligent process robot, q_i is the balance parameter of the robot’s end pose, and the robot’s end pose constraint control is used to establish an adaptive planning model for the inspection trajectory of the RPA intelligent process robot, combined with the center of gravity offset planning, the feedback constraint parameters for the

inspection trajectory center adjustment of the RPA intelligent process robot are obtained as follows:

$$q_0 = [\alpha_0, \beta_0, \gamma_0]^T \equiv [\theta_1, \theta_2, \theta_3]^T \tag{3.2}$$

Among them, $\alpha_0, \beta_0, \gamma_0$ represents the coordinates of RPA intelligent process robot in the patrol Polar coordinate system, $\theta_1, \theta_2, \theta_3$ is the phase parameter of the RPA intelligent process robot's inspection trajectory space. When measuring distance and pose, the offset correction is performed based on the calibration method of the plane template, resulting in a pose offset of $q_1 = [q_1, \dots, q_7]^T = [\theta_4, \dots, \theta_{10}]^T$; By homogeneous transformation, the pose is transformed into the robot base coordinate system, and the fuzzy information parameters of the inspection trajectory of the RPA intelligent process robot are composed of n omnidirectional motion parameters, the dynamic distribution function of the calibration trajectory is:

$$\begin{aligned} \min F(x) &= (f_1(x), f_2(x), \dots, f_m(x))^T \\ \text{s.t. } g_i &\leq 0, i = 1, 2, \dots, q \\ h_j &= 0, j = 1, 2, \dots, p \end{aligned} \tag{3.3}$$

Among them, $f_1(x), f_2(x), \dots, f_m(x)$ represents the contour sensing output parameters for the inspection trajectory inspection of RPA intelligent process robots, respectively, is the dynamic torque of the robot's end pose, h_j is the calibration feature parameter for path correction, and q and p respectively represent the calibration object positions of the inspection trajectory of the RPA intelligent process robot, based on this, a center of gravity offset planning model for online calibration of RPA intelligent process robot inspection trajectory is constructed, represented as:

$$H(s) = \frac{e^{-\tau s}}{1 + G_c(s)G_0(s)} \tag{3.4}$$

Among them, $G_c(s)$ represents the main control parameter for the center of gravity shift of the inspection trajectory of the RPA intelligent process robot; $G_0(s)$ represents the expected pose parameters of the inspection trajectory of the RPA intelligent process robot, $e^{-\tau s}$ represents the dynamic error of the inspection trajectory of the RPA intelligent process robot, for solving constrained nonlinear optimization problems, based on the correction method of center of gravity shift, the calibrated dynamic parameter distribution model in the robot's base coordinate system is obtained as follows:

$$L = J(w, e) - \sum_{i=1}^N \alpha_i \int_{i=1}^M H(r) \{w^T \phi(x_i) + b + e_i - y_i\} \tag{3.5}$$

Among them, $J(w, e)$ is the inertia function of motion along the expected path, α_i is the distribution along the edge of the expected path, is the adjustment function of the path edge, and w is the dynamic feature point for angle symmetry adjustment, $\phi(x_i)$ is the compensation torque, b is the alternating feature point in the path edge, e_i is the path offset error, y_i is the motion direction adjusted by the robot according to the path, and a discrete spatial planning method L is used to construct the inspection trajectory control model of the RPA intelligent process robot, the output is:

$$\begin{cases} K_i(d) = \sum_{r=1}^t \sum_{q=1}^{k_2} (x_{ir} - x_{irq})(x_{ir} - x_{irq})^T B_{irq} \\ F_1 = W_i^T H_2 W + f_i(d) \times \log(\frac{N}{n_i} + 0.01) \\ C_{oconst} = \sqrt{\sum_{i=1}^n \sum_{r=1}^t \sum_{p=1}^{k_1} [(x_{ir} - x'_{irp})(x_{ir} - x'_{irp})^T A_{irp}]^2} \end{cases} \tag{3.6}$$

Among them, t is the time sampling point, x_{ir} is the head torque, x_{irq} is the width required for the robot to swing, and B_{irq} is the extension direction of the path determined by the target point, A_{irp} is the directional

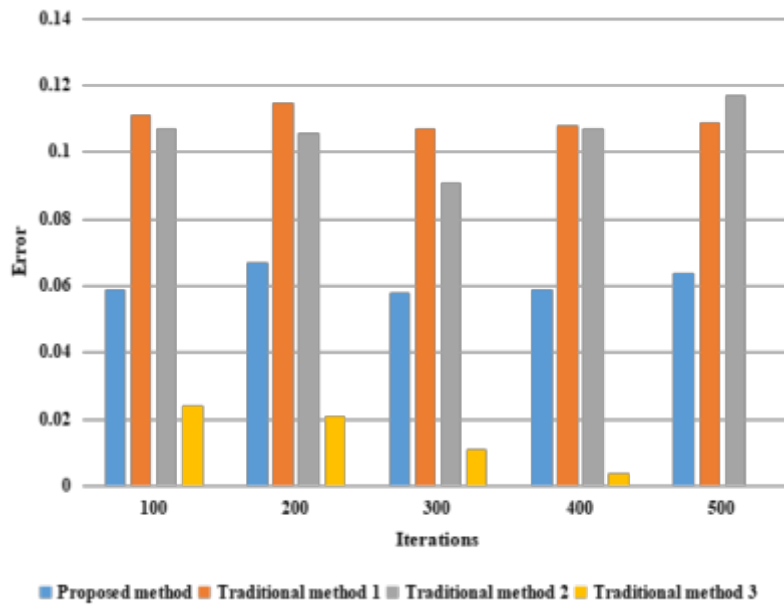


Fig. 4.1: Comparison of online correction error of intelligent robot inspection track

control parameter, W_i is the intersection point of the expected effective edge of the path, H_2 is the modeled parameter moving along the centerline of the path, and W is the alternating parameter between path edges, $f_i(d)$ is the mass of a single module, N is the target point selection parameter, n_i is the initial pose. Based on the above analysis, the RPA feedback correction algorithm is used to achieve adaptive correction of the inspection trajectory and error feedback tracking of the RPA intelligent process robot.

4. Application Achievements and Benefit Analysis. Feedback control during the trajectory correction process is carried out in LOG-CONTROL-BLOCK, using position sensors such as accelerometers and Doppler velocimeters (DVL) for data acquisition, the calibration system development and design were implemented in an integrated DSP information processing platform, and the results are shown in Figure 4.1. Analysis of Figure 4.1 shows that the error feedback performance of using this method for online calibration of RPA intelligent process robot inspection tracks is good, improving the accuracy of robot inspection[16].

The nuclear power project has been applied based on the conventional islands of Unit 1 and Unit 2 of the nuclear power plant, as well as some BOP engineering projects, with document, budget data, and item management as the entry points; Good economic benefits have been achieved through application, and the problem of duplicate data entry between the employer's IFS system and the ENPOWER document management system has been solved, greatly reducing error rates and personnel costs. It can replace manual data collection, entry, verification, and business operations; It has the characteristics of low error rate, low cost, high accuracy, compliance, and 24/7 standby; At present, the project has been applied in fields such as budget data and material management, with good application value and prospects[17,18]; It provides reference and guidance for the digital transformation work of similar nuclear power projects in the future, and its benefits in the construction of conventional nuclear power island projects are as follows:

RPA intelligent process robot, as a new software automation technology, is currently applied in nuclear power project documents, budget data, and item management; The application effect is good; through research and application, proved that RPA intelligent process robots can replace manual collection, input, verification of document data, and operation of business; Table 1 shows the labor cost savings after using RPA intelligent process robots.

5. Conclusion. Conduct online calibration of audit robot inspection trajectory, combine environmental

Table 4.1: The saved labor costs

Serial number	Year	New output value/10000 yuan	Cost savings /10000 yuan
1	2021	0	10.8
2	2022	0	43.3
3	2023	0	43.2
Three years of conservative calculation of the economic benefits generated			97.2

Note: After the above analysis, after the RPA intelligent process machine is put into use, it is expected to save 8 data or losers in the fields of budget data, item management and document management; cost (54 000 yuan / year 8 people) for 2 years + RMB 108,000 = RMB 0972,000.

parameter identification and obstacle avoidance processing to improve the robot's inspection control performance, and propose a design method for an RPA based audit robot inspection trajectory online calibration system. Establish an adaptive planning model of the audit robot's patrol path, use the Discretization space planning method to build the control model of the audit robot's patrol path, and use the RPA feedback correction algorithm to realize the adaptive correction and error feedback tracking of the audit robot's patrol path. The experimental results show that it has achieved good economic benefits through the application, solved the problem of repeated data entry between the employer's IFS system and the authorized document management system, and greatly reduced the error rate and personnel cost. It can replace manual data collection, input, verification and business operation; low error rate, low cost, high accuracy, compliance and 24 / 7 reserve; 108,000 in 2021, 432,000 in 2022, and 432,000 in 2023. Through the application research of RPA intelligent process robot technology, we have promoted the transformation and functional improvement of nuclear power project information processes, and explored a path of information management with nuclear power characteristics; Making the information management of nuclear power projects increasingly "intelligent", while improving work efficiency and reducing management costs, it also breaks the data silos with the contracting party's system, exploring a new approach and method for subsequent digital transformation and intelligent project management; In the future, it will be promoted and applied in fields such as construction task orders, financial accounting, smart warehousing, and invoice verification, which has good application value and market prospects. It is hoped that the research on the application of intelligent site system can play a reference role in practical work and make a contribution to the innovation and development of the construction industry.

I hope to do better or make breakthroughs in the following areas in the future, so that the intelligent search robot system can better serve everyone and create a new situation for us to use the Internet.

The stability and real-time performance of the system. As the number of users continues to grow and the load on the system continues to increase, the system can still maintain its stable and fast characteristics.

The expansion of business seeks businesses with more business opportunities, linking actual business with virtual internet.

REFERENCES

- [1] Guo, J. (2021). Research on the application of intelligent robots in explosive crime scenes. *International Journal of System Assurance Engineering and Management*, 14(2), 626-634.
- [2] Chuanbao, W., Jin, H. E., Liang, L. I., & Yue, Z. (2021). Application of bim technology in the construction management of oil and gas field stations. *Oil-Gas Field Surface Engineering*, 58(6), 85-89.
- [3] Khoso, A. R., Yusof, A. M., Chen, Z. S., Wang, X. J., & Memon, N. A. (2021). Embedded remote group environment through modification in macbeth-an application of contractor's selection in construction. *Journal of Civil Engineering and Management*, 47(1), 96-102.
- [4] Cai, H. (2022). Building construction operation simulation based on bim technology and intelligent robots. *Journal of Interconnection Networks*, 178(14), 88-91.
- [5] Varlamov, O. (2021). "brains" for robots: application of the mivar expert systems for implementation of autonomous intelligent robots. *Big Data Research*, 25(head-of-print), 100241.

- [6] Zhou, L., Wang, F., Wang, N., & Yuan, T. (2021). Application of industrial robots in automated production lines under the background of intelligent manufacturing. *Journal of Physics: Conference Series*, 1992(4), 042050-.
- [7] Wei, H. H., Zhang, Y., Sun, X., Chen, J., & Li, S. (2023). Intelligent robots and human–robot collaboration in the construction industry: a review. *Journal of Intelligent Construction* 33(7), 9180002-9180002.
- [8] Xie, J. Y. Y. (2021). Iot-based model for intelligent innovation practice system in higher education institutions. *Journal of intelligent & fuzzy systems: Applications in Engineering and Technology*, 40(2),98.
- [9] Dahl, M., Bengtsson, K., & Falkman, P. (2021). Application of the sequence planner control framework to an intelligent automation system with a focus on error handling,18(74),88-93.
- [10] Guo, Y. (2022). Research on the construction path of the target management system of the party building work in the tobacco industry. *Journal of Higher Education Research*, 3(4), 363-366.
- [11] Xiong, W., Xiang, Y., & Meng, K. (2021). Exploration and practice of engineering management specialty in the teaching of innovative curriculum system construction. *Francis Academic Press*,978(18),587-589.
- [12] Berco, V., Pfkukani, N. S., & Hendri, D. P. (2021). Factors influencing the adoption of building information modelling (bim) in the south african construction and built environment (cbe) from a quantity surveying perspective. *Engineering Management in Production and Services*, 13(8),447-449.
- [13] Feng, L., & Zhao, J. (2021). Research on the construction of intelligent management platform of garden landscape environment system based on remote sensing images. *Arabian Journal of Geosciences*, 14(14), 1-19.
- [14] Yang, C., & Zhang, X. (2022). Research into the application of ai robots in community home leisure interaction. *Journal of supercomputing*,58(7), 78.
- [15] C, J. H. A., & B, B. P. (2022). Study the path planning of intelligent robots and the application of blockchain technology. *Energy Reports*, 8(3), 5235-5245.
- [16] Zhao, M., Mao, Y., Hen, Q., & Zhou, Y. (2021). Research on problems and countermeasures in the application of substation intelligent inspection system. *Journal of Physics: Conference Series*, 1983(1), 012084 (7pp).
- [17] Banteng, B. S. D., & Uno, W. R. (2021). Application of time management systems in tardiness of the auditorium construction projects in the district bone bolango, gorontalo province. *IOP Conference Series: Materials Science and Engineering*, 1098(2), 022028 (4pp).
- [18] Alhusban, M., Elghaish, F., & Matarneh, S. T. (2022). The application of "deep learning" in construction site management: scientometric, thematic and critical analysis. *Construction Innovation* , 22(3), 580-603.

Edited by: B. Nagaraj M.E

Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing

Received: Dec 20, 2023

Accepted: Feb 27, 2024



OBSTACLE AVOIDANCE PATH PLANNING FOR POWER INSPECTION ROBOTS BASED ON DEEP LEARNING ALGORITHMS

YUXIN LIU^{*}, XIAOXI GE[†], HAOWEI JIA[‡], LIN YUAN[§] AND MIN ZHOU[¶]

Abstract. The current research on obstacle avoidance path planning methods for power inspection robots has problems such as poor obstacle avoidance ability and poor inspection effectiveness. Therefore, a planning method for obstacle avoidance path of power inspection robots is proposed. By utilizing motion relationships and the potential field theorem of robot motion, a three-dimensional model of the power inspection robot's route is established to determine the direction of the robot's route when obtaining action tasks. The fuzzy support vector algorithm is used to plan obstacle avoidance paths for the initialized walking path, making the inspection robot intelligent. The experimental results show that the average success rates for avoiding static and dynamic obstacles are 98.37% and 96.12%, respectively. The average time for obstacle avoidance path planning is 1.56 seconds, and it has fast, efficient, and accurate obstacle avoidance and path planning capabilities, which can improve the robot's obstacle avoidance ability and path planning efficiency for dynamic and static obstacles.

Key words: path planning, Inspection robot, three-dimensional model

1. Introduction. Entering the 21st century, with the continuous development of economies and cultures in various countries, the level of technology is also constantly improving. Among them, robotics is one of the most eye-catching development disciplines. The development of robotics has made an important contribution to the progress of social civilization and the development of market economy, and has played an important role in human's food, clothing, housing and transportation. Robots can complete various high-load, difficult, and high-precision tasks that are difficult for humans to complete, such as medical, military, agricultural, and other aspects [1]. This has greatly liberated productivity, improved human labor efficiency, and also improved the efficiency of human technological development, making significant contributions to the technological development of various countries and regions.

Nowadays, automation reform has emerged in many labor-intensive industries. The emergence of automated robots has saved human resources and greatly reduced labor costs. At the same time, labor efficiency in various industries such as manufacturing assembly lines has been greatly improved. Robots have been widely used in many fields, such as medical robots, transportation robots, etc [2]. With the continuous innovation of high-precision sensors and advanced artificial intelligence algorithms, it is not only possible to use robots in large-scale industrial production workshops, but also to complete related tasks in crowded indoor spaces and even within the human body. Some service robots, such as outdoor cleaning robots, robot nurses, and smart home assistants, have greatly improved people's quality of life. The research and development of wheeled inspection robots applied to various industries, such as warehousing and logistics, and electrical equipment inspection work, began as early as the beginning of this century. With the continuous development of modern power systems, both industrial and residential electricity consumption is significantly increasing, and the requirements for the long-term stable operation of substations in the power grid are also constantly increasing [3]. Therefore, how to complete inspection work more efficiently and accurately, the proposal of this issue further promoted the research

^{*}Zhengzhou Railway Vocational & Technical College, ZhengZhou, 450052, China (Corresponding author's e-mail: YuxinLiu3@163.com)

[†]Zhengzhou Railway Vocational & Technical College, ZhengZhou, 450052, China (XiaoxiGe53@126.com)

[‡]Zhengzhou East High Speed Rail Infrastructure Section, China Railway Zhengzhou Group Co., Ltd., ZhengZhou, 450052, China (HaoweiJia7@163.com)

[§]Beijing Institute of Science and Technology, China Railway Beijing Group Co., Ltd., Beijing, 100081, China (LinYuan771@126.com)

[¶]Zhengzhou High Speed Rail Infrastructure Section, China Railway Zhengzhou Group Co., Ltd., ZhengZhou, 450052, China (MinZhou29@163.com)

and development process of power inspection robots. In addition, the country has also invested a large amount of financial support in the use of industrial site inspection robots. Currently, China is undergoing a process of developing from traditional manufacturing to modern manufacturing. Revitalizing the manufacturing industry and realizing its industrialization are of great significance for the vigorous development of the economy. In the process of industrial development, mechanical automation is a necessary stage to achieve industrialization. It is not difficult to see from the industrialization development process of developed countries in the past that the improvement of production efficiency and the continuous expansion of industrial productivity must go through the process of mechanization, automation, intelligence, and information transformation [4]. With the rapid development of the national economy, the continuous improvement of industrial production efficiency, and the continuous increase in human resource costs, the use of automated robots instead of manual inspection has gradually become an inevitable direction for industrial equipment inspection and maintenance. The traditional manual inspection method has many shortcomings, such as large workload and low detection efficiency; The detection effect is not satisfactory, and the detection method mainly relies on visual inspection, resulting in significant errors; In some extreme meteorological environments, such as thunderstorm days, traditional manual detection methods pose safety hazards for detection personnel and cannot complete troubleshooting in a timely manner; The traditional inspection method, which mainly involves installing cameras at designated locations, has a large blind spot due to the limitations of the camera's shooting range, making it difficult to truly meet the requirements of comprehensive fault screening within the station. At the same time, due to the cumbersome design of the control system, the large number of equipment installations, and poor economic efficiency, this inspection method has a high false detection rate for faults and poses great difficulties in maintaining monitoring equipment.

2. Literature Review. The increasing amount of data in the power system greatly increases the task of power transmission, and traditional manual power grid inspections face greater risks. Adopting robots instead of manual power grid inspections can not only ensure the health of workers, but also improve inspection efficiency and create higher value economic benefits. The inspection robot must carry out path planning, that is, in an unknown environment with obstacles, plan the best running path that can avoid all obstacles, which is of great significance [5].

Abdallaoui, S conducted a comprehensive and up-to-date overview analysis and rigorous review of the safety and best path of autonomous vehicle. The focus is on sampling algorithms, node based optimization algorithms, mathematical model based algorithms, bioheuristic algorithms including neural network algorithms, and multi fusion based algorithms, which combine different methods to overcome their respective shortcomings. All of these methods consider different conditions and are used in multiple fields [6]. Xu, T proposed an improved artificial potential field method, in which the object can leave the local minimum point trapped by the algorithm while avoiding obstacles and following a shorter feasible path along the repulsive equipotential surface of local optimization. The entire obstacle avoidance process is based on an improved artificial potential field method, which is applied to the path planning action of the robotic arm, along the motion from the starting point to the target point. The simulation results of the research results show that compared with the improved artificial potential field method based on fast exploration random trees, the algorithm proposed in this paper can effectively perceive the shape of obstacles in all selected situations, and can effectively shorten the distance of the planned path by 13%-41%, significantly improving the planning efficiency [7]. Cheng, J proposed an intelligent robot food runner suitable for restaurants with lower prices but better performance. Among them, this article mainly analyzes how to use LiDAR SLAM to establish restaurant maps, positioning, and navigation, as well as how to establish obstacle avoidance and path planning. Through the ROS platform, the entire process of the intelligent robot vegetable runner is simulated and verified, which can meet the needs of restaurants [8].

Traditional power inspection robots, due to their immature technology, may collide with power equipment during the inspection process, resulting in power inspection accidents and causing losses to both the power inspection robots and the power system, therefore, the author proposes a machine learning based obstacle avoidance path planning method for power inspection robots, achieving the goal of safe inspection for inspection robots.

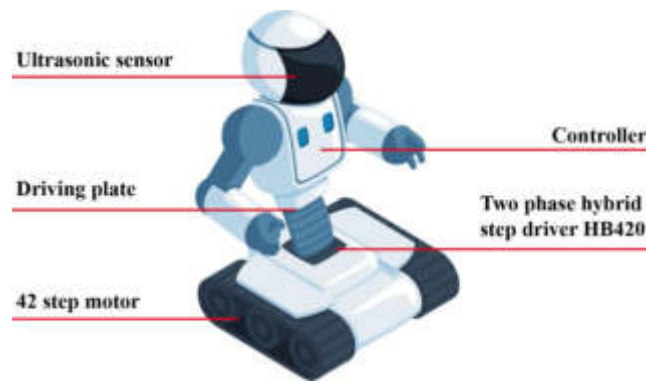


Fig. 3.1: 3D Model of Electric Power Inspection Robot

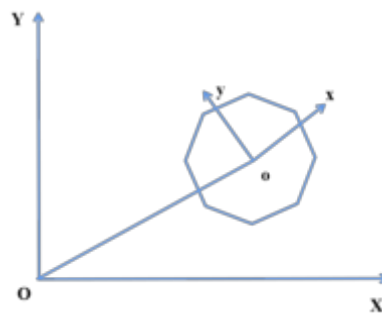


Fig. 3.2: Coordinate System of the 3D Model of the Electric Power Inspection Robot

3. Application of machine learning in obstacle avoidance path planning for power inspection robots.

3.1. Establishment of a three-dimensional model for the route of the power inspection robot

. The goal of the author's design of a three-dimensional model for the route of the power inspection robot is to ensure stable inspection, and to enable the power inspection robot to have the ability to recognize and perceive the direction of the inspection path, and to complete the inspection path planning of the inspection task in all aspects [9].

In order to improve the stability of the inspection robot, four intersecting driving wheels are designed at the bottom of the power inspection robot based on physical principles, the motion direction and period of the driving wheels are the same, the specific 3D model of the physical power inspection robot is shown in Figure 3.1.

The perception of direction is very important for power inspection robots. Once the robot's directional perception ability decreases, it will cause the inspection route of the power inspection robot to deviate from the normal inspection route, and may collide with other electronic system inspection obstacles or equipment, resulting in power inspection errors [10]. In order to solve the above problems, the author uses motion models and terrain strength to establish a three-dimensional model of the route of the power inspection robot, so as to optimize the perception angle of the center of gravity of the power robot, therefore, the author chooses the center of mass of the power robot as the center origin of the model, and the two-dimensional coordinate system diagram of the power inspection robot is shown in Figure 3.2.

Simulate the motion behavior of the power inspection machine, firstly, set the direction vector of the four driving wheels of the inspection robot as P , and then combine it with real-time environmental conditions, firstly,

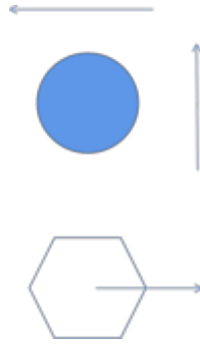


Fig. 3.3: Schematic diagram of the rotation route of the power inspection robot

calculate the relative position of the robot within the inspection range, and the calculation formula is as follows:

$$P_B = \left| \frac{x(t)}{LW}, \frac{y(t)}{LW} \right| \sum \frac{1}{L} \left[\frac{Q}{\prod} \right] P_n \quad (3.1)$$

Among them, P_B represents the starting position of robot inspection; $X(t)$ represents the directional guidance coefficient of the X-axis of the inspection robot; $Y(t)$ represents the directional guidance coefficient of the Y-axis of the inspection robot; L represents the radius of the driving wheel of the inspection robot; P_n represents the motion direction vector of the driving wheel of the inspection robot. W_x and W_y represent the weight matrices of the X-axis and Y-axis of the 3D model of the power inspection robot route, respectively [11].

The power inspection robot will perform basic inspection behavior operations during the inspection process, such as walking straight, turning, reversing, translating, and rotating, in order to improve the smoothness of the operation of the power inspection robot, the author uses the theorem of resultant force balance to constrain the robot's inspection motion behavior during the robot inspection process. The schematic diagram of the circular motion path planning for the power inspection robot is shown in Figure 3, and the formula is as follows:

$$F = \frac{f_7}{\varepsilon} \sqrt{Fx^2 + Fy^2} \quad (3.2)$$

Among them, f_7 represents the repulsive force of the motion of the electric inspection robot; F represents the combined force of the motion of the electric inspection robot; Fx^2, Fy^2 represents the motion components of the electric inspection robot on the X-axis and Y-axis, respectively; ε represents the robot's motion balance coefficient [12].

3.2. Robot obstacle avoidance path planning. Machine learning technology is widely used in various fields such as home services, industrial guidance, and military operations. Machine learning is divided into two types: Single machine machine learning technology and multi machine machine learning technology, select the best machine learning method based on the difficulty of object-oriented machine learning. The application range of single machine learning technology is relatively limited compared to multi machine machine learning technology, based on the obstacle avoidance path planning method designed by the author for power inspection robots, multi machine machine learning technology is selected, this technology can complete the planning of static and dynamic paths through learning the environment, and has a self verification process during the path planning process to avoid redundant planning paths [13].

The power inspection robot determines its own location and plans the specific path that the power inspection robot needs to inspect based on the inspection tasks sent by the control center. During the planning process, the repulsion function is used to determine the effective range of the inspection, and obstacles within the inspection range are marked using an artificial potential field method. The principle of obstacle marking is that

the artificial potential field at the location of the obstacle, combined with the field strength of the real-time environment, will emit a repulsive force outward, which affects the gravitational force of the inspection target on the inspection robot's route. The electric inspection robot determines the specific position of the inspection obstacle based on the magnitude of the gravitational force. The repulsion function is as follows:

$$U_t = \frac{k_1}{O} \quad (3.3)$$

Among them, U_t represents the repulsion function; O indicates the relative distance between the inspection robot and the obstacle; k_1 represents the coefficient [14].

After identifying the effective range and obstacles for inspection, the power inspection robot can complete the planning of obstacle avoidance routes for the first time, this route plan will eliminate the accessible routes with obstacles, but if all routes have obstacles, the obstacle avoidance function of the power inspection robot needs to be activated. The author uses the DWA sliding window method to drive the power inspection robot to avoid obstacles during operation and stably complete the inspection work, the formula for generating obstacle avoidance motion behavior is as follows:

$$y(h) = V_s \times \frac{V_a}{V_b} + \frac{ad_{path} + \beta d_{good} + \gamma d_{obstacle}}{\Delta t} \quad (3.4)$$

Among them, $y(h)$ represents the obstacle avoidance command of the power inspection robot; V_b represents the angular velocity of the power inspection robot; V_s represents the linear velocity of the motion of the power inspection robot; V_a represents the acceleration of the motion of the power inspection robot; Δt represents the inspection cycle of the inspection robot; ad_{path} represents the shortest distance between the inspection robot and the obstacle; βd_{good} represents the distance from the endpoint of the trajectory to the local target; $\gamma d_{obstacle}$ represents the maximum obstacle cost for the operation trajectory of the power inspection robot [15].

Using a fuzzy support vector model to set path planning constraints, according to the author's research objectives, setting path planning constraints to maximize the inspection range and minimize the inspection path can improve the efficiency of obstacle avoidance path planning for power inspection robots. The formula for the constraint conditions is as follows:

$$D = \sum_{i=1}^n a_i \times \frac{f(k)}{2} + \frac{y(h)}{\min_L \frac{c_1+c_2}{2}} \quad (3.5)$$

Among them, D represents the constraint condition model; c_1, c_2 represents the mean of the decision functions for the upper and lower bounds of the fuzzy support vector machine model; a_i represents the membership degree of the fuzzy control algorithm; \min_L represents the minimum path for inspection; $f(k)$ represents the kernel function of the inspection probability of the power inspection robot, and the meaning of other unknowns is the same as above. Finally, machine learning technology is used to plan the path constraints and obstacle avoidance behavior instructions for the inspection of the power robot, and the planning formula is as follows:

$$s(t) = f(x, y, z) + \omega \times maxr - Q(x, y, z) \times \mu \quad (3.6)$$

Among them, $s(t)$ represents the inspection path of the power inspection robot; $Q(x, y, z)$ represents the feature vector of inspection behavior classification; μ represents the path planning coefficient; ω represents the inner product of high-dimensional feature space vectors; $F(x, y, z)$ represents the loss function of path optimization.

4. Experimental Results and Analysis. Through the above analysis and design, the design of a machine learning based obstacle avoidance path planning method for power inspection robots has been completed, in order to verify the working performance of this method, the author utilized the obstacle avoidance path planning method for power inspection robots based on GPS navigation technology (traditional method 1) and the obstacle avoidance path planning method for power inspection robots based on carrier free communication technology (traditional method 2) to jointly complete comparative experimental testing, ensuring the scientific nature of the testing [16]. In order to improve the reliability and analyzability of the test results, the inspection

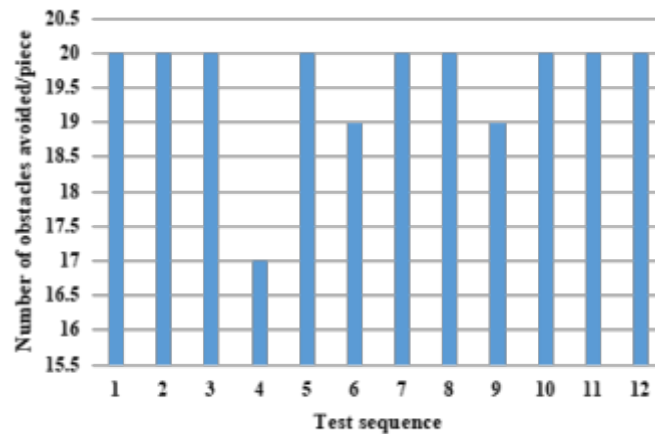


Fig. 4.1: Statistics of Robot Obstacle Avoidance Quantity

robots tested in the article are all HKD09 series inspection robots, the functions of the drivers, motors, and other accessories of this series of inspection robots are optimal and will not result in the experiment being terminated due to the inspection robot. Set at 12 meters \times The 12 meter area is a simulation environment, where 20 rectangular obstacles are unevenly distributed. The starting position and target point coordinates of the inspection robot are $[0,0]$ and $[11,11]$, respectively, with a robot step size of 0.50. 12 repeated obstacle avoidance tests were conducted based on the planned path trajectory, and the obstacle avoidance results are shown in Figure 4.1 [17].

As shown in Figure 4.1, during the obstacle avoidance test of the inspection robot in the established path, among them, the accuracy of obstacle avoidance for 9 times reached 100%, with an average obstacle avoidance rate of 97.92%, proving that the author's method has good obstacle avoidance effect [18].

Comparative experiments were conducted using the author's method to compare the effectiveness of obstacle avoidance with traditional methods 1 and 2, respectively, under the same workspace and number of obstacles, static and dynamic obstacle avoidance experiments were conducted, and the experimental results are shown in Figure 4.2.

From Figure 4.2(a), it can be seen that the author's method outperforms traditional method 1 and traditional method 2 in avoiding static obstacles, in the case of the initial two obstacles, the success rates of obstacle avoidance for the three methods are almost the same, reaching over 99.90%. However, as the number of obstacles increases, the success rates of obstacle avoidance for all three methods show a downward trend, the author's method has an average obstacle avoidance success rate of 98.37% for static obstacles, which is 8.37% and 3.49% higher than the comparison methods of traditional method 1 and traditional method 2, respectively [19]. From Figure 5b, it can be seen that when facing dynamic obstacles, the difference in obstacle avoidance success rates among the three comparison methods gradually widens as the number of obstacles increases. As the number of dynamic obstacles increases, all show significant fluctuations, the author's method has an average obstacle avoidance success rate of 96.12% for dynamic obstacles, which is 15.03% and 9.10% higher than the comparison method of traditional method 1 and traditional method 2, respectively. The results indicate that, the author's method has good obstacle avoidance ability in both static and dynamic obstacles. Under the same conditions, three methods were used to conduct multiple obstacle avoidance path planning experiments for inspection robots, and the data of 8 path planning times is shown in Figure 4.3.

From Figure 4.3, it can be seen that the author's method takes the highest time of 1.80 seconds and the lowest time of 1.40 seconds in obstacle avoidance path planning, with an average time of 1.56 seconds, the traditional method 1 takes the highest time of 3.10 seconds, while the traditional method 2 takes the highest time of 2.10 seconds, with an average time of 1.12 seconds and 0.20 seconds, respectively [20].

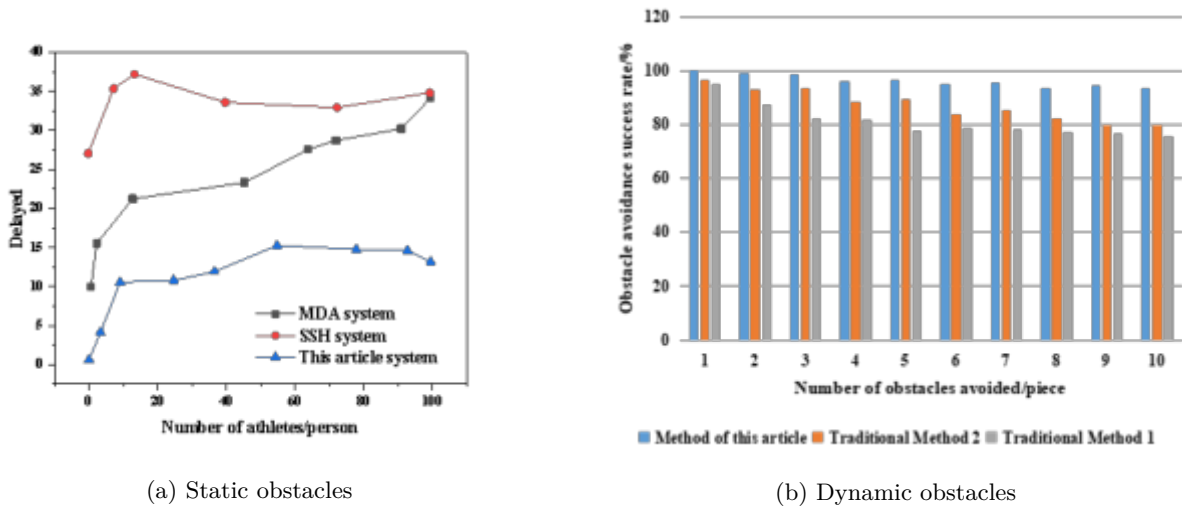


Fig. 4.2: Comparison Test Results for Obstacle Avoidance

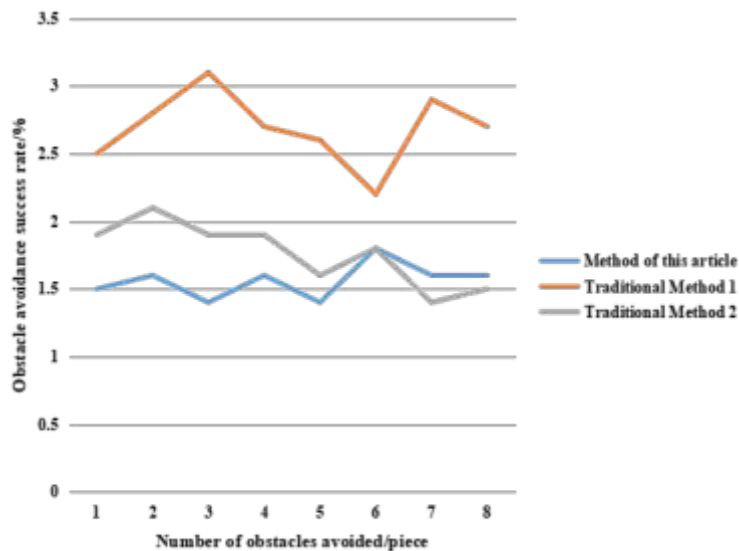


Fig. 4.3: Statistics of robot obstacle avoidance path planning time consumption

5. Conclusion. Based on fuzzy control algorithms, the directional recognition and perception ability of obstacle avoidance paths for power inspection robots has been fundamentally improved, and the robot is controlled to minimize rotation errors during turning behavior. Utilizing motion models and potential field theorems to improve the reasonable path planning ability of inspection robots, resulting in the output of the planned route with minimal obstacle avoidance and the widest effective range of inspection. Machine learning algorithms have improved the self-learning habits and intelligence of power inspection robots for inspection path planning, enabling them to achieve the goal of safe inspection. Obstacle avoidance methods based on machine learning, on the one hand, it can effectively avoid conflicts between algorithm time and accuracy in planning.

On the other hand, by adjusting the repulsive potential function, gravitational potential function, and the calculation of the resultant force, it can enhance the adaptability of the inspection robot to the environment, thereby improving the robot's obstacle avoidance ability and inspection efficiency.

REFERENCES

- [1] Zhou, Y. , Su, Y. , Xie, A. , & Kong, L. . (2021). A newly bio-inspired path planning algorithm for autonomous obstacle avoidance of uav. *Chinese Journal of Aeronautics*, 15(7), 20.
- [2] Jones, M. , & Peet, M. M. . (2021). A generalization of bellman's equation with application to path planning, obstacle avoidance and invariant set estimation. *Automatica*, 31(6), 1729-1739.
- [3] Agarwal, D. . (2021). Implementing modified swarm intelligence algorithm based on slime moulds for path planning and obstacle avoidance problem in mobile robots. *Applied Soft Computing*, 107(1),96-99.
- [4] Yehliu, K. . (2021). Path planning and obstacle avoidance for automated driving systems using rapidly-exploring random tree algorithm. *SAE International Journal of Connected and Automated Vehicles*,86(3), 4.
- [5] Yanrong, H. , & Yang, S. X. . (2021). A knowledge based genetic algorithm for path planning of a mobile robot. *Computational Intelligence and Neuroscience*, 14(3), 1-14.
- [6] Abdallaoui, S. , Aglzim, E. H. , Chaibet, A. , & A Kribèche. (2022). Thorough review analysis of safe control of autonomous vehicles: path planning and navigation techniques. *Energies*, 117(10), 12-28.
- [7] Xu, T. , Zhou, H. , Tan, S. , Li, Z. , Ju, X. , & Peng, Y. . (2022). Mechanical arm obstacle avoidance path planning based on improved artificial potential field method. *Industrial Robot*, 78(8), 11015-11050.
- [8] Cheng, J. , Liu, Z. , He, J. , Deng, Y. , & Zhang, H. . (2021). Application of simultaneous location and map construction algorithms based on lidar in the intelligent robot food runner. *Journal of Physics: Conference Series*, 1972(1), 012010-.
- [9] Xu, X. . (2021). Analysis of obstacle avoidance strategy for dual-arm robot based on speed field with improved artificial potential field algorithm. *Electronics*, 10.
- [10] Chen, G. , Sun, D. , Dong, W. , Sheng, X. , & Ding, H. . (2021). Computationally efficient trajectory planning for high speed obstacle avoidance of a quadrotor with active sensing. *IEEE Robotics and Automation Letters*, PP(99), 1-1.
- [11] Moller, T. , & Egberts, J. H. . (2021). Robot-assisted thoracic surgery-areas of application and limitations. *Der Chirurg; Zeitschrift für alle Gebiete der operativen Medizin*,85(2), 92.
- [12] Meng, H. , & Zhang, H. . (2022). Mobile robot path planning method based on deep reinforcement learning algorithm. *Journal of Circuits, Systems and Computers*, 31(15),77-79.
- [13] Miao, Z. , Zhang, X. , & Huang, G. . (2021). Research on dynamic obstacle avoidance path planning strategy of agv. *Journal of Physics Conference Series*, 2006(1), 012067.
- [14] Low, E. S. , Ong, P. , Cheng, Y. L. , & Omar, R. . (2022). Modified q-learning with distance metric and virtual target on path planning of mobile robot. *Expert Systems with Application(Aug.)*, 56(1), 91-109.
- [15] Yang, C. L. . (2021). A novel algorithm for path planning of the mobile robot in obstacle environment. *International Journal of Circuits*, 15(3), 225-235.
- [16] Zhou, H. , & Gu, M. . (2021). Application of neural network and computer in intelligent robot. *Journal of Physics: Conference Series*, 1881(3), 032028 (7pp).
- [17] Chen, Y. , & Zhou, X. . (2021). Research and implementation of robot path planning based on computer image recognition technology. *Journal of Physics: Conference Series*, 1744(2), 022097 (4pp).
- [18] Zhang, J. , Zhang, T. , Niu, Y. , Guo, Y. , Xia, J. , & Qiu, Y. , et al. (2022). Simulation and implementation of robot obstacle avoidance algorithm on ros. *Journal of Physics: Conference Series*, 2203(1), 012010-.
- [19] Wu, Z. . (2021). Decentralized path planning for multi-objective robot swarm system. *Journal of Physics: Conference Series*, 2113(1), 012002-.
- [20] Lyu, D. , Chen, Z. , Cai, Z. , & Piao, S. . (2021). Robot path planning by leveraging the graph-encoded floyd algorithm. *Future Generation Computer Systems*, 37(4), 1-9.

Edited by: B. Nagaraj M.E

Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing

Received: Dec 20, 2023

Accepted: Mar 18, 2024



ANALYSIS OF FROZEN DATA ANOMALY AND UPDATE METHOD OF ELECTROMECHANICAL ENERGY METER TERMINAL BASED ON DEEP LEARNING

FANG YAO* AND LIBIN TAN†

Abstract. In view of the lack of advanced and mature substation fault detection and facility detection technology, combined with the characteristics of the actual application environment of substation, a substation operating equipment autonomous monitoring and fault diagnosis detection system based on deep learning intelligent detection robot is proposed. That is, the deep learning algorithm, Big data analysis technology and patrol robot with HD camera are organically combined. The image information collected by the high-definition camera is fused with the data information collected by a variety of sensors, and then the fault tree and Big data analysis algorithm are used to carry out real-time intelligent detection and analysis of all equipment in the substation, and the early warning can be sent to the relevant equipment maintenance personnel in a timely manner. The experimental results indicate that, the number of input nodes in the fault tree is 7, the number of output nodes is 2, the number of center vectors is 14, the number of nodes in the basis function layer is 7, and the threshold of the basis function is set to 0.8257. In actual training, after 31 iterations, the training results can quickly converge to the target value, the training error meets the requirements, and the fault diagnosis accuracy reaches over 90%. It has been proven that the diagnostic performance of the system is good, achieving the expected design effect.

Key words: Substation, Inspection robot, Fault diagnosis, Intelligent algorithms

1. Introduction. Ensuring operation and maintenance production, as well as maintaining the safety of the power grid, is the top priority in power production work. Equipment inspection is an important part of operation and maintenance production. Conducting regular equipment inspections and inspections of substations, mastering equipment status, and promptly identifying and eliminating equipment hazards are important tasks for achieving safe, stable, and fault free operation of substations [1]. In recent years, intelligent inspection robots for substations have been widely installed in ultra-high voltage and intelligent stations. Robots are used to cooperate with or even replace operation and maintenance personnel in daily inspection work, constantly detecting the status of substation equipment. Taking ultra-high voltage substations as an example, two outdoor mobile intelligent inspection robots are equipped, responsible for the inspection of 1000kV GIS, main transformers, and high impedance equipment, as well as the inspection of 500kV GIS and 110kV equipment [2]. They can perform daily work such as red ginseng 'I-N temperature, meter reading, oil level, etc, and through preset threshold comparison, timely indicate the general, serious, and critical defects of the equipment. However, the current intelligent inspection robot system cannot automatically search for and analyze the types of faults in hidden equipment, and can only generate reports on devices whose data has exceeded the threshold. Moreover, it is currently difficult for robots to identify equipment appearance defects, in the process of equipment status evaluation, many equipment appearance damage, oil leakage and other defect information can only be obtained through manual inspection and entered into the production system [3]. In order to make greater use of intelligent inspection robots as a powerful tool, deepen the application of robot backend, and improve the efficiency of intelligent inspection robots in substation operation and maintenance work, a data processing and feedback system is designed, integrating data from the production system, online monitoring, and robot control backend, and analyzing and processing the data, automatically evaluating the status of equipment, determining the type and location of fault hazards, searching for inspection points related to faults, developing the best inspection strategy, and achieving intelligent inspection robots to independently strengthen special inspections of hidden equipment, is very challenging and feasible [4].

*Chuzhou Polytechnic, Anhui, 239000, China (Corresponding author's e-mail: FangYao37@163.com)

†Anhui University of Technology, Anhui, 243000, China (LibinTan6@126.com)

2. References. At present, the development of power equipment towards high power, high reliability, and high intelligence has increased the difficulty of daily operation, maintenance, and testing. In the trend of unmanned substations, traditional inspection methods and fault diagnosis technologies are increasingly difficult to meet the needs of complex equipment diagnosis. The traditional inspection work of substation equipment mainly relies on regular inspections by operation and maintenance personnel and infrared temperature measurement. However, due to the influence of the experience and technical level of the inspection personnel, there is often a phenomenon of missed testing [5]. At the same time, using existing testing instruments makes it difficult for testing personnel to centrally manage data, resulting in low efficiency in deep mining of historical data, which greatly restricts the development of live detection technology. The research and application of intelligent inspection robots in substations have brought new solutions to the above-mentioned problems, providing a foundation for timely, effective, comprehensive, and intelligent diagnosis and maintenance of power equipment. Many scholars at home and abroad have conducted research on intelligent inspection robots for substations [6,7].

Liao, X will use OCR technology to improve the anomaly recognition system for detecting robot equipment. Based on the collection of video information, DSP comprehensive information processing is carried out, and the detection information is analyzed using frequency domain filtering methods through the human-computer interaction interface. At the same time, pattern recognition methods were used to extract the main component features of substation detection components, and a resolution model for similar features in video surveillance images was constructed [8]. Traditional cage inspections require divers to complete, which is inefficient and dangerous. Underwater robot detection is a method to solve this problem. When the robot is in motion, the camera captures the mesh cage, replacing manual inspection. Wei, Y proposed a hybrid control strategy based on neural networks (NN) and proportional integral differential (PID) for underwater three-dimensional path tracking, overcoming the drawback of traditional feedback regulation that can only work after deviations occur [9]. The purpose of Jiang, C is to demonstrate a multi-purpose detection robot that can walk on the ground and climb on power poles. The structure design, size optimization, Kinematics analysis, experiment and algorithm of the robot are introduced. The robot consists of three adjustable modules and a series connected two degree of freedom parallel mechanism. The wheel finger mechanism of each module can open and close the wheel finger to achieve rapid movement and obstacle crossing [10].

From the above analysis, it can be seen that the current research on intelligent inspection robots for substations has certain advantages compared to traditional manual inspection methods, but they still cannot meet the requirements of automatic removal of faulty parts and foreign object removal of equipment. Their performance in autonomous tracking and intelligent diagnosis analysis also needs to be improved. In addition, most of the intelligent inspection robots mentioned above use a single grid charging method, which is not conducive to the long-term inspection work of the intelligent inspection robots, especially when monitoring key equipment point-to-point for a long time, it cannot ensure sufficient electricity.

3. Application of Robot Intelligent Inspection Technology in Fault and Defect Detection of Substation.

3.1. Intelligent inspection robot. Robot technology is a strategic technology industry in China, related to a series of cutting-edge technologies such as automatic control, image recognition, and intelligent learning. According to the "Made in China 2025" plan, industrial robots will be selected as one of the ten key fields to promote epoch-making development, promote robot standardization and modularization, expand market applications, and effectively promote the growth of the emerging robot market [11]. With the development of the times, the lack of human resources and the requirements of refined operation and maintenance of electrical equipment, intelligent inspection robots in substations are increasingly valued. This will be widely used for intelligent inspection of transmission equipment, real-time evaluation and auxiliary decision-making of power grid operation status, informatization, automation, and the establishment of interactive intelligent networks.

The intelligent intelligent inspection robot is equipped with intelligent detection equipment such as high-quality visual light cameras, infrared imaging devices, high-definition photography heads, environmental monitoring sensors, and intelligent analysis algorithm software [12]. It completes the management and control circuit of fast data collection and real-time information transmission, intelligent analysis and early warning decision feedback, replacing manual detection, achieving automatic detection and intelligent analysis of the

status of power equipment, and improving the quality of power equipment, research on the reliability of power grid and power equipment operation, and the use of power intelligent inspection robots is an important means of realizing the intelligence of power grid, and also an important direction for the development of future smart grids.

3.2. Intelligent algorithms. With the rapid development of the era of artificial intelligence and Big data, many industries are also following the development form of "machines replacing people". It is mainly divided into two application fields: Machine vision intelligent algorithm for fault tree and Big data analysis algorithm for multidimensional heterogeneous data [13].

(1) *Intelligent Algorithm for Machine Vision.* In recent years, with the deepening of industrial restructuring and the structural transformation and upgrading of modern manufacturing, more and more enterprises have implemented the "robot strategy". The application of robots in fields such as automobiles, logistics, aerospace, and even food has become increasingly widespread, driving the development of related industries.

Machine vision is a system that automatically obtains target images, analyzes and processes image features, analyzes results, obtains target knowledge, and makes decisions. Moving object testing technology is one of the functions of machine vision systems, this is the process of serializing image change regions and extracting moving targets from background images. The main purpose of machine vision research is to provide convincing data sources for subsequent object extraction and tracking in image arrangement compared to camera moving targets. Machine vision algorithms generally target specific application scenarios, there is currently no universal algorithm applicable to any situation. That is to say, all machine vision algorithms have their own applicability [14].

(2) *Big data analysis and processing intelligence.* The field of algorithm Big data involves a wide range. It deepens the Big data that occurs in the industrial field. With the deep integration of informatization and industrialization, information technology has penetrated into all parts of the industrial chain of various industries. For example, bar codes, two-dimensional codes, communication and identification, industrial sensors, industrial automatic control systems, industrial networks, etc., enterprise resource planning, Computer-aided design, Computer-aided manufacturing, Computer-aided engineering, etc. are widely used in enterprises. The application of next-generation information technologies such as the Internet, mobile Internet, and Internet of Things in the industrial field has brought enterprises into a new stage of development, and data is becoming increasingly abundant, especially in manufacturing enterprises where production lines are running at high speeds and a large amount of data is generated in industrial equipment. Models and algorithms are the two core issues of Big data analysis. The research on Big data analysis models can be divided into three levels [15]. Descriptive analysis, predictive analysis, and normative analysis. Descriptive analysis is the analysis and exploration of historical data, explaining what has happened. This stage includes discovering a set of data rules, mining related rules, describing model discovery, and visual analysis of data rules. Predictive analysis is used to predict future probabilities and trends.

3.3. Intelligent inspection fault diagnosis system. The intelligent patrol fault detection system uses fault tree vision algorithm, Big data analysis technology and intelligent patrol robot with high-definition camera. Through the fault tree, the image information collected by the HD camera carried by the intelligent patrol robot is fused with the data information collected by various sensors, and then through the Big data analysis algorithm, the real-time intelligent fault detection and analysis of all equipment in the substation is carried out. The overall diagram of the intelligent inspection fault detection system is shown in Figure 3.1.

(1) *Using Fault Tree Machine Vision Algorithm to Determine Fault Information.* Machine vision involves related technologies such as optical imaging, visual information processing, artificial intelligence, and mechatronics. It is a necessary technology for many highly automated industries to achieve intelligence. Machine vision technology has a series of advantages such as high accuracy and strong real-time efficiency, and is one of the important driving forces for intelligent robots. With the continuous improvement of various technologies and the increasing demand for high-quality products in the manufacturing industry, image processing has been mainly used for industrial electronic assembly error detection and is gradually applied in manufacturing, monitoring, visual navigation, communication and other applications. Therefore, studying imaging technology is of great significance for promoting the industrial development of intelligent industrial robots [16].



Fig. 3.1: Block diagram of fault detection system of intelligent inspection robot

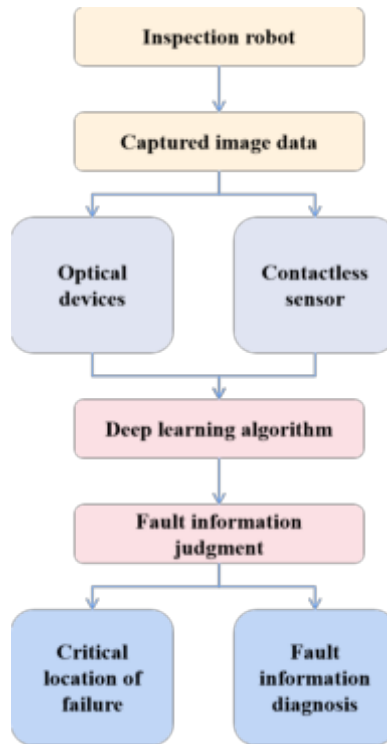


Fig. 3.2: Schematic diagram of fault position determination by intelligent inspection system

The organic combination of fault tree machine vision algorithm and intelligent inspection robot can enable the intelligent inspection robot to flexibly and intelligently locate the key positions of all equipment faults during substation inspection, ensuring that maintenance personnel can timely maintain and handle the key positions where faults occur (Figure 3.2).

(2) *Use Big data to analyze fault information.* With the rapid development of the era of Big data and artificial intelligence, the organic integration of industrial automation and Big data and other technologies can promote the industry to move towards digitalization, intelligent transformation and integration with the era of Big data. The power generation system of all equipment in the substation is complex and highly centralized [17].

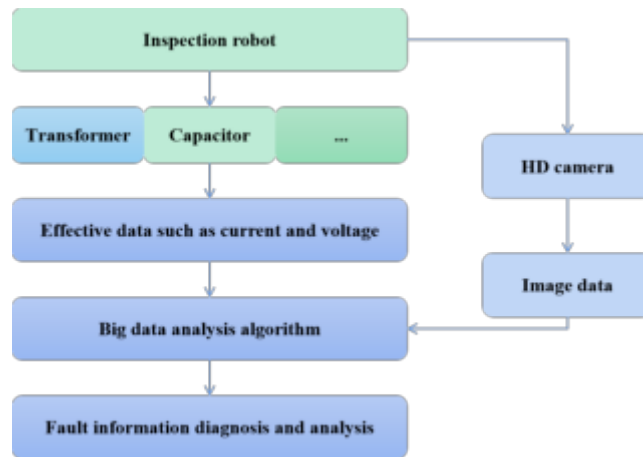


Fig. 3.3: Schematic diagram of fault information diagnosis of intelligent inspection system

Collect data transmitted by sensors on devices through intelligent inspection robots, and combine the results processed by machine vision. Through Big data analysis technology, targeted algorithms are adopted to establish various data mining models and equipment analysis models for substations, and real-time early warning and diagnosis of substation equipment faults and operation modes (Figure 3.3). The intelligent patrol robot combines the image data collected by machine vision technology with the operation data (such as real-time current and voltage data) of transformers, high-voltage circuit breakers, disconnectors, capacitors, reactors and other equipment in the substation to generate a large number of real-time effective data. After Big data analysis algorithm, it can accurately analyze whether there is fault information of the equipment at the moment. If there is any fault information, relevant equipment maintenance personnel can be notified in a timely manner through the early warning system of the intelligent inspection robot [18].

3.4. Fault Tree Diagnosis Principle. The model consists of three layers: input layer, intermediate layer, and output layer. During fault diagnosis, the data decision table is first trained as the training sample of the fault tree to obtain their respective connection weights and thresholds, and then the corresponding connection weights are stored to form a knowledge base. Finally, the trained fault Tree model model is used for fault location and diagnosis. Before working on the fault tree, the first step is to establish a fault knowledge base based on experimental data and expert experience. In order to obtain initial data, the system uses hardware circuits to obtain radar detection signals, and then uses fault trees for shallow empirical reasoning [19]. Then, fault diagnosis is carried out by combining fault trees with expert systems. The network inputs the fault phenomena of the diagnosed object, and the network outputs the probability of the diagnosed object's failure. When constructing the model, the number of nodes in each layer of the fault tree is mainly set based on the empirical formula of previous radar faults, and adjusted based on the training results.

The input layer implements nonlinear mapping from $x \rightarrow \phi_i(x)$, and the output layer implements Linear map from $\phi_i(x) \rightarrow y_k$, namely:

$$y_k = \sum_j^k \omega_{kj} \phi_j(X) + \theta \quad (3.1)$$

$$j = 1, 2, \dots, h$$

In the formula, k is the number of output nodes; ω_{kj} is the output weight value; θ is the threshold value; $(x_1, x_2, \dots, x_n)^T$. The kernel function of the hidden layer node will produce a certain response to the input signal locally. When the input signal is close to the central range of the kernel function, the hidden layer node will produce a larger output. The kernel function often used is the Gaussian function. Fault tree analysis (FTA)

is a method to describe the causal relationship. It qualitatively describes the causal relationship between the layers of fault propagation. It can use the Minimum cut set to find possible fault sources, and is effectively applied to various complex analysis and diagnosis situations. This method applies certain decision conditions to conduct in-depth analysis of specific conditional states, revealing the relationship and correlation between conditions and events, and expressing them through graphical means. The fault tree graph can clearly list the association and logical relationship between the major faults and specific Glitch of the system [20].

The fault tree is established through the following steps:

1. Determine the top event. In the backend analysis system, it generally refers to the type of fault, which can be a large category of faults or specific faults.
2. Analyze the top event, screen the various reasons that trigger the top event, and associate these identified reasons with the top event through logical gates, forming the upper input of the top event.
3. Analyze the causes of the top events, decompose these events again, and identify their input events.
4. Repeat the calculation layer by layer until it can not progress again, that is, get the bottom event, and build and complete the fault Tree model.

Fault tree is a powerful tool for establishing correlations between data and mining causal relationships, but it is difficult to automatically search for relevant knowledge through a large and complex substation operation and maintenance database, the workload of building fault trees through manual experience is too huge. So it is necessary to introduce rough set technology to obtain knowledge of data classification and attribute association for constructing fault trees, which can be used to conveniently construct fault trees.

If there is a bottom event $B_i(i = 1, 2, \dots, n)$ and its state is $x_i(t)$ at a certain time, then:

$$x_i = \begin{cases} 1, \text{The bottom event occurs at time } t \\ 0, \text{The bottom event did not occur at time } t \end{cases} \quad (3.2)$$

The probability of triggering the bottom event at this time is:

$$p_i(t) = E[x_i(t)] = p[x_i(t) = 1] \quad (3.3)$$

If the top event in the fault tree is triggered as M, and its state is M [X (t)] at a certain time, then

$$M[X(t)]_i \begin{cases} 1, \text{The bottom event occurs at time } t \\ 0, \text{The bottom event did not occur at time } t \end{cases} \quad (3.4)$$

And the probability of M at t is

$$p_1 = E\{M[X(T)]\} = p\{M[X(T)] = 1\} \quad (3.5)$$

Converting the fault tree into a structural function can facilitate data calculation and correlation analysis. If the gates and the three OR gates in the fault tree are T1T2T3T4, respectively,then

$$T_4 = B_5 + B_6 \quad (3.6)$$

$$T_3 = B_3 + B_4 \quad (3.7)$$

$$T_2 = B_1 + B_2 \quad (3.8)$$

$$T_1 = T_2T_3T_4 \quad (3.9)$$

Indicates the likelihood of an event occurring, including:

$$p_{and} = \prod_{i=1}^n p_i \quad (3.10)$$

Table 4.1: Typical Fault State Model of P100 Unit

State model	Sample number
P101 board fault	001001
P102 board fault	010010
P103 board fault	101101
Azimuth drive fault	011011
High and low drive failure	110110
15 MHz clock failure	001011
PRF signal failure	100011
Equipment is normal	000000

Table 4.2: Typical fault state model of the P 200 unit

State model	Sample number
24 V power failure	000011
P201 board fault	000110
P202 board fault	001101
P203 board fault	011010
High voltage power supply 200 V fault	000111
Serial communication failure	100111
Equipment is normal	000000

$$p_{or} = 1 - \prod_{i=1}^n (1 - p_i) \quad (3.11)$$

In summary, the event probability can be conveniently calculated through the structure of the fault tree. The system can determine the probability of each bottom event based on the obtained event probability, and make fault judgments and equipment evaluation and maintenance strategies for the most likely bottom events. However, some faults have high importance and high risk factors, and the probability of them appearing in the bottom event is often very small. Therefore, the fault coefficient is set as $F_i(u) = I_i P_i$.

4. System Experiment Results and Analysis. The radar is composed of three independent unit modules: P100, P200, and P300. In order to verify the effectiveness of fault tree for fault diagnosis, the typical faults of the radar P100 unit are taken to establish a sample training model, and the samples are initialized. The typical fault state model of the P100 unit is shown in Table 4.1.

In actual training, after 29 iterations, the training results can quickly converge to the target value, the training error meets the requirements, and the fault diagnosis accuracy reaches over 90%. In order to verify the fault diagnosis effect of the system on other units, the typical faults of the radar P200 unit are taken to establish a sample training model, and the samples are initialized. The typical fault state model of the P200 unit is shown in Table 4.2.

The number of input nodes in the fault tree is 7, the number of output nodes is 2, the number of center vectors is 14, the number of nodes in the basis function layer is 7, and the threshold of the basis function is set to 0.8257. In actual training, after 31 iterations, the training results can quickly converge to the target value, the training error meets the requirements, and the fault diagnosis accuracy reaches over 90% [21]. Each radar unit was divided into 25 sets of fault and non fault samples, with a total of 225 fault samples to test the diagnostic performance of the system. The diagnostic results are shown in Figure 4.1.

From the diagnostic results data, it can be seen that the fault diagnosis accuracy of all three units is above 90%, indicating that the diagnostic performance of the system is good and meets the expected design effect [22].

5. Conclusion. This article is based on the fact that current technologies such as fault detection and diagnosis for substation equipment have not yet entered full intelligence. By combining the practical application

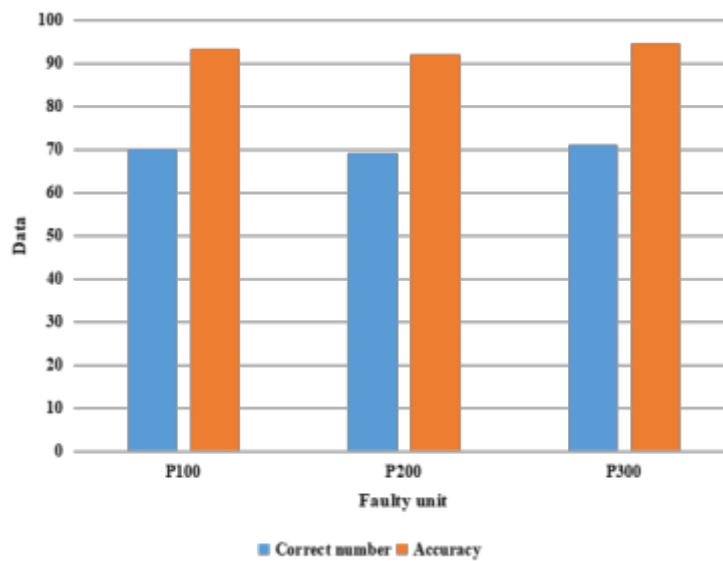


Fig. 4.1: Troubleshooting results

environment characteristics of substations, an intelligent intelligent inspection robot autonomous monitoring and fault diagnosis detection system is proposed. The fault tree, Big data information analysis technology and carrying high-definition camera are organically combined, the fault tree model model is used to diagnose the fault, and a large amount of simulation data is used to provide experimental samples for the fault tree.

6. Acknowledgement. Anhui Provincial Department of Education University Scientific Research Project: Transfer learning based defect detection method for substation with data scarcity.

REFERENCES

- [1] Wang, Y., Mai, Y., & Wen, W. (2022). Research on path planning algorithm of intelligent inspection robot. *Journal of Physics: Conference Series*, 2181(1), 012006-.
- [2] Su, L., Yang, X., Cao, B., Wang, Y., Li, X., & Lu, W. (2021). Development and application of substation intelligent inspection robot supporting deep learning accelerating. *Journal of Physics: Conference Series*, 1754(1), 012170-.
- [3] Gan, X., Geng, X., Xiong, Z., Wu, Z., Du, S., & Gao, Y., et al. (2021). Application of 5g communication technology on intelligent inspection in 750kv substation. *Journal of Physics: Conference Series*, 1983(1), 012089 (8pp).
- [4] Luo, L., Ma, R., Li, Y., Yang, F., & Qiu, Z. (2021). Image recognition technology with its application in defect detection and diagnosis analysis of substation equipment. *Hindawi Limited*,87(7),987-989.
- [5] Han, S., Yang, F., Jiang, H., Yang, G., & Wang, D. (2021). A smart thermography camera and application in the diagnosis of electrical equipment. *IEEE Transactions on Instrumentation and Measurement*, PP(99), 1-1.
- [6] Yao, Y., & Li, S. (2022). Design and analysis of intelligent robot based on internet of things technology. *Computational intelligence and neuroscience*, 2022(1), 7304180.
- [7] A, Y. C., & Envelope, J. W. B. (2022). Application of measuring intelligent robot in building deformation monitoring. *Procedia Computer Science*, 208(9), 206-210.
- [8] Liao, X., Xie, K., & Qiu, Z. (2021). Joint inspection of hd video and robot in substation based on ocr technology. *Mobile Information Systems*,74(5),96-102.
- [9] Wei, Y., An, D., Liu, J., Wu, Y., Li, W., & Wei, Q., et al. (2022). Intelligent control method of underwater inspection robot in netcage. *Aquaculture Research*, 53(5), 1928-1938.
- [10] Jiang, C., Ye, C., Zang, Y., & Yu, S. (2022). Structure design and optimization of ground moving and pole climbing inspection robot. *Assembly Automation*,85(42-2),97-102.
- [11] Zhao, M., Mao, Y., Hen, Q., & Zhou, Y. (2021). Research on problems and countermeasures in the application of substation intelligent inspection system. *Journal of Physics: Conference Series*, 1983(1), 012084 (7pp).
- [12] Jia, X., Yuan, W., Li, H., Jiang, S., & Zhang, Y. (2021). Application of environment-perception intelligent control technology in the inspection robot of coal conveyance corridor in thermal power plant. *IOP Conference Series: Earth and Environmental Science*, 772(1), 012056 (6pp).

- [13] Chen, N., & Wang, Y. (2021). Design and collaborative operation of multimobile inspection robots in smart microgrids. *Complexity*, 2021(11), 1-11.
- [14] Sun, T., Ye, L., Xie, J., & Fan, H. (2021). Research and application of substation cable trench inspection robot communication system. *Journal of Physics: Conference Series*, 741(3), 96-105.
- [15] Zhang, S., Zhang, Y., Cao, S., Li, B., Qi, X., & Li, S. (2022). Design and application of intelligent patrol system in substation. *Journal of Physics: Conference Series*, 2237(1), 012017-.
- [16] Zhang, K., Tan, L., Chen, S., & Zhang, D. (2021). Research on intelligent operation and maintenance technology of primary equipment in substation. *IOP Conference Series: Earth and Environmental Science*, 98(8), 48-56.
- [17] Zhou, H., & Gu, M. (2021). Application of neural network and computer in intelligent robot. *Journal of Physics: Conference Series*, 1881(3), 032028 (7pp).
- [18] Bauer, P., Schmitt, S., Dirr, J., Magaa, A., & Reinhart, G. (2022). Intelligent predetection of projected reference markers for robot-based inspection systems. *Production Engineering*, 16(5), 719-734.
- [19] Zu, W., Li, Z., & Nie, L. (2022). Research on the core algorithm of wireless charging technology for substation patrol robot based on electromagnetic resonance. *Nonlinear Optics, Quantum Optics*, 85(3/4), 55.
- [20] Yuan, C., Xiong, B., Li, X., Sang, X., & Kong, Q. (2022). A novel intelligent inspection robot with deep stereo vision for three-dimensional concrete damage detection and quantification. *Structural health monitoring*, 96(3), 21.
- [21] Pan, Q., Zhang, M., & Zhou, H. (2021). Application of augmented reality (ar) technology in power grid emergency training. *Journal of Physics: Conference Series*, 2074(1), 012095.
- [22] Zhao, Z., Yang, J., Xi, H., Wang, J., & Gao, B. (2021). Research on mobile terminal technology supporting intelligent maintenance of substation. *Journal of Physics: Conference Series*, 1802(3), 032139 (6pp).

Edited by: B. Nagaraj M.E

Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing

Received: Dec 25, 2023

Accepted: Mar 18, 2024



THE APPLICATION OF DEEP LEARNING INTELLIGENT ROBOTS IN THE DESIGN AND IMPLEMENTATION OF INFORMATION RETRIEVAL SYSTEMS

YUQI MIAO*

Abstract. Traditional information retrieval algorithms ignore user needs and are unable to obtain user gaze coordinates and gaze times, resulting in low retrieval accuracy. The author proposes a new interactive information retrieval algorithm for this purpose. Divide eye tracking technology evaluation indicators, visually process eye movement information, obtain user gaze coordinates and gaze time, and calculate the influence coefficients of each gaze area and each point in the area. Weighted visual words are accumulated to get a visual word list with the weight of the associated area, and visual word list and Rocchio algorithm are combined to build a hidden Relevance feedback retrieval model in semantic space to judge information retrieval preferences. Intelligent robot is introduced, and Jensen Shannon divergence is used to calculate the Kullback–Leibler divergence distance between the probability distributions of document sets, calculate the similarity matching, and complete the interactive information retrieval. The simulation results prove that, due to the introduction of intelligent robot strategy in this method, the amount of irrelevant retrieval information is reduced by matching user needs and retrieval results. Therefore, the amount of messages generated by information retrieval is significantly lower than that of the two literature methods, without adding additional network load. The network system is in a stable operation state, and users can quickly grasp their own required information, and the retrieval speed is also improved to a certain extent. It is proved that the proposed algorithm has high retrieval accuracy, can effectively reduce network load and achieve high-quality human-computer interactive information retrieval.

Key words: Intelligent robots, Information retrieval, System design and implementation, eye tracking

1. Introduction. Big data is actually a multi information fusion, through sorting and summarizing effective information, a large amount of effective data information is further processed and analyzed, and application information in the data is extracted, through extensive data processing, effective analysis can be conducted on various issues [1]. Big data has the following four characteristics: Firstly, there is a trend of diversification in data types, with uncertain data sources and a wide range of data types; Second, the storage space of Big data is very large, and the capacity often exceeds 10000GB; Third, Big data has high requirements for the authenticity of data, and it also needs data with a certain degree of real-time [2]. Fourth, the data structure of Big data is very complex, and a single storage mode cannot meet the needs of Big data. Gradually increasing and becoming more complex, artificial intelligence, as an emerging development discipline, cannot develop without the support of internet technology and communication devices [3]. Moreover, artificial intelligence can process information quickly and accurately, which humans cannot achieve. Artificial intelligence technology is also an extension and expansion of internet technology, especially in the application of artificial intelligence in information retrieval, which further increases the connection with the internet. However, there is also a certain mutual restriction relationship between artificial intelligence systems and the internet. In short, on the basis of Internet technology, AI technology can be effectively applied to information retrieval in the context of Big data [4].

After hundreds of years of research and exploration, artificial intelligence has also gained new characteristics and significance of the times. Given the rich scientific and technological knowledge contained in artificial intelligence technology, it is a very complex technical task and also involves psychology [5]. After completing the imitation of humans, artificial intelligence efficiently and accurately completes information retrieval work, greatly increasing work efficiency and social production efficiency. However, in the context of Big data, the work of artificial intelligence is becoming more and more difficult, and the technical characteristics of artificial intelligence should also make corresponding adjustments and changes to meet the current characteristics of the times.

*Hefei preschool education college (YuqiMiao7@126.com)

2. References. With the wide application of modern communication and network technologies such as network communication, Internet of Things technology and cloud computing technology, various structured and unstructured data resources have shown explosive growth, marking a new stage of the Big data era. And text data dominates all forms of information resources, with a large amount of data presented in the form of text [6]. As the main manifestation of internet information, massive amounts of text information have become a key research object in the fields of computer science and intelligence. Therefore, text processing technology is the key way to use information in the era of Big data, in which Text retrieval is an important foundation and premise of text processing technology. How to accurately and comprehensively retrieve the information required by readers in massive text information is the key issue and research hotspot of Text retrieval technology development [7].

Xu, Y provide a method for manufacturing magnetic nanorobots, which is an intelligent robot system updated from traditional autonomous experimental platforms. Nanorobots synthesized by robots have uniformly sized samples, which can significantly reduce time costs [8]. Rahimi, T introduced a new topology structure of the proposed converter, which has the following advantages: (i) the topology structure of the converter is based on traditional boost and buck boost converters, which leads to its simplicity; (ii) The voltage gain of the converter provides a higher value through the lower value of the duty cycle; (iii) Due to the use of efficient traditional topology in its structure, the efficiency of the converter remains high for a large duty cycle interval; (iv) In addition to the increase in voltage gain, the current/voltage stress of semiconductors remains at a relatively low level; (v) The continuous input current of the converter reduces the current stress of the capacitor in the input filter [9]. Zeinoddini Meymand, H designed and implemented a speed control controller for permanent magnet synchronous motors based on an intelligent neural network. Firstly, an accurate mathematical model of permanent magnet synchronous motor was presented, and then, by designing a controller, we applied the challenge of wind turbine simulation. The designed controller was first implemented on the Arm Cortex-M microcontroller and tested on laboratory PMSM [10]. In order to solve the application problems of the traditional methods mentioned above, the author proposes an interactive information retrieval algorithm based on intelligent robots. Combining eye tracking technology, divide eye tracking technology evaluation indicators, visually process eye movement information, and obtain user gaze coordinates and gaze time. And the author first creates four eye movement evaluation indicators: gaze, scanning, pupil dilation, and scanning path.

3. Interactive Information Retrieval. Incorporating user behavior into the retrieval system can effectively achieve human-machine interaction for information retrieval. At present, Relevance feedback has two modes: explicit Relevance feedback and implicit Relevance feedback. Displaying Relevance feedback requires users to make a lot of preparations and inform users of the impact of their behavior on information retrieval in advance; In the implicit Relevance feedback mode, users do not need to consider the impact of their own behavior on the search results, but only need to pay attention to whether the search behavior meets their own needs, which can greatly reduce the workload of users, and the accuracy of the search results is also high [11].

3.1. Classification of eye tracking feature indicators. Eye trackers are tools for implementing eye tracking technology, which can be classified into three types: Helmet mounted eye trackers, desktop eye trackers, and eyeglass eye trackers. Eye tracking is divided into four categories: Gaze, scanning, pupil dilation, and scanning path. Gazing indicates the length of time the eyes stay at a fixed point; Scanning refers to the rapid movement or delay of the eyes between fixation points; Pupil dilation is used to describe the level of interest of users when browsing information; The scanning path is a trajectory formed by the rapid movement of both eyes between fixation points [12].

3.2. Implicit Relevance feedback of search page. Use multiple circles to describe the range of fixation points, with the diameter of the circle indicating the fixation time and the connecting line indicating the fixation trajectory. For each user's interest region, the region's fixation time is represented as

$$FD(i) = \sum_{e \in AOI(i)} T(e) \quad (3.1)$$

In the formula, e represents a fixation event, $T(e)$ is the user's fixation time for event e , and i is the index of the region of interest (AOI). The corresponding coordinates of the fixation point in the region of interest are

$$\begin{aligned} FiA_x(j) &= F_x(j) - AOI_{x1}(i), F(j) \in AOI(i) \\ FiA_y(j) &= F_y(j) - AOI_{y1}(i), F(j) \in AOI(i) \end{aligned} \tag{3.2}$$

In the formula, $AOI_{x1}(i)$ represents the x-coordinate of the upper left corner of the region of interest, and $AOI_{y1}(i)$ represents the y-coordinate of the upper left corner of the region of interest. The influence area of each user's fixation point is

$$\begin{cases} FiA_x(j) - r \leq IA_x(j) \leq FiA_x(j) + r \\ FiA_y(j) - r \leq IA_y(j) \leq FiA_y(j) + r \end{cases} \tag{3.3}$$

In the equation, r is the radius of influence. The calculation process is as follows

$$r = p \cdot F_{time}(j) \tag{3.4}$$

In the formula, p represents the regulatory factor, and $F_{time}(j)$ represents the fixation time of the fixation point.

Set a fixation threshold t , if the fixation time of a user's interest area is higher than t , the information relative to this area is considered as related information, and vice versa is considered as unrelated images. Express the measurement criteria for evaluating user interest as

$$M(i) = \begin{cases} 1, if FD(i) \geq t \\ 0, if FD(i) < t \end{cases} \tag{3.5}$$

According to the user's fixation time for different information, if the information correlation degree $h(i)$ is specified, the coupling relationship between fixation time and correlation degree is

$$k(i) = \frac{FD(i)}{\sum_{M(i)=1} FD(i)} \tag{3.6}$$

Based on the above information, a fixation point influence area can be obtained, and the size of this influence area is proportional to the fixation time. Record the influence coefficients of each point in the affected area as

$$IF(x, y) = e^{-\frac{((x - F_x)^2 + (y - F_y)^2)}{2\delta^2}} \tag{3.7}$$

Based on the initial search results viewed by the user, calculate the influence coefficients of each fixation point's influence area and each point in the area. Extracting visual words from each region, weighting and accumulating visual words, can obtain a visual word list that covers all associated regions and includes weights. The visual word list is the expression form of semantic space

$$word = \sum_{i \in FiA} word(i) \cdot IF \tag{3.8}$$

In order to obtain more accurate user retrieval preferences, relevant information is reordered, and the reordering process can be seen as the process of forming a visual word list of user retrieval intentions, as shown in Figure 3.1. If there are M related regions, the initial visual word list for each related region is

$$G(i) = (w_1, w_2, \dots, w_c) \tag{3.9}$$

In the equation, $G(j)$ represents the visual vocabulary, and w_n represents the vocabulary in the vocabulary [13]. The weight $WA(i)$ of each region of interest is

$$WA(i) = \frac{FD(i)}{\sum_i FD(i)} \tag{3.10}$$

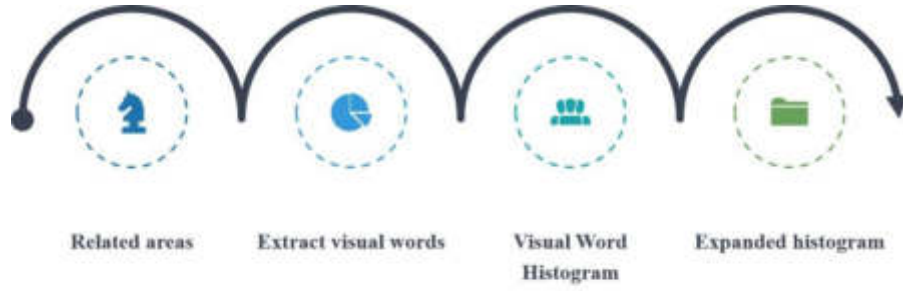


Fig. 3.1: Step of the implicit correlation feedback algorithm on the retrieval page

The improved visual word list for the relevant areas is

$$G^e(i) = (w_1^e, w_2^e, \dots, w_i^e, \dots, w_c^e) \quad (3.11)$$

where

$$w_i^e = WA(n) \cdot F_{time}(m) \cdot IF \quad (3.12)$$

In the formula, $WA(n)$ represents the range of interest of the relevant region, and $F_{time}(m)$ represents the corresponding fixation time of the relevant region. The expanded search visual word histogram H_i^e is

$$H_i^e = \sum_{j=0}^{M-1} G^e(j) \quad (3.13)$$

After using the above process to obtain new visual words, the Rocchio algorithm is integrated, and the implicit Relevance feedback retrieval model in semantic space is recorded as

$$\vec{q} = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in |D_r|} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in |D_{nr}|} \vec{d}_j \quad (3.14)$$

In the formula, \vec{q}_0 represents the user's initial search vector, D_r , D_{nr} represent the set of known related and unrelated search contents, and α, β, γ is the corresponding weight [14]. During information retrieval, the system needs to interact with users for many times, that is, it has multiple pages of implicit Relevance feedback, each feedback will generate a corresponding retrieval strategy, introduce new information vectors into the original retrieval vector, and eliminate+irrelevant vectors, thus improving Equation 3.14 to

$$\vec{q}_{m+1} = \alpha \vec{q}_m + \beta \frac{1}{|D_r|} \sum_{\vec{d}_j \in |D_r|} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{\vec{d}_j \in |D_{nr}|} \vec{d}_j \quad (3.15)$$

The search vector \vec{q}_{m+1} in Equation 3.15 is determined by the search vector \vec{q}_m during the m-th search and the related and unrelated search content vectors fed back in the macro results of this search.

3.3. Interactive Information Retrieval Algorithm Based on Requirements Mining . It can be seen from the implicit Relevance feedback model of equation (15) that each improvement of retrieval method is obtained on the premise of Relevance feedback of the previous retrieval results [15]. Requirement mining refers to starting from the real needs of users, the system judges their needs, and obtains the information they need. From content structure - spatial navigation construction - information content presentation, this series is expressed in an interactive logical form, as shown in Figure 3.2.

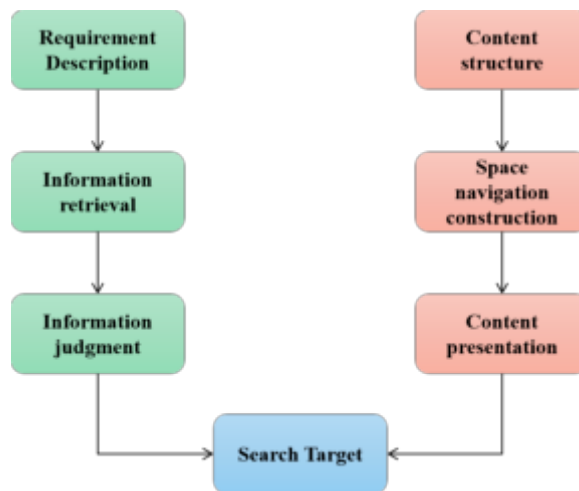


Fig. 3.2: Interactive information retrieval logical relationship

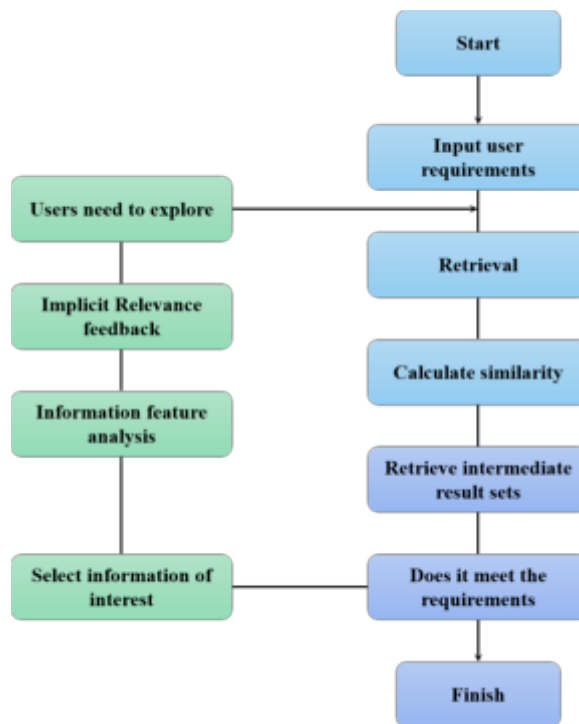


Fig. 3.3: Interactive information retrieval process based on demand mining

We introduce requirements mining conditions and design an interactive information retrieval process as shown in Figure 3.3.

In interactive retrieval systems, information retrieval refers to the similarity matching between retrieval vectors that describe information requirements and different document vectors within the system.

Currently, the cosine of the vector angle is widely used. The calculation formula for measuring the similarity

between two documents using this method is:

$$\cos(P, Q) = \frac{P \times Q}{|P| \times |Q|} = \frac{\sum_{i=1}^n \text{freq}(w_i|P)\text{freq}(w_i|Q)}{\sqrt{\sum_{i=1}^n \text{freq}(w_i|P)^2} \times \sqrt{\sum_{i=1}^n \text{freq}(w_i|Q)^2}} \quad (3.16)$$

In the formula, P and Q represent the vectors of two documents in sequence, $\text{freq}(w_i|P), \text{freq}(w_i|Q)$ represents the component in the vector, which is the frequency at which the user retrieves terms within this document.

However, in practical calculations, it has been found that the vector angle cosine method has a high computational complexity and cannot achieve fast retrieval targets. For this reason, the Jensen Shannon divergence method is used to compensate for its shortcomings. Calculate the Kullback–Leibler divergence distance between the probability distributions of two document sets, and determine the similarity between documents [16]. If the Kullback–Leibler divergence distance is shorter, the document similarity is greater, and vice versa. The derivation formula for Jensen Shannon divergence is

$$D_{js}(P||Q) = \frac{1}{2}D_{KL}(P||R) + \frac{1}{2}D_{KL}(Q||R) \quad (3.17)$$

$$R = \frac{1}{2}(P + Q) \quad (3.18)$$

where, D_{KL} represents the Kullback–Leibler divergence of P and Q probability distribution.

$$O = (o_1, o_2, \dots, o_n) \quad (3.19)$$

According to the Kullback–Leibler divergence theorem, design a probability vector o as in Equation 3.19, then the information entropy of the vector is

$$H(O) = - \sum_{i=1}^n o_i \cdot \lg o_i \quad (3.20)$$

Regarding the vocabulary set $W = \{w_1, w_2, \dots, w_n\}$, o_i can be used as , and the number of occurrences in the document, then

$$o_i = \frac{\text{freq}(w_i|o)}{\sum_{i=1}^n \text{freq}(w_i|o)} \quad (3.21)$$

If information entropy is used to describe Jensen Shannon divergence, Equation 3.17 can be transformed into

$$D_{js}(P||Q) = H(R) - \frac{1}{2}(H(P(w_i)) + H(Q(w_i))) \quad (3.22)$$

In the formula, H is the information entropy function, and R is the composite vector of P and Q. The author fully integrates the two strategies of implicit Relevance feedback and demand mining under eye tracking technology, calculates the matching degree of user needs and retrieval results by using Equation 3.22, and completes the intelligence and accuracy of information retrieval while effectively tracking user retrieval preferences.

3.4. Search engines. The fundamental idea of an intelligent search robot system is to combine two major internet applications, information retrieval and instant messaging. The most important and common manifestation for information retrieval is search engines, according to the intelligent search robot designed by the author, it also uses search engines to achieve the search characteristics of the system. In this section, we first provide a brief overview of search engines [17].

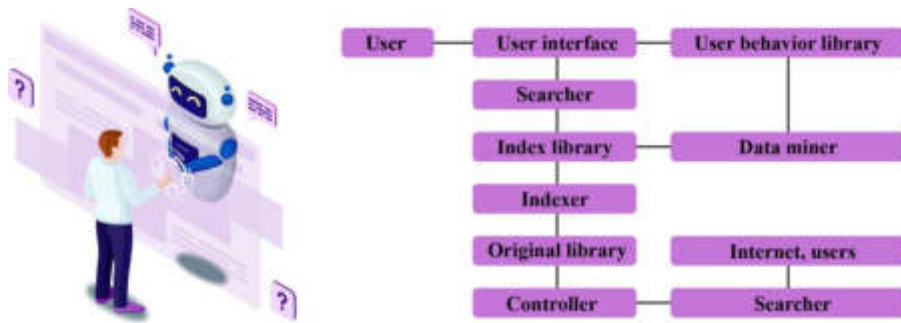


Fig. 3.4: Search engine architecture

Modern large-scale high-quality search engines generally use a three stage working mode for information collection, information processing, and query services.

Information gathering: Each independent search engine has an information gathering program, which is a web crawler for search engines. Web crawlers continuously crawl web pages along hyperlinks. In theory, starting from a certain range of web pages, they can collect the vast majority of web pages. In addition to web crawlers, users can also input information through robots for the search of this system.

Information processing: Search engines need to do a lot of processing work after collecting information in order to provide retrieval services. The most important thing is to extract keywords and establish index files. It should be pointed out that the processing of Chinese also requires the execution of Chinese word segmentation.

According to the above requirements and workflow, the architecture of the search engine is roughly shown in Figure 3.4.

4. Simulation research. In order to detect the true information retrieval performance of the proposed algorithm, simulation analysis was conducted, and traditional method 1 and traditional method 2 were compared. Using recall and precision indicators to measure the quality of a retrieval algorithm, recall represents the ratio of the number of relevant documents retrieved to the total number of relevant documents in the system document library, highlighting the comprehensiveness of the retrieval algorithm [18]. The calculation formula is

$$\text{Recall rate} = \frac{\text{Retrieved similar documents}}{\text{All detailed documents in the database}} \tag{4.1}$$

The precision ratio represents the ratio of the number of relevant documents retrieved to the total number of documents retrieved, highlighting the correctness of the retrieval algorithm. The calculation formula is

$$\text{Accuracy} = \frac{\text{Retrieved similar documents}}{\text{Retrieve all documents obtained}} \tag{4.2}$$

The comparison of recall and precision of the three methods is shown in Figure 4.1.

From Figure 4.1, it can be seen that when the recall rate is between 20% and 60%, the traditional method 1 and method 2 exhibit significant precision jitter. However, this method outperforms the other two methods in terms of precision as the recall rate gradually increases. This is because this method uses eye tracking technology to timely capture user retrieval preferences, and this interactive strategy can maximize the accuracy of information retrieval [19].

The message volume of the information retrieval process is the average message volume that meets each retrieval request. This indicator is used to verify the stability of the method’s operation, thereby reflecting the efficiency of the method’s retrieval. The comparison results of message volume simulation for the three methods of information retrieval process are shown in Figure 4.2.

Therefore, the amount of messages triggered by information retrieval is significantly lower than that of the two literature methods, without adding additional network load. The network system is in a stable operation

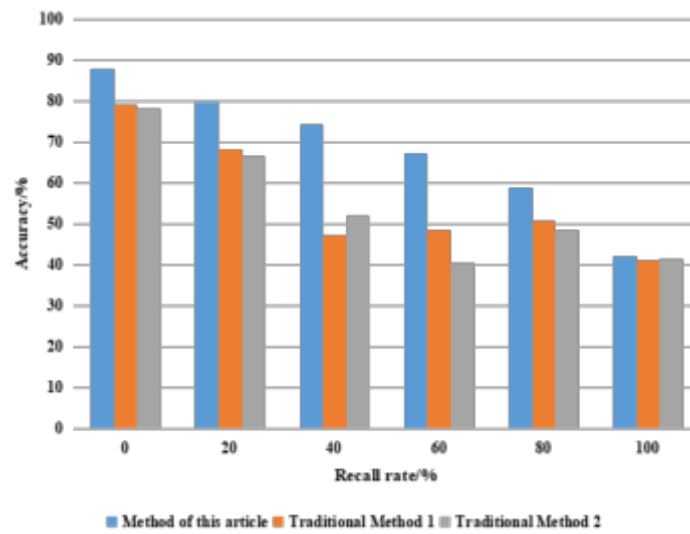


Fig. 4.1: Compare the recall and precision of the three methods

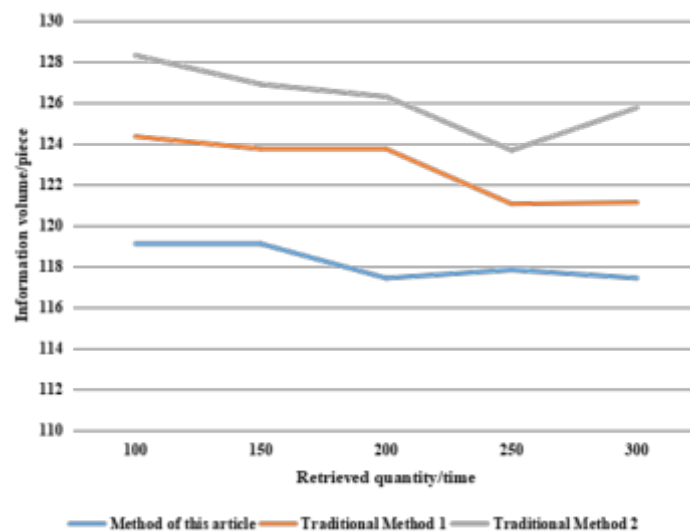


Fig. 4.2: Comparison of the message volume of the information retrieval process

state, and users can quickly grasp their own required information, and the retrieval speed is also improved to a certain extent [20].

5. Conclusion. In order to effectively improve the accuracy of interactive information retrieval and provide users with a better service experience, this study combines the theory of human gaze behavior and proposes a new interactive information retrieval algorithm. This method can concentrate on presenting cognitive features in the process of information retrieval, evaluate users' actual retrieval needs, and ultimately present an ideal human-machine interactive retrieval mode, bringing new exploration ideas for future research in the field of interactive information retrieval.

6. Acknowledgement. The study was supported by the Education Department of Anhui Province“‘The Case of ideological and political education in the course of information retrieval”, project number htytyldsr21)”

REFERENCES

- [1] Varlamov, O. (2021). "brains" for robots: application of the mivar expert systems for implementation of autonomous intelligent robots. *Big Data Research*, 25(ahead-of-print), 100241.
- [2] Guo, J. (2021). Research on the application of intelligent robots in explosive crime scenes. *International Journal of System Assurance Engineering and Management*, 14(2), 626-634.
- [3] Hosono, K., Maki, A., Watanabe, Y., Takada, H., & Sato, K. (2021). Implementation and evaluation of load balancing mechanism with multiple edge server cooperation for dynamic map system. *IEEE Transactions on Intelligent Transportation Systems*, PP(99), 1-11.
- [4] Zhu, L. (2021). Research on the design and application of ideological and political education platform in colleges and universities based on moodle. *Journal of Intelligent and Fuzzy Systems*,98(3), 1-8.
- [5] Bian, L., Zhang, J., Cui, Q., Chen, X., & Wang, S. (2021). Research on the realization and application of intelligent iot platform for electrical equipment under industrial internet. *Journal of Physics: Conference Series*, 1982(1), 012078.
- [6] Wei, H. H., Zhang, Y., Sun, X., Chen, J., & Li, S. (2023). Intelligent robots and human–robot collaboration in the construction industry: a review. *Journal of Intelligent Construction*, 18(8),9180002-9180002.
- [7] Santhanaraj, K. K., Ramya, M. M., & Dinakaran, D. (2021). A survey of assistive robots and systems for elderly care. *Journal of Enabling Technologies*, ahead-of-print(ahead-of-print),97(1),974-976.
- [8] Xu, Y., Gao, Y., Liu, R., Li, J., & Zhu, X. (2021). Robots built robots: nanorobots customized by intelligent robot. *Crystal Growth & Design*,84(5),749-754.
- [9] Rahimi, T., Islam, M. R., Gholizadeh, H., Mahdizadeh, S., & Afjei, E. (2021). Design and implementation of a high step-up dc-dc converter based on the conventional boost and buck-boost converters with high value of the efficiency suitable for renewable application. *Sustainability*, 13(5),99-103.
- [10] Zeinoddini-Meymand, H., Kamel, S., & Khan, B. (2022). Design and implementation of a novel intelligent strategy for the permanent magnet synchronous motor emulation. *Complexity*, 74(8),941-946.
- [11] Wisniewski, R. (2021). Design of petri net-based cyber-physical systems oriented on the implementation in field programmable gate arrays. *Energies*, 14(7),87-89.
- [12] Lv, X., & Li, M. (2021). Application and research of the intelligent management system based on internet of things technology in the era of big data. *Mobile Information Systems*, 2021(16), 1-6.
- [13] Jia, J., & He, Y. (2021). The design, implementation and pilot application of an intelligent online proctoring system for online exams. *Interactive Technology and Smart Education*, ahead-of-print(ahead-of-print),47(7),58-61.
- [14] Yang, Y., Xie, Y., Liu, J., Jiang, P., & Chen, Y. (2023). Self-pumping actuation module and its application in untethered soft robots. *Journal of Intelligent & Robotic Systems*, 108(2),854-865.
- [15] C, J. H. A., & B, B. P. (2022). Study the path planning of intelligent robots and the application of blockchain technology. *Energy Reports*, 8(7), 5235-5245.
- [16] A, L. F., A, X. L., A, C. G., & B, B. J. (2021). Path control of panoramic visual recognition for intelligent robots based-edge computing. *Computer Communications*,987(1),45-49.
- [17] Zhou, L., Wang, F., Wang, N., & Yuan, T. (2021). Application of industrial robots in automated production lines under the background of intelligent manufacturing. *Journal of Physics: Conference Series*, 1992(4), 042050.
- [18] Zhao, M., Mao, Y., Hen, Q., & Zhou, Y. (2021). Research on problems and countermeasures in the application of substation intelligent inspection system. *Journal of Physics: Conference Series*, 1983(1), 012084 (7pp).
- [19] Kandhofer, M., Steinbauer, G., Lassnig, J., Menzinger, M., & Szalay, I. (2021). Edlris: a european driving license for robots and intelligent systems. *KI - Künstliche Intelligenz*,74(3),98-103.
- [20] Yang, R., Mo, Q., Li, Y., Liu, Y., & Hu, R. (2021). Application of 3d vision intelligent calibration and imaging technology for industrial robots. *Journal of Physics: Conference Series*, 2082(1), 012004.

Edited by: B. Nagaraj M.E

Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing

Received: Dec 25, 2023

Accepted: Mar 18, 2024



APPLICATION OF MEASUREMENT ROBOTS BASED ON DEEP LEARNING IN BUILDING TILT STABILITY MONITORING

WEI ZHANG* AND JUNHUA LI†

Abstract. In order to better understand the application of measuring robots in monitoring the stability of building inclination, this paper proposes a deep learning based algorithm for analyzing the trend of building inclination using measuring robots. The author first proposes a method of using measuring robots to monitor the tilting stability of buildings. Based on the principle of laser ranging, construct a robot structure and laser ranging model to achieve information exchange between different units. Secondly, based on simulation analysis, the position relationship between the laser and the building was marked with triangular coordinates, and the inclination of the building was marked with robot benchmarks. The simulation process was designed. By using the forward crossing method to evaluate the accuracy of tilt monitoring and comprehensively monitoring the three-dimensional angle of free station setting, the problem of low monitoring accuracy in traditional methods has been effectively solved. Finally, the experimental results show that using this method, the accuracy of building tilt monitoring can reach 99%, which is 7% higher than traditional measurement methods. Due to the low efficiency, high cost, and low accuracy of traditional manual monitoring work, it can no longer meet the requirements of modern engineering measurement. Therefore, high-precision measurement robots are used for tilt stability monitoring. Compared with traditional monitoring, high-precision measurement robots can achieve high-precision, high-efficiency, and low-cost monitoring with faster speed, higher efficiency, and strong automation capabilities.

Key words: Measuring robots, building tilting, stability monitoring

1. Introduction. Monitoring the tilt stability of buildings is a highly technical task that requires regular, timed, and quantitative observation of buildings. By monitoring buildings, timely detection of displacement changes in buildings can be carried out in order to take effective measures. In recent years, with the increasing scale and quantity of construction projects, slope stability monitoring has gradually become an important task in engineering surveying. Due to the low efficiency, high cost, and low accuracy of traditional manual monitoring work, it can no longer meet the requirements of modern engineering measurement. Therefore, high-precision measurement robots are used for tilt stability monitoring. Compared with manual monitoring, high-precision measurement robots can achieve high-precision, high-efficiency, and low-cost monitoring with faster speed, higher efficiency, and strong automation capabilities. The design of a measurement robot includes a motor, controller, and angle measurement system. Measurement robots can be fixed on a planar coordinate measuring machine through a machine or robotic arm installed on the ground, and then control the motion of the robot's four robotic arms on the ground to measure according to the predetermined route and orientation. Adopting a dual turntable system for fast conversion. At the same time, the measurement robot uses Total station and electronic angle measuring system to realize three-dimensional tilt measurement of buildings. The Total station can directly obtain the three-dimensional coordinates and direction information of the measured object, and the angle measuring system can collect the two-dimensional coordinates of the measured object, and then convert them into three-dimensional coordinates. Install sensors on buildings to monitor them in real-time. Sensors can be installed at different locations in buildings, and data collection and analysis can be carried out through a data acquisition software system to achieve dynamic monitoring of buildings. The terminal consists of a data acquisition module, wireless communication module, power switch module, and SD card. It can communicate wirelessly with measurement robots to achieve data collection and transmission. The terminal also has an RS485 interface, which can be connected to an industrial computer to achieve remote control of

*Department of Construction Engineering, Hebei Vocational University of Industry and Technology, Shijiazhuang, Hebei, 050091, China (WeiZhang938@163.com)

†Basic course teaching Department, Hebei Vocational University of Industry and Technology, Shijiazhuang, Hebei, 050091, China (JunhuaLi5@126.com)



Fig. 1.1: Robot monitoring of building tilt stability

the measurement robot. At the same time, the terminal also has the characteristics of low power consumption, strong anti-interference ability, and long service life. This terminal sets up a wireless communication link between the measuring robot and the building to collect the inclination value of the building. At the same time, two data acquisition devices are also set up, one for collecting the coordinates of control points and reference points, and the other for measuring communication between robots and buildings [1,2] (Figure 1.1).

2. Literature Review. With the development of economy and society, the height and number of buildings are gradually increasing, the construction process of new buildings will inevitably cause changes in the stability of existing high-rise buildings, therefore, it is necessary to conduct Deformation monitoring on existing high-rise buildings around the site. The tilt of buildings will pose a threat to people's life safety. Therefore, the tilt stability monitoring of Tower block is also one of the main contents of Deformation monitoring. When monitoring the tilt stability of buildings, the measurement robot should be combined with the structure and surrounding environment of the building to avoid defects such as cracks and holes in the building walls. If there are cracks, holes, and other defects on the walls of a building, traditional repair methods should be used to repair them to ensure the safety of the building. If high-precision repair methods are used to repair it, it can shorten the repair time. Especially in recent years, with the development of high-tech technologies such as sensor technology, image processing technology, and communication technology, the performance of measurement robots has become increasingly high. In addition, with the continuous development of computer equipment such as electronics and computers, measurement robots have gradually shifted from traditional measurement to intelligent measurement.

Scholars have conducted some research on this issue. Giacoppo, G. A. proposed a method for measuring the inclination of buildings in mining subsidence areas based on point cloud feature line extraction, which can measure and analyze the subsidence trajectory through point cloud features. Although this method has certain applicability in tilt measurement of settlement buildings, laser point clouds are easily affected by ground obstacles and have low accuracy in tilt measurement of urban buildings [3]. Tian, Y. M. proposed a study on the dynamic deformation laws and internal forces of ground and buildings based on numerical analysis of Metro Line 3. By constructing a numerical analysis model, the impact of ground motion on buildings is studied, deformation data is collected, and inclination degree is measured. This method achieved the estimation of inclination measurement values for transportation buildings through model numerical analysis, but did not conduct building simulation analysis, and the reliability of the measurement results needs to be verified [4]. Zhou, T. proposed the application of inclination sensor in precision detection of parallel robots. This method utilizes inclination sensor to collect monitoring data and transmit it to the parallel robot system for measurement. The above three methods rely on traditional mechanical probes for measurement. Although the mechanical probe type measuring instrument is simple to use, its range of use is limited due to limitations caused by working principles, and human factors can cause large errors, which can easily lead to low accuracy of tilt monitoring [5]. In summary, the author proposes a method for monitoring the tilt stability of buildings using measuring robots, which can obtain various spatial dimensions of the building from all directions and achieve monitoring without contact, effectively improving measurement accuracy.

3. Research methods.

3.1. Monitoring of inclination stability of measuring robots.

3.1.1. Tilt stability monitoring process. The detailed process for monitoring tilt stability is as follows:

1. Monitoring points should be set up on the platform as needed, and control points should be placed in high-rise buildings. The selection of points should be firm and not affected by mining.
2. The auxiliary observation mark of the measuring robot is set at the stable position of the building, and a cloud platform point distribution map with equal inclination is drawn based on the obtained monitoring data [6,7].
3. In the adjustment calculation of monitoring data, the Centroid coordinates of the detection layer of the building are calculated based on the measured data; According to the accuracy requirements of servo Total station, the mean square error value of horizontal observation $m_0 = \pm 0.5$ is selected.
4. Benchmarks and observation points should be set up for monitoring the inclination of buildings through coordinate positioning.
5. For the same building tilt observation object, it is necessary to set two or more reference points at different positions within the observation range of the measuring robot.

In order to ensure the accuracy and stability of building tilt monitoring, it is necessary to select stable and long-term monitoring positions for stability inspection.

3.2. Simulation of robot benchmark labeling building inclination. Because the end effector of the measuring robot is a laser sensor, the laser sensor is connected to the robotic arm through 01, and the robotic arm of the measuring robot is simplified into a planar robotic arm model. The description of the position and posture of the laser sensor using coordinate annotation is more intuitive [8].

Let X be the laser sensor, therefore, the robot base coordinate system is Equation 3.1.

$$\begin{cases} x_e = l_1 \cos \theta_1 + l_3 \cos(\theta_1 + \theta_2) \phi + l_4 \cos(\theta_1 + \theta_2 + \theta_4) \\ y_e = l_1 \sin \theta_1 + l_3 \sin(\theta_1 + \theta_2) \phi + l_4 \sin(\theta_1 + \theta_2 + \theta_4) \end{cases} \quad (3.1)$$

The Kinematics parameters obtained by simulation are basically in line with the actual situation, which can be used to describe them and verify the theoretical feasibility of the proposed method. According to the tilt simulation process (as shown in Figure 3.1), complete the tilt simulation research of the robot benchmark labeling building.

3.3. Accuracy Appraisal of Building Tilt Monitoring. The servo Total station with high precision and strong automatic tracking ability is selected, and the front crossing mode is adopted to realize the three-dimensional accurate observation of building inclination observation. This method can effectively weaken the influence of the error of conventional observation methods on the observation accuracy. The calculation formula for inclination error (2) is:

$$i = \frac{\sqrt{(x_1 - x_2)^2 - (y_1 - y_2)^2}}{|z_1 z_2|} < 0.05 \quad (3.2)$$

In Equation 3.2, x_1 , y_1 , and z_1 respectively represent the coordinate errors of the upper half of the horizontal axis of the coordinate system; x_2 , y_2 , and z_2 respectively represent the coordinate errors in the lower half of the horizontal axis of the coordinate system. Select the set of data with the highest error in the coordinate system to evaluate the accuracy of building tilt monitoring. If the monitoring error is less than 0.05, it indicates that the monitoring results of this method are relatively accurate .

3.4. Precision Analysis of Monitoring Schemes.

3.4.1. 3D free station adjustment calculation model. Before conducting long-term monitoring, analyze the accuracy of this plan based on three types of observation values and monitoring methods. When conducting accuracy analysis based on the instrument accuracy and observation method used, the mean square error in horizontal direction observation is $m_y = \pm 0.5''$, and the mean square error in horizontal distance observation is $m_s = \pm 0.5mm$, with a vertical right angle observation error of $m_g = \pm 0.7''$ [9, 10].

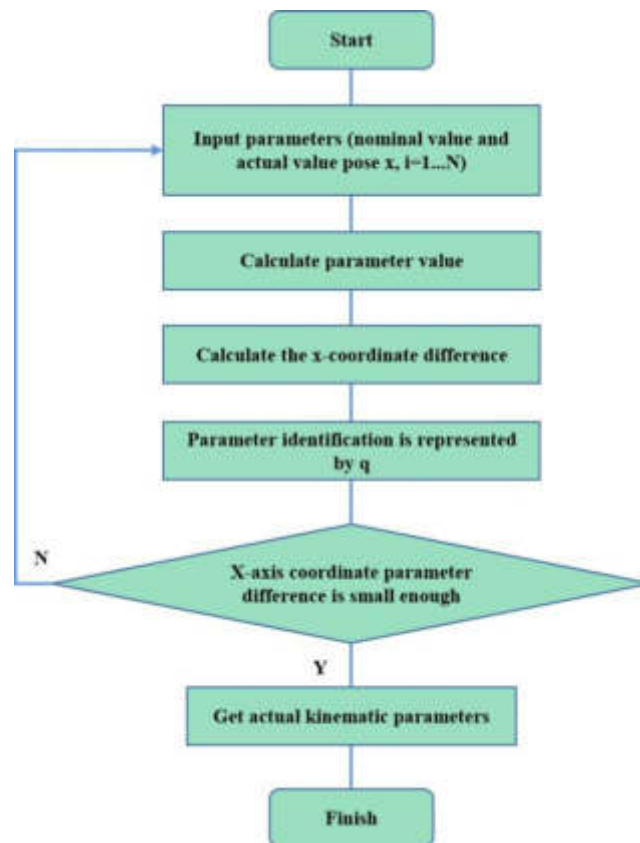


Fig. 3.1: Tilt Simulation Process

According to the principle of indirect adjustment, list the error equations for the three measurements in this scheme.

1. The error equation is shown in Equation 3.3

$$V = B\delta_x + L \quad (3.3)$$

In the equation:

- V - Correction matrix of observation values;
- B - coefficient matrix of observation error equation;
- gx - unknown number correction matrix;
- L-constant matrix of observation error equation.

2. The coefficient matrix formula of the normal Equation 3.4

$$N = B^T P B \quad (3.4)$$

3. Expected results of errors.

By writing VB program code to solve the coefficient matrix of the unknown number matrix, the expected error value is finally obtained. The expected error results are shown in Table 3.1.

From the expected results in Table 3.1, it can be seen that if the mean of two independent observations is taken as the final observation result, the measurement accuracy of the monitoring project will meet the requirement of $\leq 1\text{mm}$.

4. Actual measurement and data analysis

Table 3.1: Expected Results of Errors

Monitoring point number	Mean square error in X coordinate	Y coordinate mean square error	Mean square error of a point	Mean square error of elevation
S1	± 0.07	± 0.08	± 0.08	± 0.05
S2	± 0.07	± 0.07	± 0.09	± 0.06
1	± 0.19	± 0.39	± 0.43	± 0.16
2	± 0.19	± 0.39	± 0.42	± 0.21
3	± 0.23	± 0.23	± 0.47	± 0.15
4	± 0.34	± 0.21	± 0.45	± 0.14

Table 3.2: Results of Measured Data

1	X(m)	Y(m)	H(m)	type
2	936.25	963.45	148.36	undetermined
3	936.12	954.12	148312	undetermined
4	985.12	960.12	148.32	undetermined
S1	947.21	945.13	110.25	undetermined
S2	856.21	935.74	110.74	fixed

When observing on the actual site, the position of the measuring station remains fixed. Leica TM30 surveying robot is used to set up on S1 and s2 stations respectively for observation. Each station automatically observes 3 sets of measurements, and simultaneously observes the horizontal angle, distance and vertical angle. After collecting data from monitoring points, perform 3D adjustment model calculations, and perform overall rigorous adjustment calculations on all observation data [11,12]. The measured data is shown in Table 3.2.

3.5. Main parameters of measuring robots. The main parameters of the measuring robot include: Total station coordinate system, angle measuring accuracy, distance measuring accuracy and angle measuring speed. The coordinate system of Total station includes three coordinate systems, namely X, Y and Z direction coordinate systems, namely ground coordinate system, distance coordinate system and angle coordinate system. The distance measurement accuracy of Total station is 1 mm, and the angle measurement accuracy is 0.02° ; The ranging speed is 6-10m/s. Angle measurement accuracy refers to the ratio of the difference between the angles of Total station in two observation directions and the difference between the angles of these two observation directions, expressed in%. In practical applications, in order to ensure the accuracy of measurement, direct reading method and indirect reading method are generally used to calculate the coordinates of the measuring station. The direct reading method refers to directly reading out the coordinates of the measuring station; The indirect reading method refers to obtaining the coordinates of a given point by substituting the station coordinates into the known point coordinates, provided that the station coordinates are already known [13,14].

In the actual survey, the angle measurement accuracy needs to be corrected due to the deviation of Total station observation value. Generally, the error of station coordinates can be divided into two situations: One is the systematic error of station coordinates, which can be corrected by setting reference points; The other type is the system error of angle measurement accuracy, which mainly comes from the system error generated by the measuring robot itself, its size is related to the coordinate system error and angle measurement accuracy of the measuring station. In order to ensure the accuracy and reliability of measurement robots, the angle measurement accuracy is generally controlled within 1° . For high-precision monitoring projects, additional compensation is required.

3.5.1. Measurement method.

1. Turn on the power of the automated measurement robot and adjust the working temperature of the laser ranging sensor to $-20^\circ C \sim +40^\circ C$.

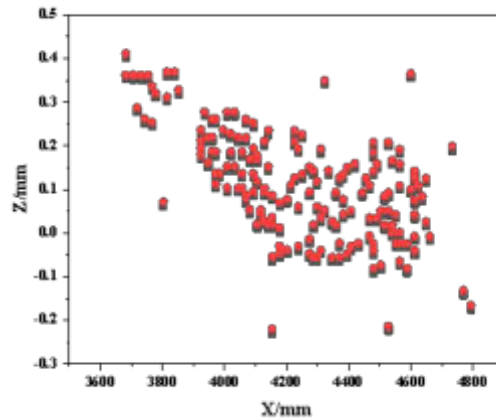


Fig. 4.1: 3D Display Results of Actual Building Inclination

2. Use a handheld computer to read the coordinates of the reference point, then place the measuring robot on the reference point and start the automatic tracking system.
3. Calculate the tilt angle and relative position of the building based on the sensor data on the measuring robot and the laser ranging sensor data.
4. Based on the laser ranging sensor data on the measuring robot and the coordinates of the measuring station, calculate the building tilt angle and relative position, and then compare the calculated building tilt angle and relative position with the reference point to calculate the error of the building tilt angle [15].

4. Experiments.

4.1. Experimental Environment. Taking a building as an example, through the method of establishing a model, the scanning angle between the coordinates of the measuring points and the light surface of each Launch pad is measured to obtain the required angle, so as to determine the relationship between the distance measurement problem functions of high-rise buildings; Utilize a large amount of random data to statistically analyze the coordinates of the measured point, and then use a three-dimensional discrete point cloud to represent the coordinates of the measured point. Through the Monte Carlo simulation of 3000 samples, the point cloud distribution is obtained, which is similar to the ellipsoidal distribution. The use of measuring robots for rapid monitoring of the inclination of buildings and the analysis of the accuracy of monitoring results using three-dimensional measurement methods have led to conclusions of universal significance.

4.2. Experimental process. The distance between the measurement target and the dual emission station will have a certain impact on the measurement results. Several monitoring points are selected within the range of 1100mm to 6000mm, with each selected point spaced at 110mm intervals, and experiments are conducted in sequence. Draw a 3D cloud spatial point map of the actual building inclination, a 3D display cloud spatial point map of building inclination under monitoring by measuring robots, and a 3D display cloud spatial electrical state of building inclination under traditional mechanical probe measurement. Compare the inclination and draw an experimental conclusion [16].

4.3. Experimental results. If the error distribution is introduced into the sensor, the direction of laser ranging can be changed. Therefore, the point cloud obtained by Monte Carlo method takes the laser ranging error into consideration, and the distribution accuracy of the point cloud obtained by laser ranging method is very high. The three-dimensional display of the actual building inclination is shown in Figure 4.1.

Using the point cloud data in Figure 4.1 as supporting data, traditional mechanical probe measurement

Table 4.1: Comparison of Monitoring Accuracy between Two Methods

	Number of monitored dimensional groups (piece)	Actual total number of groups	Detection accuracy%
Traditional mechanical probe measurement	2765	2800	92
Measurement robot mechanical probe measurement	2985	2800	99

and measurement robot monitoring methods were used to compare and analyze the monitoring of building inclination. The three-dimensional point cloud spatial distribution formed by the actual inclination of the building is used as a control group. Under traditional mechanical probe measurement, three-dimensional modeling is used to demonstrate the monitoring accuracy of building inclination, which is lower compared to measurement robots. The results of building tilt monitoring based on measuring robots are better matched with the actual situation. In order to further verify the higher monitoring accuracy of the author's research method, the monitoring accuracy of the two methods was compared and analyzed, and the comparison results are shown in Table 4.1 [17].

From Table 4.1, it can be seen that the monitoring accuracy of building inclination based on traditional mechanical probe measurement is 92%, and the monitoring accuracy based on measurement robots is 99%, which is 7% higher than the former. It can be seen that the monitoring accuracy based on measurement robots is higher.

5. Discussion. Improving the accuracy of measurement data by measuring robots

According to the actual situation of buildings, the following measures can be taken to improve the accuracy of measurement data:

1. When using high-precision measurement robots for monitoring, it is necessary to isolate the boundaries, walls, and other areas of the building from the surrounding environment to ensure the accuracy of measurement data.
2. When monitoring buildings, it is necessary to strictly follow the construction plan for monitoring, and the internal structure of the building must be the main measurement object during measurement.
3. When conducting measurements, it is necessary to ensure that the monitoring environment is dry, clean, and tidy, and to repair cracks, holes, and other defects on the building's exterior walls and walls.
4. When conducting measurements, it is necessary to ensure that the measuring robot has good stability and reliability. If there are problems in their work, they should be promptly contacted by the staff to replace them [18].
5. When monitoring buildings, it is necessary to ensure that the instruments and equipment used can meet the relevant detection requirements.

When using measurement robots for building tilt stability monitoring, it is necessary to choose appropriate measurement methods based on the actual situation to improve measurement efficiency. For example, Total station and Dumpy level can be used for slope stability monitoring. In order to ensure the safety of buildings, various parts of the building should be monitored. Due to the fact that using measurement robots can reduce the number of measurements, it can improve work efficiency.

In addition, the use of measurement robots can improve the accuracy and accuracy of monitoring, thereby reducing monitoring costs. When using high-precision measurement robots for building tilt stability monitoring, it can not only improve work efficiency but also reduce monitoring costs. In traditional building tilt stability monitoring, it is usually necessary to measure and calculate the results multiple times. Therefore, it is necessary to measure multiple monitoring points simultaneously and calculate the results [19,20].

6. Conclusion. The use of measuring robots to monitor the inclination of buildings has broken the research process of traditional building inclination observation methods, greatly improving monitoring efficiency. The use of monitoring data can quickly calculate the inclination of building materials, and the monitoring results have high accuracy and reliability. The robot measurement method can deeply measure the inclination of buildings,

and use 3D models to display the monitoring accuracy of building inclination. Compared with traditional methods, it has higher accuracy and is in line with the actual situation, resulting in better measurement results. The use of measurement robots to monitor the inclination of buildings has a significant impact on the measurement results. In the following research, heuristic search algorithms such as genetic algorithms will be used to further optimize the design and layout of monitoring points, thereby improving monitoring accuracy.

REFERENCES

- [1] Tian, J., Jia, L., & Mu, S. (2021). Research of intelligent temperature measuring robot system. *Journal of Physics: Conference Series*, 1757(1), 012152 (5pp).
- [2] Tanaka, K., Nakaya, D., Kondo, Y., & Yoshida, I. (2021). A study on optimal voltage of electromagnet for precision measuring robot during surface roughness measurement by vibration analysis. *International journal of automation technology*36(4), 15.
- [3] Giacoppo, G. A., Mammel, R., & Pott, P. P. (2021). Finding the curved pathway of the large intestine for robot-aided colonoscopy. *Current Directions in Biomedical Engineering*, 7(2), 215-218.
- [4] Tian, Y. M., Xiong, J., Wang, G., Jiang, C., Zhang, S., & Li, W. L. (2021). A novel position and orientation correction method for specific robot poses by measuring an array of standard balls. *Measurement Science and Technology*, 32(12), 125014 (12pp).
- [5] Zhou, T., & Tao, L. (2021). Measuring the micro deformation and crack of rock surface and make robot intelligent walking path from pulse lidar profiler. *Journal of Physics: Conference Series*, 1813(1), 012062 (4pp).
- [6] Velentza, A. M., Fachantidis, N., & Lefkos, I. (2021). Learn with surprise from a robot professor. *Computers & Education*, 173(3), 104272.
- [7] Mu, S., Shibata, S., & Yamamoto, T. (2022). Development of a user-following mobile robot with a stand-up assistance function. *Cognitive Robotics*, 2, (8)3-95.
- [8] Zhang, H. W. Q. (2021). A novel method to identify dh parameters of the rigid serial-link robot based on a geometry model. *Industrial Robot*, 48(1)42.
- [9] Hu, C., Xing, R., Hu, S., & Li, J. (2021). Application of portable measuring equipment in the field of industrial measurement. *Journal of Physics: Conference Series*, 1965(1), 012153-.
- [10] Chiwande, S. N., & Ohol, S. S. (2021). Comparative need analysis of industrial robot calibration methodologies. *IOP Conference Series Materials Science and Engineering*, 1(0)12, 012009.
- [11] Schneckenburger, M., Almeida, R., Hfler, S., & Brret, R. (2022). Material removal by slurry erosion in the robot polishing of optics by polishing slurry nozzles. *Wear*, 494-495, 2(0)4257-.
- [12] Manteuffel, C., Dirksen, N., & Hartwig, T. (2021). From extra to actor: facilitating automated conditioning in animal-robot interaction. *Computers and Electronics in Agriculture*, 1(9)1, 106496-.
- [13] Cheng, Z. S. T. R. (2021). A novel robot-assisted electrical impedance scanning system for subsurface object detection. *Measurement Science & Technology*, 32(8)12.
- [14] Bui, K. D., Wamsley, C. A., Shofer, F. S., Kolson, D. L., & Johnson, M. J. (2021). Robot-based assessment of hiv-related motor and cognitive impairment for neurorehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, PP(99), 1-1.
- [15] Hiller, J., Landstorfer, P., Marx, P., & Herbst, M. (2021). Evaluation of the impact of faulty scanning trajectories in robot-based x-ray computed tomography. *Measurement Science and Technology*, 32(1), 015401 (13pp).
- [16] Wang, C., Wang, W., Wei, Z., Yang, H., Wang, L., & Chen, Z., et al. (2021). Multi-point calibration method for articulated arm coordinate measuring machine based on an observability index. *Measurement Science & Technology*36(12), 32.
- [17] Ba, K. X., Song, Y. H., Shi, Y. P., Wang, C. Y., Ma, G. L., & Wang, Y., et al. (2022). A novel one-dimensional force sensor calibration method to improve the contact force solution accuracy for legged robot. *Mechanism and Machine Theory*, 16(9), 104685-.
- [18] Lanevski, D., Manocheri, F., & Ikonen, E. (2022). Gonioreflectometer for measuring 3d spectral brdf of horizontally aligned samples with traceability to si. *Metrologia*, 59(2), 025006-.
- [19] Wang, Z., Garrett, C. R., Kaelbling, L. P., & Tomás Lozano-Pérez. (2021). Learning compositional models of robot skills for task and motion planning:. *The International Journal of Robotics Research*, 40(6-7), 866-894.
- [20] Mondal, S., & Dutta, A. K. (2021). Technical study of the effect of laser engraving using uarm swift pro robot. *IAES International Journal of Robotics and Automation (IJRA)*, 10(3), 182-191.

Edited by: B. Nagaraj M.E

Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing

Received: Dec 26, 2023

Accepted: Mar 18, 2024



APPLICATION OF SOFTWARE ROBOTS AND DEEP LEARNING IN REAL TIME PROCESSING OF E-COMMERCE ORDERS

WENBO NIU*, YIBO HU† AND WEI ZHANG‡

Abstract. The development of e-commerce faces many problems, among which the real-time information exchange between customers and websites is the most urgent. To study the application of software robots in the real-time processing of e-commerce orders. According to the working mechanism of ALICE, JAVAPROGRAMD, AIML technology, database technology, knowledge base principle, Rule of inference, and reasoning strategies are used to design the structure, workflow, knowledge processing flow, and Inference engine supporting human-computer automatic negotiation of the software robot. Establishing a rule base is the acquisition, induction, and organization of knowledge related to order processing. It requires learning and analyzing the problems and related materials to be solved, simulating the actual process of order processing, and extracting useful processing rules and processes from it. The rules in the system are not organized in a simple list form, but have a certain hierarchical structure. This way, when using these rules for reasoning, the hierarchy is clear and it is easy to modify and add rules. The hierarchical structure of the rules referred to here is to simulate the thinking activities of the human brain in the process of order processing, and organize them in a certain way. Based on the basic rules of order processing, the process of order processing, the rational performance of the order and the possibility of necessary events, so as to realize the structure of the reasoning machine. Expands the application of Chatbot ALICE, and provides a new tool for real-time online negotiation and negotiation of e-commerce order processing. The core part of the e-commerce order processing software robot is the human-machine automatic interaction module. Database module, knowledge base module and rule base module provide the necessary data, knowledge and rule basis for its realization, while the implementation of reasoning machine provides the possibility for the intelligence and dynamic interaction of human-machine automatic interaction module.

Key words: Software robots, E-commerce, Real time processing of orders, Online negotiation, Negotiate reasoning

1. Introduction. With the development of computer technology and network technology, the development of e-commerce advances rapidly. E-commerce can reduce costs, reduce inventory, save time, and so on, but it also faces many problems, among them, the order processing transaction in the high cost, low efficiency is especially obvious. The widespread market of e-commerce has led to increasing demands and expectations from people. At present, the problems faced by e-commerce websites in order processing transactions and the resulting defects of high cost and low efficiency are becoming increasingly apparent [1]. Currently, after consumers place orders through business websites, merchants will confirm the authenticity of the order and negotiate specific matters on the order through phone or email. In the face of fierce global market competition, every merchant should respond promptly and quickly. However, in the above-mentioned order processing process, several aspects are far from meeting the needs of rapidly developing information technology and increasingly competitive online commerce [2]. Therefore, it is extremely important to design man-machine dialogue software robot based on rules and reasoning mechanism, which can free people from order processing; increase the real-time, accuracy and dynamic interaction of order processing; reduce the operating cost of merchants; realize the characteristic service of merchants and the personalized needs of buyers. The research and implementation of this real-time and online man-machine negotiation platform is one of the challenging research topics in e-commerce order processing. Firstly, when a consumer places an order, they are faced with a pre designed business website by the merchant, which includes static information such as description information of various products, prices of products, and fixed delivery times in fixed areas. Consumers have only two choices: accept or not, and cannot make personalized special requirements based on their own situation. Moreover, consumers are not aware at the time whether the order can be fulfilled in a timely manner, that is, the order is fulfilled. Secondly,

*The School of Business, Xi'an International University, Xi'an, 710077, China (Corresponding author, WenboNiu7@163.com)

†The School of Business, Xi'an International University, Xi'an, 710077, China (YiboHu9@126.com)

‡The School of Business, Xi'an International University, Xi'an, 710077, China (WeiZhang93@126.com)

although online commerce based on phone and email can solve negotiation issues after placing an order, not all consumers are willing or have time to interact with merchants after placing an order. In addition, in order to improve online sales and market competitiveness, businesses need to provide online and offline services 24 hours a day, 7 days a week. This undoubtedly increases the operating costs of businesses, and due to the negligence of staff operations, timely and accurate services may not be possible. E-commerce is a further extension of human commodity buying and selling activities, which is changing the business operation mode of enterprises and people's economic lifestyle [3]. However, e-commerce, which is currently in its early stages of development, only provides functions such as information dissemination, the use of electronic currency, the sale and purchase of fixed price goods, and the fixed delivery capacity in fixed areas, lacking the intelligent negotiation part in traditional business activities. Secondly, the marketing philosophy has shifted from traditional 4P (Product, Price, Place, Promotion) to 4C (Customer, Convention, Cost, Communication), and this shift in marketing philosophy has put forward new demands for the operation of business websites [4]. The 4C theory needs to consider both consumer needs and corporate profits, so business websites should be customer-centric and constantly consider consumer needs while ensuring their own profits. Therefore, in order to solve the problems in the order processing process of e-commerce websites, providing efficient e-commerce transaction platforms with online negotiation mechanisms for merchants and customers can promote the further development of e-commerce, at the same time, it can realize personified real-time processing of orders. Therefore, it is necessary to design a human-computer dialogue software robot for e-commerce order processing based on rules and reasoning mechanism. Software robots are not fatigued and are based on a powerful real-time and dynamically updated database and rule base in the backend, enabling e-commerce websites to provide real-time, accurate, and error free online order processing services, which can promote the automation and intelligence of product buying and selling in e-commerce. Therefore, the implementation of human-machine dialogue software robots can free people from the process of order processing; Increase the real-time, accuracy, and dynamic interactivity of order processing; Reduce the operating costs of businesses: Achieve unique services for businesses and personalized customer needs, thereby increasing consumer satisfaction. The research and implementation of this real-time and online human-machine negotiation platform is one of the challenging research topics in e-commerce order processing [5,6]. At present, the research on negotiation reasoning mainly focuses on automatic negotiation model, interactive multi-objective negotiation model, negotiation support system, virtual reality technology, anthropomorphic human-computer interaction Agent, business auction system based on Agent technology, online bidding system based on XML technology, etc. The auction on Internet was so great success that some researchers believe that the auction is the only effective negotiation mechanism in e-commerce. The information integration in the real-time order processing system is also deeply concerned. The application of ALICE research is also very extensive.

This paper adopts JAVAPROGRAMD and AIML technology, using database technology, knowledge base principle, reasoning rules and reasoning strategy, and designs the structure, workflow, knowledge processing process of the software robot and the reasoning machine to support man-machine automatic negotiation. It has expanded the application of chatbot ALICE and provides a new tool for real-time online negotiation and negotiation of e-commerce order processing.

2. Literature Review. In order to solve the problems in the process of order processing of e-commerce websites, it is extremely important to provide efficient man-machine dialogue software robot with online negotiation mechanism to design man-machine dialogue software robot based on rules and reasoning mechanism.

At present, due to the research and implementation of internal and external e-commerce order processing, there is a research on the order processing process of the supply chain based on Finite-state machine. "Most of the order processing is mainly online, that is, not real-time online order processing. For example, the world's largest Chinese online book and video store, Dangdang.com, eGuo.com Mall, Gome Appliances, TOM Mall, Joyo Network, and so on, the shopping process of these websites generally includes: Logging in, selecting products, registering, logging in, filling out orders, and confirming payment methods, all these websites provide only the function of placing orders online through the network environment, without providing timely online negotiation and order confirmation when customers place orders, in this case, the so-called order processing is carried out manually offline, and the order processing personnel check the content of the order, including whether the goods ordered by the customer have sufficient supply and whether they can be delivered to the customer's

designated receiving location on time, which is the effective fulfillment of the order [7]. In this case, whether the order placed by the customer can be effectively fulfilled depends on whether the customer can provide the processing result of an order through email or phone on the order after the above processing is completed offline. At present, the real-time and online part of e-commerce online shopping includes the rationality verification of order content and the provision of online electronic payment function when placing orders. Among them, the rationality verification of order content does not refer to the verification of information that affects order fulfillment, but rather to the verification of the validity of the data by the business website according to certain rules when filling out the order, including the type of data, the number of digits of the ID number, whether the contact phone number and email address are correct, and so on. Online payment is based on financial electronic networks, using commercial electronic tools and various transaction cards as media, and using computer and communication technology as means to store electronic data in the computer system of banks, and it is a means of circulation and payment through electronic information transmission through computer network systems. Online payments circulate through electronic currency, which is a cash currency that exists in electronic digital form [8]. At present, research on online electronic payments mainly focuses on payment gateways, electronic banking, electronic wallets, electronic checks, and so on. The various online payment methods launched by the financial system are becoming increasingly perfect, which has invisibly played a role in promoting the rapid growth of online shopping. The C2it online payment service launched by Citibank in the United States can help customers establish an online account, and customers can easily complete payment procedures without entering their credit card number or bank account information each time they make a payment. Nevertheless, online electronic payment is also carried out when the order has not been confirmed to be fulfilled, adopting a mechanism of payment before delivery. The condition for its smooth execution is that both parties assume that the order will be fulfilled, but this process undoubtedly increases the uncertainty of order processing [9].

Electronic commerce order processing software robot is the core part of the man-machine automatic interaction module, database module, knowledge base module and the corresponding rule base module, provides the necessary data, knowledge and rules basis, and the implementation of the reasoning machine automatic interaction module provides the intelligence, dynamic interaction. The first step of establishing the rule base is to acquire, summarize and organize the knowledge of order processing. It is necessary to study and analyze the problems and relevant materials to be solved, simulate the actual processing process of the order, and extract useful processing rules and processes from it. The rules in the system are not organized in the form of a simple list, but in a certain way, based on the basic rules of order processing, with the process of order processing, the rational performance of orders and the possibility of necessary events as the reasoning mechanism, so as to realize the structure of the reasoning machine.

Software robots actually refer to a program that can simulate an operator's operation, and can automatically complete a certain operation without human participation, thus replacing some of the human work. The current research on software robots mainly focuses on search engine software robots (Uluka, Spider, CyBot, MetaCrawler), shopping software robots (AcseBookfinder), and chat software robots (Eliza, ALICE). The shopping software robot applied in e-commerce is particularly eye-catching, as it not only helps buyers find the most cost-effective price, but also automatically orders this product when the best performance price ratio is found. This software robot has certain advantages when ordering goods, but when customers cannot determine whether their orders can be fulfilled in a timely manner in order to ensure that they can purchase the goods, they will place the same order on multiple websites. Therefore, orders that are not processed in a timely manner can cause false demand for businesses [10,11]. The real-time order processing software robot is mainly responsible for obtaining customer ordering information, logistics and distribution arrangement results of e-commerce websites, and providing customers with negotiation information about the positioning results of supply points and the selection results of third-party logistics and distribution centers.

With the development of computer technology and network technology, the development of e-commerce has made rapid progress. E-commerce can reduce costs, reduce inventory, and save time, but it also faces many problems, among them, the high cost and low efficiency in order processing transactions are particularly evident. Currently, after consumers place orders on the website, merchants will confirm the authenticity of the order and negotiate specific matters on the order through phone calls, text messages, and emails. In the face of fierce global market competition, every merchant should make timely and rapid responses, and in the above

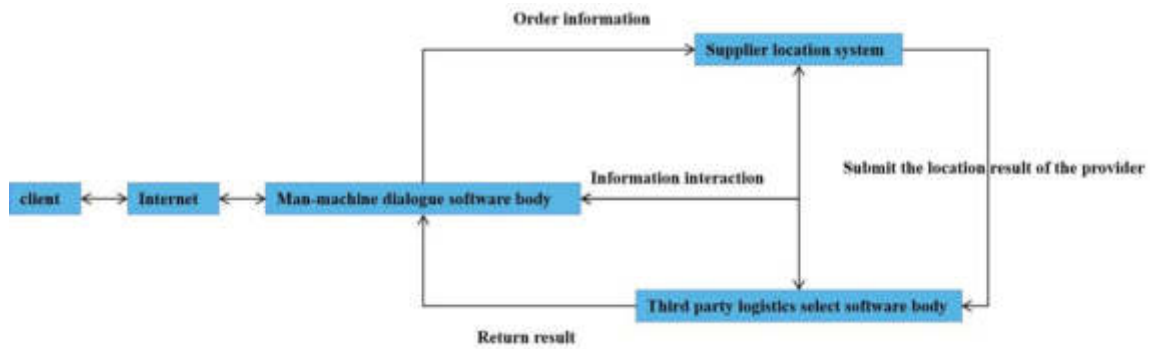


Fig. 3.1: Real time processing process of e-commerce orders

processing process, it is far from meeting the needs of online commerce. With the development of computer technology and network technology, the development of e-commerce has made rapid progress. E-commerce can reduce costs, reduce inventory, and save time, but there are also many problems, among which the high cost and low efficiency in order processing transactions are particularly evident. Currently, after consumers place orders on the website, merchants will confirm the authenticity of the order and negotiate specific matters on the order through phone calls, text messages, and emails. In the face of fierce global market competition, every merchant should make timely and rapid responses, and in the above processing process, it is far from meeting the needs of online commerce. Therefore, the structure of the human-machine dialogue software robot, inference rules and inference machines supporting human-machine automatic negotiation, have been designed, expanding the application scope of the chat robot ALICE, and providing new tools for real-time online negotiation and negotiation of e-commerce order processing [12].

3. Methods.

3.1. Determine system boundaries . Determining system boundaries means identifying what is inside and outside the system, and determining their relationships, but it is necessary to consider interface or information transfer issues within and outside the system. The location and information interaction relationship of this system in real-time processing of e-commerce orders are shown in Figure 3.1 [13].

This system is a man-machine dialogue system, in the running process of the Web application, from the customer login business website, to select goods with the server software robot after many negotiations, the system needs to remember to interact with it customers which one, what interaction behavior, namely the customer behavior, this requires business website web application ability to record and track the session. Even though the customer visits hundreds of web pages on the website and orders dozens of goods, after many negotiations with the software robot on the server side, the system can remember all the goods the customer wants to buy and the specific customers and relevant information they interact with. However, the construction and development of the website is based on HTTP protocol, and HTTP itself is a stateless protocol, namely there is no memory function, which means that it cannot each customer a request with another request, they use different ports at different times, so the server has no association about the request. In order to enable the system to remember that the customer who chose the product and the customer who negotiated with it are the same customer, and to record the relevant information in the negotiation process, the conversation tracking mechanism should be adopted to solve this problem.

Session tracking is the process of recording a customer logically associated with different access requests over a period of time. Sessions can be tracked through the customer's only-ID at each service request. Each session is identified by a unique session ID, used to track multiple requests sent from the same customer to the server, and to associate the customer with his session data.

3.2. System design of human-machine dialogue software robot. The system is divided into two parts for design and development, including web applications and EJBs. The system adopts Uml for system

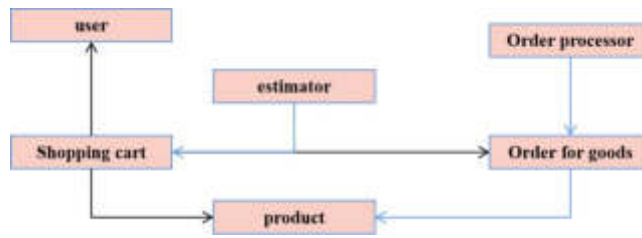


Fig. 3.2: Component Relationship Diagram

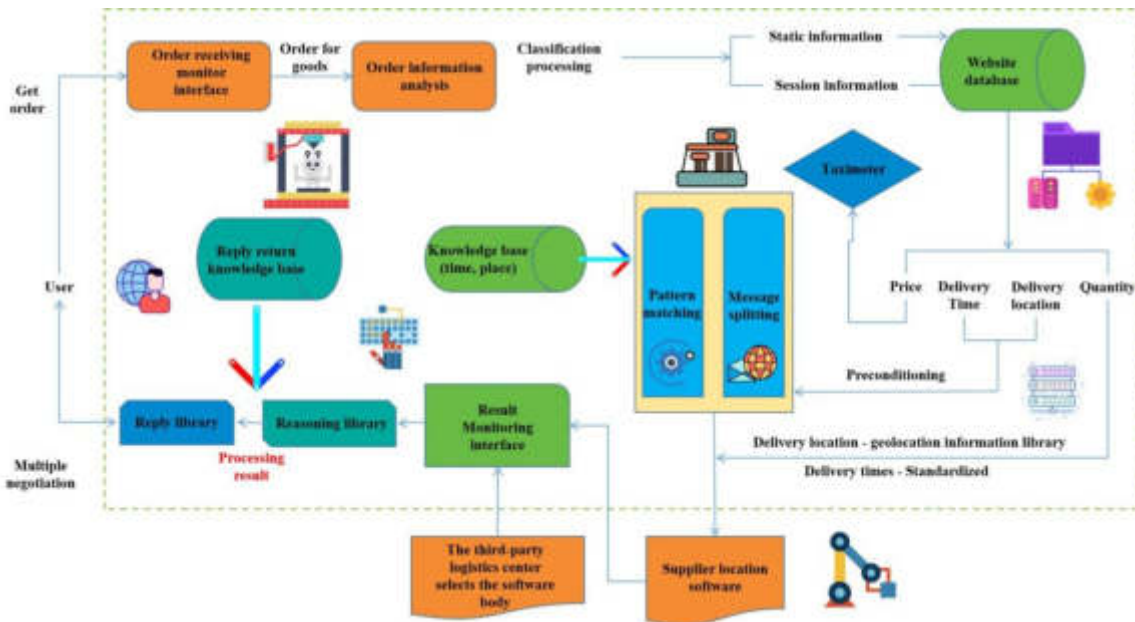


Fig. 3.3: Structure of human-machine dialogue software robot

analysis, while applying a modular design concept, and adopts Java. Jsp for system implementation [14].

(1) *Design of System EJBs Model.* The data objects involved in this system include User, Good, and Order, and three EntityBeans (UserEJBs, BookEJBs, and OrderEJBs, all of which are CMP types) were designed, each EntityBean has a table corresponding to it in the database, and each instance of EntityBean corresponds to a row of data in the Table.

The system accesses entity beans in the form of session beans. In this system, two types of session beans are used, stateful session beans representing shopping carts and stateless session beans representing evaluators. The component relationship is shown in Figure 3.2 [15].

(2) *The structure of human-machine dialogue software robot.* The main task of implementing human-machine dialogue software robots is to create a, when interested in purchasing goods, but unable to determine whether the merchant has sufficient supply of goods and can fulfill the order at the designated time and location, real-time online negotiation is conducted with the merchant to sign the order under mutually acceptable conditions. The structure of the human-machine dialogue software robot is shown in Figure 3.3 [16].

(3) *Design of Robot Analysis Mechanism for Human Machine Dialogue Software.* Information analysis is one of the most important parts of the personification real-time processing of e-commerce orders, mainly providing four functions: 1) Analyze the original information of orders received by the website, including the analysis of delivery time and location; 2) Analyze the processing results of the supply point positioning software body;

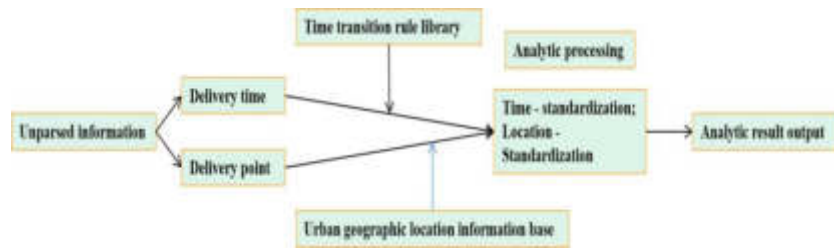


Fig. 3.4: Information Analysis Mechanism Structure of Original Orders

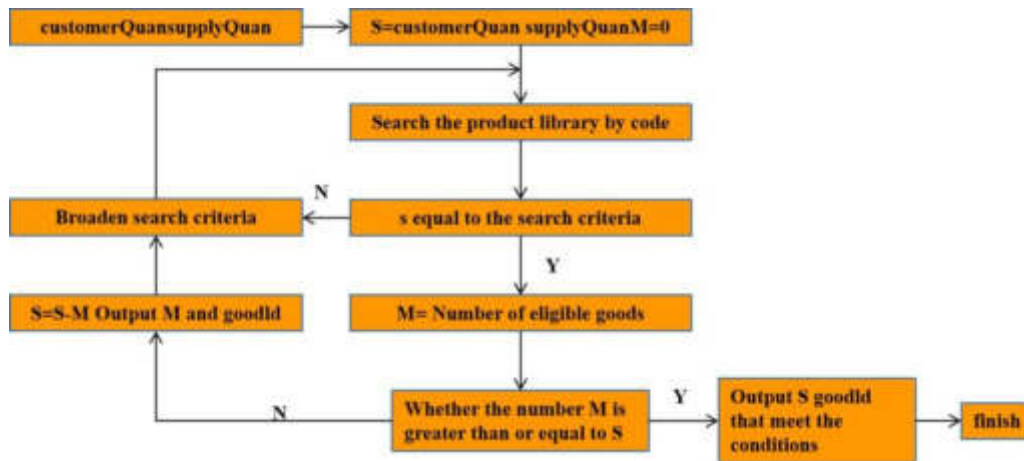


Fig. 3.5: Product Replacement Search Process

3) Analyze the processing results of selecting software entities for third-party logistics centers; 4) Combine the response and return to the knowledge base to perform secondary parsing on the processing results of 2) and 3), in order to obtain a humanized human-machine negotiation result. Therefore, in the information input interface section, it is necessary to implement the classification of received information, and input corresponding parsing rules based on the classified results for parsing, while providing the parsing results. When parsing the original order, the parsing mechanism structure is shown in Figure 3.4 [17].

(4) Reasoning Rule Design. If the user wants to buy 100<JSP Introduction and Improvement), after the first human-machine negotiation, the feedback information shows that the customerQuan (quantity of goods ordered by the user) is greater than the supplyQuan (quantity of goods available in the supply point positioning software), and the ten digit code of "JSP Introduction and Improvement" is N002C1P1W3l0, then search for books that match or approximate N002C1P1W3 in the product library, as shown in Figure 3.5 [18].

In the process of product substitution, the substitution rules used are from precise matching to approximate matching. The rule execution process is as follows: When substitution behavior occurs, first search according to the taxi system code of the book that the user wants to purchase. For example, N002C1P1W3, if the quantity requirement is met, the search will be stopped. Otherwise, the matching conditions will be relaxed and similar matching rules will be implemented. Regarding the different importance of the four types of coding rules in the ten digit system, while ensuring that the category of the ordered book remains unchanged, perform similar matching in the following order, namely N002C1P1... - N002C1... - N002... During the search process, as long as the required quantity is met, exit the search and output the product ID and quantity that meet the conditions [19].

Table 4.1: Function Table of Java Class in Inference Machine

Java class	function
resSupply. java	Call the supply point positioning software body to select the supply point, and return the quantity and time of goods provided by the supply point
judgeSupply. java	Judge the processing results of the supply point positioning software, that is, judge the quantity and time. If the quantity is not enough, replace the product. If the time cannot be met, negotiate the time and call the corresponding user negotiation interface
goodReplace. java	The product substitution module mainly performs substitution queries based on the coding rules of the product, and the query process is dynamic, from precise matching to approximate matching as needed
timeReplace. java	Time change negotiation may occur after the supplier locates the software body, or after the third-party logistics center selects the software body
resTpl. java	Call the third-party logistics center software to perform logistics delivery and return the available logistics delivery time
judgeTpl java	Judge the processing results of the third-party logistics center software body to determine whether there has been a time change negotiation. If necessary, call the corresponding negotiation interface
orderOk. java	Accept order, update order6 data table, delete temporary data table
orderEnd. java	Cancel the order, update the order6 data table, and delete the temporary data table

4. Implementation of inference engine in human-machine negotiation. The core part of the e-commerce order processing software robot is the human-machine automatic interaction module. The database module, knowledge base module, and rule base module provide necessary data, knowledge, and rule basis for its implementation, while the implementation of the inference machine provides the possibility for the intelligence and dynamic interaction of the human-machine automatic interaction module.

Establishing a rule base is the acquisition, induction, and organization of knowledge related to order processing, it requires learning and analyzing the problems and related materials to be solved, simulating the actual process of order processing, and extracting useful processing rules and processes from it. The rules in the system are not organized in a simple list form, but have a certain hierarchical structure. This way, when using these rules for reasoning, the hierarchy is clear and it is easy to modify and add rules. When modifying, you only need to modify the rules at a specific level without modifying or changing the rules at other levels. The hierarchical structure of the rules referred to here simulates the thinking activities of the human brain during the order processing process, and organized in a certain way, based on the basic rules of order processing, the inference mechanism is based on the process of order processing, the reasonable fulfillment of orders, and the possibility of necessary events, to achieve the structure of the inference machine. When designing the inference engine structure, the inference process that handles the same problem is implemented as a class, and the processing situations with the same probability of occurrence are implemented as classes at the same level. The general form of rules in this system is: (IF<state judgment>) (THEN<operation 1>ELSE<operation 2>). The inference engine of this system is written in Java language. The description of the Java classes inside the inference machine is shown in Table 4.1 [20].

5. Conclusion. With the development of computer technology and network technology, the development of e-commerce has made rapid progress. E-commerce can reduce costs, reduce inventory, and save time, but there are also many problems, among which the high cost and low efficiency in order processing transactions are particularly evident. Currently, after consumers place orders on the website, merchants will confirm the authenticity of the order and negotiate specific matters on the order through phone calls, text messages, and emails. In the face of fierce global market competition, every merchant should make timely and rapid responses, and in the above processing process, it is far from meeting the needs of online commerce. Therefore, the structure of the human-machine dialogue software robot, inference rules supporting human-machine automatic negotiation, and inference machine have been designed, expanding the application scope of the chat robot ALICE, provides new tools for real-time online negotiation and negotiation of e-commerce order processing. The core part of the e-commerce order processing software robot is the human-machine automatic interaction module. Database module, knowledge base module and rule base module provide the necessary data, knowledge and rule basis for its realization, while the implementation of reasoning machine provides the possibility for the intelligence and dynamic interaction of human-machine automatic interaction module. How to use natural

language to negotiate with human-machine dialogue software robots for order processing in e-commerce is a challenging research direction with broad application prospects, and further in-depth research is needed.

6. Acknowledgement. Shaanxi Provincial Social Science Foundation of China: Research on the Circulation Mechanism of "Agriculture Consumer Connection" of Shaanxi Characteristic Agricultural Products (No. 2022D052).

REFERENCES

- [1] Zhang, C., & Ren, M. (2021). Customer service robot model based on e-commerce dual-channel channel supply coordination and compensation strategy in the perspective of big data. *International Journal of System Assurance Engineering and Management*, 14(2), 591-601.
- [2] Zolkin, A. L., Munister, V. D., Lavrov, E. A., Aygumov, K. G., & Saradzheva, V. (2021). Creation of a software and hardware product of a real-time system for collecting, accounting and managing data transmission of an intelligent transport system in context of the iot. *Journal of Physics: Conference Series*, 2094(5), 052059-.
- [3] Glukhov, M. A., Glukhova, E. D., Marunkov, P. A., & Barulin, A. S. (2021). Application of in-house software to improve the design process of multifunctional aircraft indicators and control panels. *Journal of Physics: Conference Series*, 1864(1), 012118-.
- [4] Liu, C., & Liu, R. (2021). Application of bp neural network in cross-border e-commerce web pages quality evaluation. *Journal of Physics Conference Series*, 1774(1), 012015.
- [5] Ma, Y., Fan, X., Cai, J., Tao, J., & Yang, Q. (2021). Application of sensor data information cognitive computing algorithm in adaptive control of wheeled robot. *IEEE Sensors Journal*, PP(99), 1-1.
- [6] Ni, B., Cao, J., Mao, Z., Cheng, L., Shen, H., & Chen, J. (2021). Research on real-time sensing technology of aerospace interactive behavior based on infrared array force tactile. *Journal of Physics: Conference Series*, 2078(1), 012065-.
- [7] Buachoom, A., Wuttisela, K., & Wuttirom, S. (2021). Development of a simple line-follower robot with constant acceleration motion. *Journal of Physics Conference Series*, 1719(1), 012091.
- [8] Balázs Gyenge, Zoltán Máté, Vida, I., Bilan, Y., & László Vasa. (2021). A new strategic marketing management model for the specificities of e-commerce in the supply chain. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(4), 1136-1149.
- [9] Wang, S. (2021). Study on the application of computer security technology in e-commerce. *Journal of Physics Conference Series*, 1915(4), 042044.
- [10] Gururaj, P. (2021). Artificial intelligence-application in the field of e-commerce. *International Journal of Research - GRANTHAALAYAH*, 9(4), 170-177.
- [11] Heriyandi, A., Reza, R. F., & Albar, C. N. (2021). Designing user interface of web-based e-commerce application. *Journal of Physics: Conference Series*, 1764(1), 012187-.
- [12] Qu, L., & Dai, Y. (2021). Research on the path of improving the service quality of rural e-commerce logistics in jilin province based on computer. *Journal of Physics Conference Series*, 1744(3), 032144.
- [13] Sardjono, W., Selviyanti, E., Mukhlis, M., & Tohir, M. (2021). Global issues: utilization of e-commerce and increased use of mobile commerce application as a result of the covid-19 pandemic. *Journal of Physics: Conference Series*, 1832(1), 012024 (6pp).
- [14] Yin, L., & Lu, Z. (2021). The integrated application of computer software in architectural space design. *Journal of Physics: Conference Series*, 1812(1), 012033 (4pp).
- [15] Lorenc, A., & Burinskiene, A. (2021). Improve the orders picking in e-commerce by using wms data and bigdata analysis. *FME Transactions*, 49(1), 233-243.
- [16] Kimr, N., Bodunkov, N., & Sinyavskaya, J. (2021). Hardware and software structure for a social robot capable of situation analysis. *Journal of Physics: Conference Series*, 2096(1), 012070-.
- [17] Clément.Cormi, Parpex, G., Julio, C., Ecartot, F., Laplanche, D., & Vannieuwenhuyse, G., et al. (2022). Understanding the surgeon's behaviour during robot-assisted surgery: protocol for the qualitative behav'robot study. *BMJ open*, 12(4), e056002.
- [18] Bulej, V., Barto, M., Tlach, V., M Bohuřk, & Wiecek, D. (2021). Simulation of manipulation task using irvision aided robot control in fanuc roboguide software. *IOP Conference Series: Materials Science and Engineering*, 1199(1), 012091-.
- [19] Ji, Z., Yuan, X., Lin, M., & Huang, X. (2021). Dynamic analysis of a 6-dof wheeled mobile robot. *Journal of Physics: Conference Series*, 1948(1), 012082-.
- [20] He, D., Lu, J., Zhou, Q., & Yin, H. (2021). Vision-based robot identification and tracking. *Journal of Physics: Conference Series*, 1971(1), 012033 (6pp).

Edited by: B. Nagaraj M.E

Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing

Received: Dec 26, 2023

Accepted: Mar 18, 2024



INTELLIGENT ALGORITHM OPERATION AND DATA MANAGEMENT OF ELECTROMECHANICAL ENGINEERING POWER COMMUNICATION NETWORK BASED ON THE INTERNET OF THINGS

YING LI* WENJING QU† AND ZHENQIANG ZHANG‡

Abstract. With the rapid development and gradual popularization of Internet of Things technology, it is necessary to provide necessary Operation and Maintenance (O&M) services and data control for electromechanical engineering to maintain the normal function of the power system. This paper does the following research to carry out orderly and standardized management of related communication resources in power communication networks, conduct closed-loop and process-oriented management of communication O&M work, and ensure the safe, stable, and economical operation of power grids in electromechanical engineering. Firstly, the technology selection, algorithm implementation, solutions, and other related technologies that may be used in the platform design and implementation process are introduced and selected. Secondly, the research and call logic of the design and implementation of related algorithms for intelligent O&M are introduced. It includes the design and implementation of anomaly detection algorithms to monitor equipment health status and the design and construction of fault diagnosis algorithms for abnormal analysis. Finally, the simulation experiment of the proposed processing scheme is carried out on the Mininet simulation network to prove that the proposed scheme can provide a better anonymization effect when introducing low latency. The results show that the design of the gateway system realizes the module applications of the system, such as user, data storage, O&M, and fault management. Based on the technical selection, the algorithm is implemented and optimized, and the call logic of the algorithm is implemented in the O&M module. Simulation verifies that the anonymization algorithm can complete the mapping without introducing an additional delay of more than 3%.

Key words: Internet of things, electromechanical engineering, power communication networks, intelligent operation and maintenance, data control

1. Introduction. The safety and stability control system and dispatching automation system of the power communication network and power system are important supports for maintaining the safe production of the power system. Also, it is the basis for grid dispatching automation, marketization of power grid operation, and modernization of management and provides important maintenance means to ensure the safe, stable, and economic operation of the power grid. Orderly and standardized management of relevant communication resources in power communication networks, centralized and intelligent monitoring of communication equipment and communication services, closed-loop and process-oriented management of communication Operation and Maintenance (O&M) work, and the formation of intelligent, integrated, and automated intelligent full-process management and control network of power communication network are effective means to improve work efficiency and ensure the safe operation of communication equipment [1].

This paper confirms and analyzes the O&M requirements and O&M status of electromechanical equipment and studies the implementation direction and system architecture innovation of intelligent O&M and Internet of Things (IoT) systems. Then, the overall system is designed and implemented with the idea of microservices, and the architecture pattern of microservices is formed. Platform services are provided through the architecture pattern, forming a low-coupling structure to meet the needs of convenience and scalability. Next, the box-plot, time series anomaly detection algorithm, and isolated forest algorithm are studied and presented separately. After testing and evaluating the characteristics of the three algorithms, a hierarchical intelligent detection

*Department of Intelligent Engineering, Shijiazhuang Posts and Telecommunications Technology Vocational College, Shijiazhuang, Hebei, 050031, China. (Corresponding author's e-mail:YingLi36@126.com)

†Department of Intelligent Engineering, Shijiazhuang Posts and Telecommunications Technology Vocational College, Shijiazhuang, Hebei, 050031, China.(WenjingQu7@163.com)

‡Department of Intelligent Engineering, Shijiazhuang Posts and Telecommunications Technology Vocational College, Shijiazhuang, Hebei, 050031, China.(ZhenqiangZhang7@126.com)

mechanism is designed to cope with the needs of anomaly detection in different situations. After anomalies are detected, a Decision Tree (DT) algorithm is used to troubleshoot based on anomalous data. It realizes the intelligent O&M of the system.

Data control opens up some data center interfaces. In this case, the data center of the power communication network will be significantly more likely to be attacked, and the attack methods will be very different. Data center traffic needs to be dynamically processed to prevent grid data from being improperly stolen or leaked during transmission. Through the centralized dynamic management and control capability provided by Software Defined Network (SDN), the efficient and fine-grained anonymization of data transmission is completed to achieve the purpose of security protection of power communication networks.

2. Literature Review. The IoT is a new industrial concept based on the concept of the Internet. It consists of the Internet as the core of interaction and the basis of communication. IoT involves industrial, home, medical, and logistics aspects. It is a network that interconnects the entire society. According to Muteba's research, the number of IoT devices worldwide was expected to reach 22 billion by 2025 [2]. Jalali said that the domestic output of the IoT industry was growing at a rapid rate of 20% per year, and the types and number of IoT devices were growing exponentially [3].

Algorithmic Intellectual Property Operations is an O&M with algorithms as the core. It is the latest O&M mode formed by O&M services with the continuous development of computing power, equipment, and concepts. At present, domestic and international intelligent O&M services are often integrated on large-scale platforms. For example, in the Amazon Web Services provided by Wei, the intelligent O&M function is only the overall edge function of the platform. The O&M function is relatively simple. The platform includes many functions not required by intelligent O&M services, which are prone to additional overhead [4]. You proposed a new architectural pattern that is different from traditional monolithic architecture. Its main feature is the division and treatment of system functions. The system's overall function is split, and the single function of the split is realized in a monolithic architecture pattern [5].

Network traffic control includes private information such as IP addresses and network ports of network users. If maliciously detected and analyzed by the outside world, it will cause the leakage of this information, violating user privacy and even trade secrets. Hammoudeh proposed the idea of using the programmability of SDN for anonymous services. It realized the hiding of network resources through address mapping [6]. Later, Iordache proposed some other anonymization algorithms. These algorithms achieve the purpose of shielding private information from the outside world by hashing information such as IP addresses to varying degrees. Meanwhile, some algorithms reduce the impact of anonymization operations on performance by caching [7].

Compared with foreign network management research and network management system development, China started late. After more than ten years of construction, China's network management standardization research has also achieved great results. However, the current intelligent O&M platform pays little attention to the emerging industry of the IoT. Besides, the current domestic and international intelligent O&M services are relatively simple O&M functions. The platform as a whole includes many functions that are not required by intelligent O&M services, which is prone to additional overhead. It can be seen that the intelligent O&M platforms on the market cannot meet the requirements. So, the innovation points here are as follows. First, the intelligent O&M mode is combined with the IoT industry to realize an intelligent O&M platform focusing on IoT devices' monitoring and O&M services. Second, it is very necessary and meaningful research to ensure the lightweight and scalability of the management and control platform. At present, the lack of this part of the content on the market should be filled to meet the overall development trend of the IoT. Third, anomaly detection depends on the implementation and optimization of algorithms. However, no one algorithm is suitable for all anomaly detection scenarios. Therefore, this paper selects and combines algorithms to achieve the optimal detection function.

3. Methods and materials.

3.1. Technologies related to the intelligent O&M platform of the power grid.

3.1.1. Microservices architecture. The microservice architecture realizes convenient development and flexible deployment and improves the system's scalability by redefining the separation and communication of services [8]. It is proposed that the power grid intelligent O&M platform has requirements for lightweight and

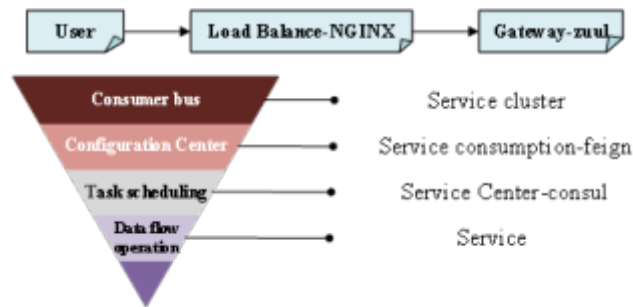


Fig. 3.1: Spring Cloud framework architecture diagram (NGINX (engine x): High-performance HTTP and reverse proxy web servers)

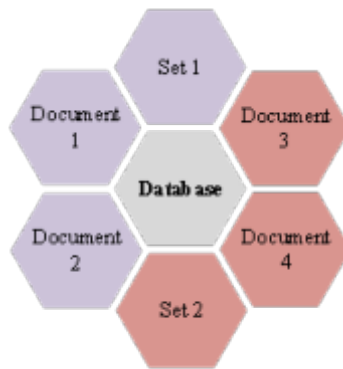


Fig. 3.2: Database storage structure

flexibility. The microservice architecture can realize the design of the system platform and further extend the system for this. As a result, the microservice system architecture is used to build the server-side system.

Spring Cloud is a mature framework that combines various mainstream frameworks. The implementation of individual services relies on the Spring Boot framework of the same company. Figure 1 shows the Spring Cloud framework architecture.

The Spring Cloud framework integrates with current mainstream third-party components or frameworks to provide microservices infrastructure capabilities. It encapsulates third-party components or frameworks during the integration process, shields the internal complex configuration and implementation principles, and forms a simple, reliable, and mainstream microservice system framework. There are specific requirements for lightweight deployment in the scenario of intelligent O&M for IoT devices. The non-intrusive Istio framework is not suitable for additional deployment tasks. The system studied here is designed and implemented based on the Java language, so there is no need for the Thrift system for microservice communication between multiple languages. Finally, Spring Cloud integrates many of the current mainstream components and frameworks. Additionally, it is widely used. As for stability and community considerations, Dubbo, as a recently restarted project, is far inferior to Spring Cloud.

3.1.2. Relational databases. Relational databases store data in a relational model. The data location is located by rows and columns inside the database. Several rows and columns form a table, and multiple tables form a database [9]. Figure 3.2 displays the database storage structure.

The basic data storage unit for Mongo (Humongous) is a document. Its structure uses a Binary Serialized Document Format similar to JavaScript Object Notation. As a mainstream document-based database, Mongo

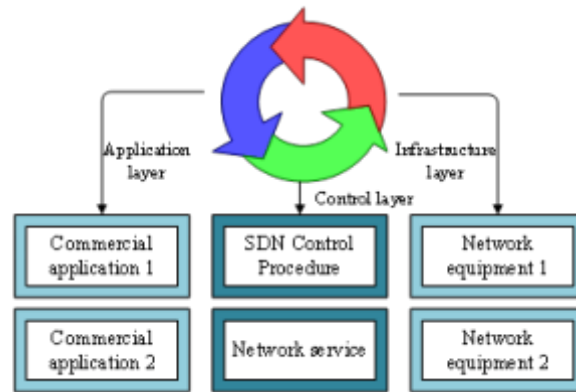


Fig. 3.3: SDN three-layer network architecture

is convenient to store and read quickly. It is suitable for large-scale data storage.

3.1.3. Intelligent O&M algorithm.

1. Isolation Forest (iForest) is a fast anomaly detection algorithm. The basic idea of the algorithm is ensemble learning, so it has certain advantages in linear time complexity and accuracy [10]. The rationale for iForest is to define anomalies as “outliers that are easily isolated.” Each data point is divided into a separate space by constantly cutting the data space from different characteristics. Multiple cuts are required, and sparsely distributed points can be divided into space by a few cuts [11].
2. DT is a commonly used classification method. Through the classification results of known data and the corresponding occurrence probability, the logical judgment of if-then-else is constructed by the effect of factoring. The classification method of sample data is obtained. DT is an algorithm based on probabilistic analysis [12]. In constructing if-then-else logic, binary trees are usually used as logical structures, so the resulting algorithm is called a DT.
3. Box-plot is a statistical map based on statistical principles to show data distribution. The box-plot can display the distribution characteristics of the data to judge the two data with a small probability of abnormalities to achieve anomaly detection [13]. The plot of the box-plot is achieved by quartiles, which reduce the influence of extreme values in the sample data with quartiles. The distance judges anomalies from the upper and lower quartiles.

3.2. Technologies related to power grid data management and control.

3.2.1. Data center software definition. The existing power communication network architecture is a three-layer architecture built on the extended tree protocol. The transmission of data packets is completed through various transmission protocols [14]. However, with the increasing scale of the IoT and massive data, the routing tables in routers are becoming increasingly complex, which brings many problems to the current network framework. When tuning network devices, network administrators must configure each switch or router one by one using the command line [15]. The SDN concept is designed to solve this type of problem. Figure 3 presents the SDN three-layer network architecture.

SDN is a network virtualization technology. It strips the control functions of traditional switches and separates the data plane and control plane. SDN entrusts the function of decision control to the control layer and uses the open interface provided by the control layer to complete the delivery of decisions [16]. The network layer data of data center traffic is processed through the SDN network, combined with pooled IP. The data center’s security is ensured by decoupling the internal organization of the data center from the outside world. Figure 4 shows the specific structure diagram of the communication data center.

The network layer mainly has two parts, and one is SDN-based data center control network feeding. The central node discovery and the message transmission are part of the Peer-to-Peer network composed of blockchain

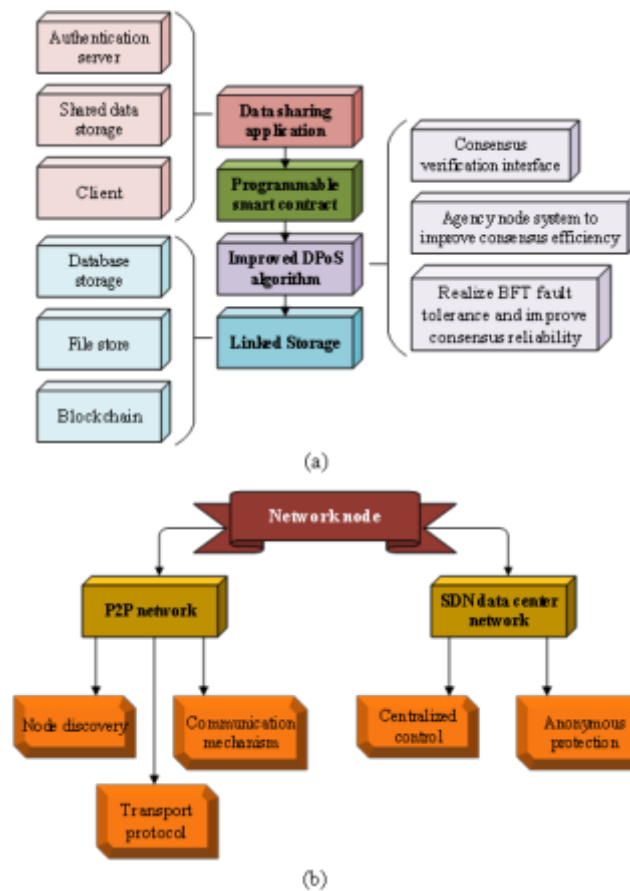


Fig. 3.4: Specific structure diagram of the communication data center

nodes responsible for node discovery, message propagation, and transmission protocol parsing in blockchain nodes [17]. The consensus layer and the data layer form the core of the blockchain data protection mechanism. The consensus layer improves the Delegated Proof of Stake (DPoS) algorithm. Based on the DPoS algorithm proxy system, the consensus layer combines the Practical Byzantine Fault Tolerance algorithm to improve the consensus efficiency and provide reliable node consensus.

3.2.2. Data copyright supervision mechanism. The center interacts with institutions in real time to fetch the returned data. These two methods are shown in Figure 3.5 for the data flow process.

The data itself has the characteristics of reproducibility and easy processing. In addition, the “see and own” nature of data makes it easier to own data than traditional goods. Agencies can move data at a lower cost [18].

4. Model design.

4.1. Simulation design of intelligent O&M service of power communication network. In the process of system implementation, for the consideration of applicability, the minimum available basic functions of anomaly detection should be provided for various types of IoT devices. As a classical algorithm with simple implementation, stable performance, and low requirements for data characteristics, box-plot has specific usability and wide applicability, which is suitable for the needs of this scenario. Therefore, the box-plot is selected as the basic algorithm for system anomaly detection. Table 4.1 records the relationship between the number of box-plot scenarios and the fluctuation of the detection accuracy of a single indicator.

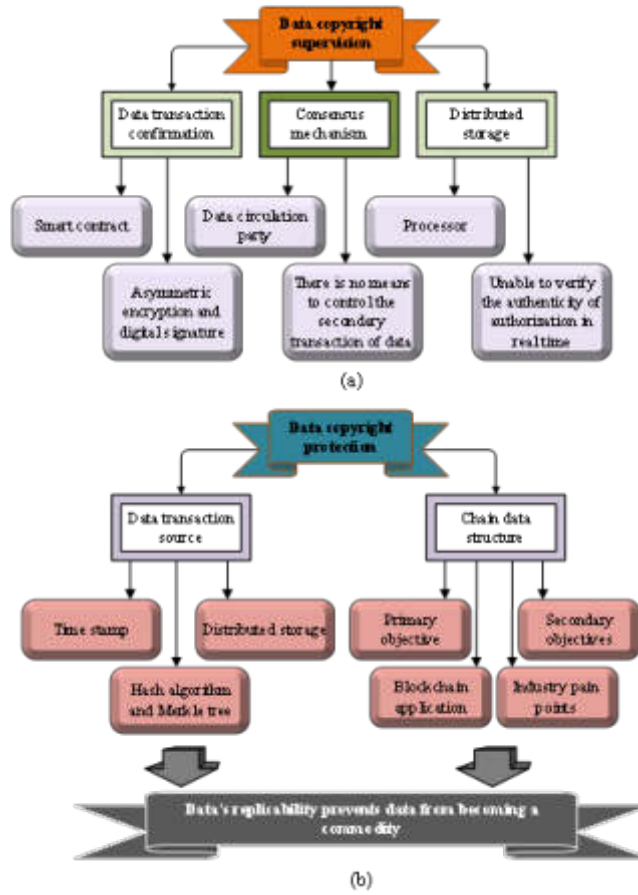


Fig. 3.5: Central interaction and real-time interactive data flow architecture diagram

Table 4.1: Box-plot sample number and detection fluctuation table

Number of scenes	50	250	750	1,500	2,500	3,500	4,500	5,500
Correct rate variance	0.045	0.017	0.014	0.006	0.002	0.006	0.012	0.008

Box-plot calculation formulas have inherent flaws. The data features cannot be effectively distinguished when the sample size is small. When the sample size is large, overfitting problems are prone to occur.

The data synchronization scheme in the data distribution scenario is designed, and the combination of synchronous calls and asynchronous callbacks is decided. Based on the message queuing transaction solution, the synchronous/asynchronous call method is improved to ensure that data can be successfully distributed to each node. Data is successfully synchronized to ensure data consistency. The specific implementation logic is demonstrated in Figure 4.1.

1. The first is the transmission of the O&M module. As the main functional module of the system, the O&M module has a higher priority than the data module. As data is the most critical dependency element of the O&M module, the real-time and accuracy of data transmission must be guaranteed.
2. The second is the transmission of data modules. The data module has a lower priority than the O&M module in system function priority. There are no requirements for data real-time. Therefore, the method of combining asynchronous call and callback confirmation is used to ensure the successful

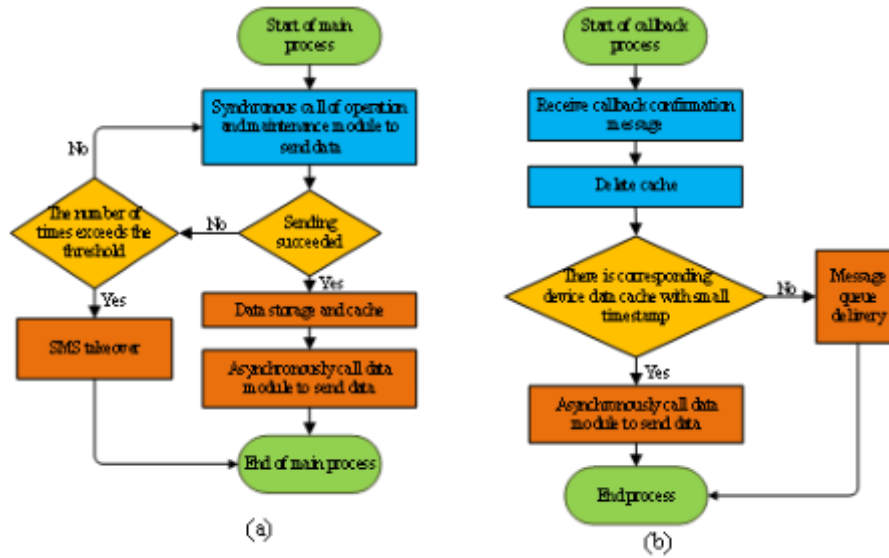


Fig. 4.1: Synchronous and asynchronous call logic flowchart (subscriber management system)

transmission of data and data consistency.

- The callback part receives an acknowledgment request after the data module receives the data. The acknowledgment request contains the device number and data timestamp. After the callback function receives the acknowledgment request, the corresponding data is deleted from the Redis cache. The data with the timestamp before the current callback data time cut-off is retrieved from the collection. It is the data that occurred before but still did not receive the callback acknowledgment.

The combination of synchronous transmission and asynchronous transmission ensures data consistency during data distribution between O&M modules and data modules, saving system overhead and shortening the time of data synchronization.

4.2. Data control center anonymization processing experimental simulation. This experiment studies the data anonymization guarantee scheme in the electromechanical engineering power system environment. The coupling relationship between network traffic is reduced by hashing and encrypting information such as account addresses and data payloads of data packets. The risk of improper in-depth analysis of data packets by the outside world is reduced to achieve the purpose of not leaking the internal information of the data control center.

The data anonymization solution is based on the SDN network. The pooled IP address is mapped in the egress switch of the data center to achieve the purpose of anonymization. Figure 4.2 shows the architecture of the anonymous service system.

SDN networks have the characteristics of centralized management and programmability, which can be used to deliver data at the data center egress gateway without the introduction of additional hardware devices. The topology and network information inside the data center can be completely shielded from the outside. The anonymization service component AnonyService is mounted on the SDN controller of the data center network. The main function of this component is to complete the mapping between Routable IP Address (RIPA) and Machine IP Address (MIPA) and the corresponding flow table delivery. MIPA is the actual address of the data center server, which is only perceived inside the data center and is shielded from the outside world. RIPA is an IP address that the outside world can access. After the controller calculates the mapping relationship between the two, it generates a flow table and delivers it.

When new traffic arrives, it determines whether the current anonymized flow table exceeds the given threshold. If it is exceeded, the timeout table is dropped. After that, determine whether the threshold has

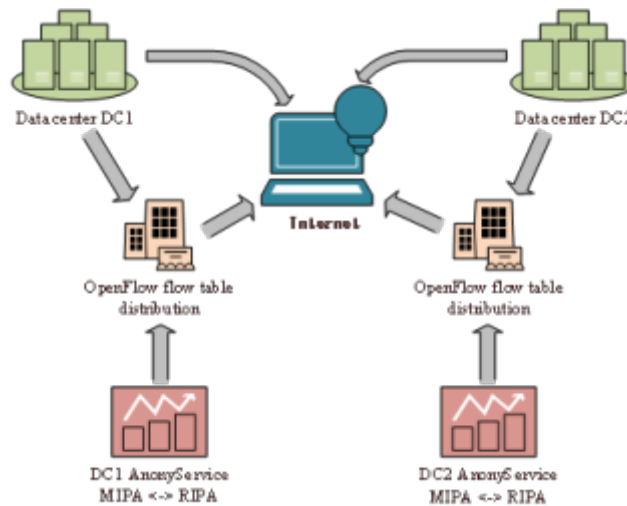


Fig. 4.2: Data center anonymous service system architecture

Table 4.2: Mapping algorithm pseudocode

```

Start
Input: Network traffic sequence S
Output: Anonymized flow table
Begin
If AnonyService.use >= max_capacity then
    AnonyService.delete_by_time);
    If AnonyService.use >= max_capacity then
        AnonyService.destory);
        If flow_entry != null;
            AnonyService.destory();
            AnonyService.init(max_capacity);
            index=(src_ip.hashcode+timestamp)%anonyIPs.current_size;
            new_src_ip=anonyIPs.get(index);
            flow_entry=AnonyService.generate_flow_entry(src_ip,
            new_src_ip, timestamp);insert_entry(flow_entry);
        End if
    End if
End if
End
    
```

been exceeded. If it is still exceeded, a new discarded current mapping is created. The flow table is emptied, and the process is repeated. Table 2 illustrates the mapping algorithm pseudocode.

The main operations of this mapping algorithm are as follows. The first is service initialization, which is responsible for creating the initial flow table and configuring the flow table threshold. The second is invalid stream table cleanup, which is responsible for clearing stream table entries that have not been matched within a period. Unlike the expiration time specified during the flow table configuration, this cleanup is an overall active cleanup to reduce the space footprint within the switch. The third is flow table emptying. If the switch space is still occupied after the invalid flow table cleanup, the flow table emptying operation will begin. The memory space occupied by the flow table is reclaimed, and a service initialization is triggered after recycling. The fourth is flow table lookup, which is mainly to find whether there is already a corresponding flow table

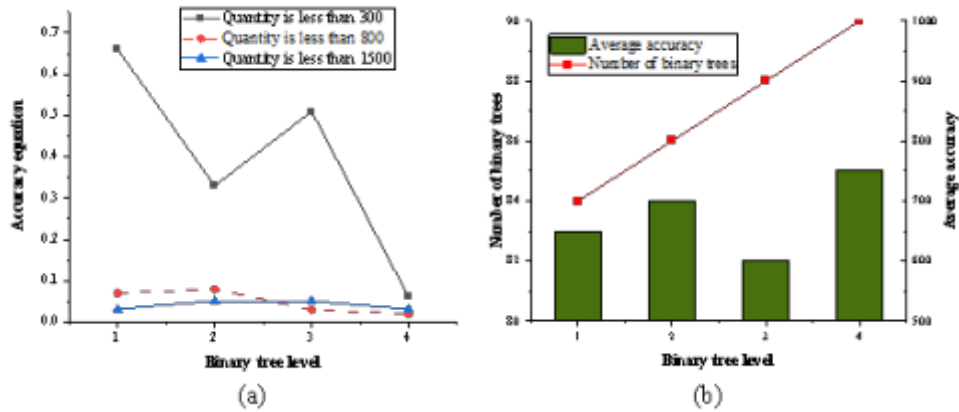


Fig. 5.1: Detection accuracy and quantitative bounds of samples under different quantities

entry based on the anonymous source IP address according to the need. If there is one, return it directly. The fifth is flow table insertion. This part is mainly for new traffic. It inserts the anonymized IP address into the switch flow table.

5. Results.

5.1. Intelligent operation platform algorithm implementation and optimization. The total number of trees in the forest affects the accuracy of the system's detection. The detection results obtained by ensemble learning with enough isolated trees are more accurate. In addition, the total number of trees also affects the system's performance. Too many trees cause additional overhead to the system and affect the system's overall performance. Therefore, the number of trees in isolated forests needs to be studied to achieve optimal detection results and minimize the waste of system resources. The most stable quantitative boundary is obtained by studying the fluctuation of the detection accuracy of samples under different quantities. The results of the study after several tests are plotted in Figure 5.1.

From Figure 5.1, when the number of binary trees is less than 700, the variance of the accuracy rate is at a large value, indicating that the anomaly detection is relatively unstable. When the number of binary trees reaches 700, the accuracy variance is relatively small. The variance corresponding to the subsequent quantity also fluctuates within a small range, and the variance as a whole shows the nature of convergence and stability. When the number reaches 700, the detection effect of the system is stable. Adding more trees in the future does not significantly improve the detection effect, which will cause additional unnecessary overhead to the system. Therefore, 700 is selected as the number of binary trees in implementing the iForest algorithm.

Apache Jmeter is used to perform network stress tests on the system. The concurrent threads are created. Users are impersonated to access server ports. A stress test is conducted. The test environment is an i5 processor and 8G memory in a Windows environment. The final stress test results are revealed in Figure 5.2.

From Figure 5.2, when the number of concurrencies exceeds 2,000, the proportion of error requests and system response time increase significantly. According to the test results, it is concluded that when the concurrent requests are less than 2,000, the system's high availability can be guaranteed. Creating a cluster can improve system availability when requests are more than 2,000.

The test data set is used to detect anomaly detection performance and test the accuracy of different levels of anomaly detection function. The test results are shown in Figure 5.3.

5.2. Power communication network data analysis. Iperf simulated network traffic is tested to study the impact of anonymization schemes on response times for external requests. During the simulation, a delay of 1 ms between the client and the data center is set. Different traffic rates are sent through control flow

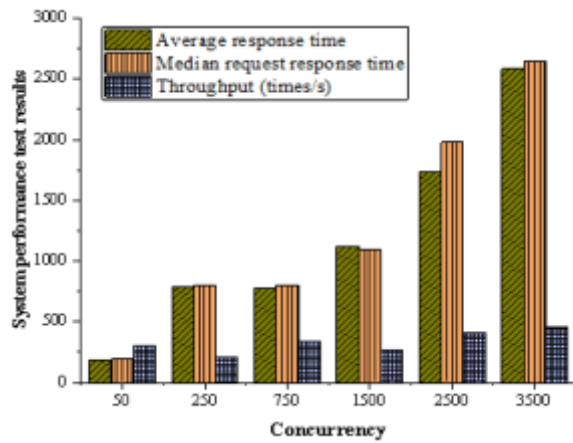


Fig. 5.2: Final stress test results

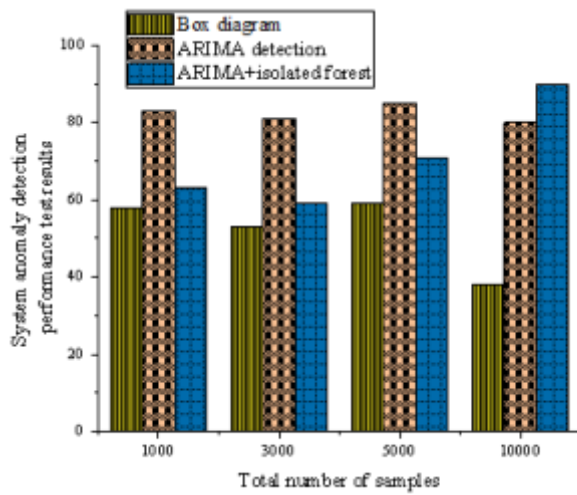


Fig. 5.3: According to the test results, the box-plot has a specific usability when the sample data is small, which is suitable for early anomaly detection. Autoregressive Integrated Moving Average Model (ARIMA) detection accuracy is relatively stable and remains high. Combining ARIMA with iForest will reduce the accuracy due to false positives of the iForest algorithm when the sample size is small.

clients to test responses under various network stresses. For each flow rate, experiment one min and repeat the experiment ten times. The delay of the flow rate in each experiment is counted. Figure 5.4 shows the delay statistics.

From Figure 5.4, with or without anonymization, the response time is slightly over 1 ms. As data traffic grows, response times tend to rise. This shows that the network delay (1 ms) of the outgoing client access data center, and the internal response time of the data center is very low, about the order of 10us. The experiment is relatively small, so the delay and jitter variation of data transmission are very small. However, it can be seen overall that the loss of response time caused by anonymization processing is very small. When the data flow rate is less than 50 Mbps, the internal latency of the data center without anonymization is about 0.039 ms, and the internal delay of the anonymized data center is about 0.041 ms. The response time increases by about 5%. This performance loss is very low and within the acceptable range. The test scale is relatively small,

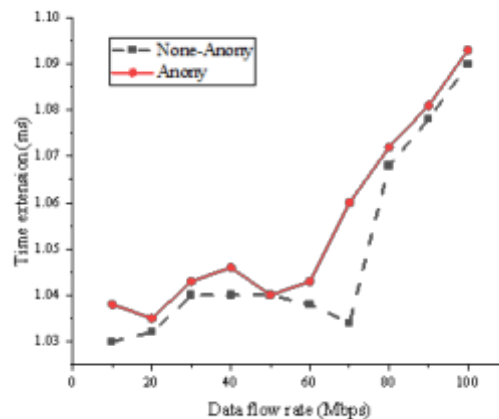


Fig. 5.4: Traffic delay statistics for each experiment

so the internal latency of the data center is relatively low. Iperf statistics are not accurate enough, and the performance impact should be lower in practice.

6. Conclusion. This paper studies the power communication network supporting the IoT. Based on this research purpose, the research and design are carried out by combining the current relatively complete intelligent O&M and data management concepts. The research results are as follows: (1) When the binary trees reach 700, the detection effect of the system is stable. The subsequent addition of more trees does not significantly improve the detection effect, which will cause additional unnecessary overhead to the system. So, 700 is selected as the number of binary trees in implementing the iForest algorithm. (2) When the concurrent requests are less than 2,000, the system's high availability can be guaranteed. When the number of requests exceeds 2,000, the system availability can be improved by adding servers and creating clusters. (3) When the algorithm data is increased to a specific value, the iForest algorithm improves the overall recognition rate of the system anomaly detection function. Therefore, it can be considered that the graded detection method adopted by the system is feasible and effective. (4) The internal latency of the anonymized data center is about 0.041 ms, and the response time increases by about 5%. This performance loss is low and within acceptable limits. However, due to the current status quo of software and hardware, there is still room for further optimization and improvement of the platform. First, data timing has a high dependence on the network state. Once network fluctuations occur, there will be a disorder in the data transmission process. Second, when providing O&M functions for different devices in the current system, it is necessary to redesign the algorithm and modify the source code. In the subsequent work, it is essential to improve the system and carry out additional design and implementation of functions such as data timing verification and retransmission start. Moreover, it is necessary to design and implement the algorithm training configuration module to realize the algorithm's automatic training and configuration call algorithm functions to improve the intelligence and automation of the system.

REFERENCES

- [1] Yang D., Wei H., Zhu Y., et al. (2018) Virtual private cloud based power-dispatching automation system—Architecture and application [J]. *IEEE Transactions on Industrial Informatics*, 15(3): 1756-1766.
- [2] Muteba F., Djouani K., Olwal T. (2019) A comparative Survey Study on LPWA IoT Technologies: Design, considerations, challenges and solutions [J]. *Procedia Computer Science*, 155: 636-641.
- [3] Jalali M. S., Kaiser J. P., Siegel M., et al. (2019). The internet of things promises new benefits and risks: a systematic analysis of adoption dynamics of IoT products. *IEEE Security & Privacy*, 17(2), 39-48.
- [4] Wei Y., Peng M., Liu Y. (2020) Intent-based networks for 6G: Insights and challenges [J]. *Digital Communications and Networks*, 6(3): 270-280.

- [5] You X., Zhang C., Tan X., et al. (2019) AI for 5G: research directions and paradigms [J]. *Science China Information Sciences*, 62(2): 1-13.
- [6] Hammoudeh M., Epiphaniou G., Belguith S., et al. (2020) A service-oriented approach for sensing in the Internet of Things: intelligent transportation systems and privacy use cases [J]. *IEEE Sensors Journal*, 21(14): 15753-15761.
- [7] Iordache D. (2021) Database-Web Interface Vulnerabilities [J]. *STRATEGIES XXI-Security and Defense Faculty*, 17(1): 279-287.
- [8] Li N., Liu G., Zhang H., et al. Micro-service-based radio access network [J]. *China Communications*, 2022, 19(3): 1-15.
- [9] Lu Y., Liu C., Kevin I., et al. (2020) Digital Twin-driven smart manufacturing: Connotation, reference model, applications and research issues [J]. *Robotics and Computer-Integrated Manufacturing*, 61: 101837.
- [10] Barbariol T., Chiara F D., Marcato D., et al. (2022) A review of tree-based approaches for anomaly detection [J]. *Control Charts and Machine Learning for Anomaly Detection in Manufacturing*, 2022: 149-185.
- [11] Ning X., Li F., Tian G., et al. (2018) An efficient outlier removal method for scattered point cloud data [J]. *PloS one*, 13(8): e0201280.
- [12] Shah D., Patel S., Bharti S K. (2020) Heart disease prediction using machine learning techniques [J]. *SN Computer Science*, 1(6): 1-6.
- [13] Mishra P., Pandey C M., Singh U., et al. (2019) Descriptive statistics and normality tests for statistical data [J]. *Annals of cardiac anaesthesia*, 22(1): 67.
- [14] Khan W Z., Rehman M H., Zangoti H M., et al. (2020) Industrial internet of things: Recent advances, enabling technologies and open challenges [J]. *Computers & Electrical Engineering*, 81: 106522.
- [15] Alabady S A., Al-Turjman F., Din S. (2020) A novel security model for cooperative virtual networks in the IoT era [J]. *International Journal of Parallel Programming*, 48(2): 280-295.
- [16] Hyun J., Van Tu N., Yoo J H., et al. Real-time and fine-grained network monitoring using in-band network telemetry [J]. *International Journal of Network Management*, 2019, 29(6): e2080.
- [17] Liu X., Jaekel A. (2019) Congestion control in V2V safety communication: Problem, analysis, approaches [J]. *Electronics*, 8(5): 540.
- [18] Wilson J P., Butler K., Gao S., et al. (2021) A five-star guide for achieving replicability and reproducibility when working with GIS software and algorithms [J]. *Annals of the American Association of Geographers*, 111(5): 1311-1317.

Edited by: B. Nagaraj M.E

Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing

Received: Dec 27, 2023

Accepted: Mar 20, 2024



ANALYSIS OF ABNORMAL FREEZING DATA AND UPDATING ALGORITHM FOR ELECTROMECHANICAL ENERGY METER TERMINALS

SHUZHI ZHAO*, YUE DU†, SHANSHAN HE‡, JIAO BIAN§ AND JIABO SHI¶

Abstract. Due to the rapid development of science and information technology, electricity information acquisition system has been widely used in the electricity data collection of users. With the massive electricity data collection, it is difficult to adopt traditional data processing methods to meet the abnormal data processing. In order to effectively mine abnormal information in electricity consumption data, an anomaly detection model based on the Isolation Forest (iForest) algorithm is proposed. Firstly, the daily load curve with strong regularity is used as the characteristic index of anomaly monitoring, and the users with abnormal electricity consumption data are preliminarily screened. Secondly, on the basis of electrical variables, the suspected abnormal users are further analyzed, and the anomaly identification model of electricity data is established to automatically classify the voltage at the metering point. Moreover, combined with the current data, the abnormality of the electric energy metering device is identified, and then the validity of the model is determined through on-site verification. Finally, according to the participating voltage of the fault phase, the 96-point voltage data frozen during the failure period is analyzed and the correction coefficient is adjusted. The results reveal that the electricity data detection model based on the iForest algorithm has significant advantages in computational efficiency. Through the cumulative recall and Precision-Recall (P-R) curves of the model, it is found that the majority of abnormal users can be detected only by detecting a few users with high abnormal scores, which shows that the model has high efficiency. The decision tree algorithm combined with the current data can effectively identify the anomalies of the energy metering device, which verifies the validity of the anomaly identification model of the electricity consumption data.

Key words: Energy meter, Electricity information acquisition system, Abnormal electricity consumption data, Isolated Forest algorithm, Decision Tree algorithm

1. Introduction. As one of the most important technical bases in the power system, the electric energy metering technology is constantly improving with the improvement of technical and service levels [1,2]. As the basic equipment in the power automation system, electric energy metering plays a vital role in the power system, mainly responsible for data reading, processing, and storage. The application of big data technology promotes the integrity and accuracy of power-metering terminal data. The electric energy metering technology is mainly used in the electricity information acquisition system (hereinafter referred to as the “EIAS”), which is the basic platform for the collection of users’ electric power information and can acquire and real-time monitor the power information of all power consumers. The measurement anomaly detection function is principally to monitor and analyze the collected data in real-time, and report the abnormal power, load, voltage, and current of electricity users to the EIAS [3,4].

At present, there are more and more studies on the abnormality of terminal data of energy meters in the power system. Hasan et al. (2019) proposed a power theft detection system based on the Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) architecture. An LSTM model based on CNN is implemented for data classification in a smart grid. The results denoted that the proposed scheme can better classify most ordinary users and a few users with abnormal electricity consumption data [5]. Ji et al. (2021)

*Tangshan Power Supply Company, State Grid Jibei Electric Power Co., Ltd., Tangshan, Hebei, 063000, China. (Corresponding author’s e-mail:ShuzhiZhao7@163.com)

†Tangshan Power Supply Company, State Grid Jibei Electric Power Co., Ltd., Tangshan, Hebei, 063000, China.(YueDu13@126.com)

‡Tangshan Power Supply Company, State Grid Jibei Electric Power Co., Ltd., Tangshan, Hebei, 063000, China.(ShanshanHe9@163.com)

§Tangshan Power Supply Company, State Grid Jibei Electric Power Co., Ltd., Tangshan, Hebei, 063000, China.(JiaoBian19@126.com)

¶Tangshan Power Supply Company, State Grid Jibei Electric Power Co., Ltd., Tangshan, Hebei, 063000, China.(JiaboShi52@163.com)

Table 2.1: Data field description of EIAS

Name	Implication
ID	Meter number
DATA_DATE	Data acquisition time
DATA_TYPE	Power type
ORG_NO	Number of the power supply company
DATA_WHOLE_FLAG	Data collection success identification
R1-R96	96-point load data
CONS_ID	User ID
CONS_NAME	User name
CONS_SORT_CODE	User type
ELEC_ADDR	User address
METER_ID	Number of measuring points
ASSET_NO	Asset number of electricity meter

put forward an estimation method for real-time robust auxiliary state for power system prediction based on the Bayesian framework, deep learning, and Gaussian mixture model (GMM). By combining anomaly detection technology in machine learning (ML) with GMM, abnormal data in measurement information can be accurately determined and deleted. Numerical simulations on IEEE 118-node and IEEE 300-node test systems show that the proposed method has high accuracy and robustness [6]. Liu et al. (2021) raised a framework based on general data mining, which can extract typical power load patterns and discover the insightful information hidden in the patterns. The proposed framework was applied to analyze the time series electricity data of three practical office buildings in Chongqing, and its validity was verified [7]. Yan and Wen (2021) proposed a power theft detector using metering data based on extreme gradient lift. Preprocessing of measurement data, including recovery of missing and wrong values and normalization. Compared with 8 ML methods such as support vector machine and Decision Tree (DT), the proposed method can detect power theft with higher accuracy or a lower false alarm rate. Experimental results also manifested that the proposed method was robust when the data was unbalanced [8].

To sum up, the current anomaly detection algorithm for the energy meter's power system mainly constructs a normal data model and identifies data inconsistent with the model as abnormal data, which leads to excessive redundancy of information and low computational efficiency. At the same time, it can be found that mining abnormal electricity consumption data is helpful to improve the efficiency of power enterprises. To improve computing efficiency, an anomaly detection model of electricity data is established based on the Isolation Forest (iForest) algorithm, and users suspected to be abnormal are identified through preliminary screening combined with a daily load curve. Then, according to the electrical variable of the abnormal user, the voltage is accurately classified by the DT algorithm and determined by combining the current data, identifying the abnormal electric energy metering device. Ultimately, the calculation method of the correction coefficient of electric quantity is put forward.

2. Abnormal Detection of Electricity Consumption Data Based on the iForest Algorithm.

2.1. The data basis of the model.

2.1.1. Choice of information fields for electricity consumption data. Abnormal electricity consumption data shows high line loss, large loss of electricity sold, etc. The above phenomena are mainly caused by line and equipment quality, management loopholes, energy meter quality, and abnormal consumption behavior. Since the daily load curve is characterized by strong regularity and a more obvious shape, anomalies in electro-data can be found more easily. Therefore, the daily load curve is used to analyze electric power data [9]. The fields of the initial electricity load data set selected from the EIAS are exhibited in Table 2.1:

The data set used above retains the meter number, user type, power type, 96-point load data, and other fields [10]. The user type is three non-resident users, and the power type is positively active. Now, the main

Table 2.2: Types of dirty data

Types	Missing value	Duplicate value	Minimax	Load burr	Impact load
Presentation	The table of data has NA or blank	The data of the user's electricity load appears to be repeated at some point	Too large or too small power load data	Data increases or decreases suddenly between adjacent time periods	The meter reads down over a continuous period of time

special transformer terminals used in the system are the 96 protocol and the 05 protocol. The 96 protocol can only be used to collect the voltage, current, and electricity data of the user's energy meter at zero. The special transformer terminal of the 05 protocol collects the voltage, current, and energy data of the user's energy meter every 15 minutes, with a total of 96 points in 24 hours. To facilitate data description, the 96-point load data are reduced to 24 points.

2.1.2. Data cleaning. In the process of dirty data processing, the types of dirty data should be analyzed and summarized first, and then targeted processing should be carried out according to its manifestation. After obtaining the power load data of industrial users, dirty data is identified by data specification principles. Common types of dirty data are outlined in Table 2.2:

In the processing of dirty data, firstly, the redundant data in the data set should be deleted. In a data set, the customer name and time uniquely determine a data record. If multiple records are the same, the redundant data needs to be deleted. Secondly, it is necessary to maintain the integrity of the data set. The problem of missing data in the data set must be properly handled according to the current situation. Every user must have the electricity reading data for every hour per day. If there is a small amount of missing data, the severity of the missing should be analyzed. The missing severity is as follows:

1. The curve is missing 20% of its reading points;
2. The curve continuously misses more than 2 consecutive readings.

If the data missing reaches the above two conditions, the user will be excluded from the research range, and the remaining load curves containing missing data will be repaired by the multi-stage Lagrange interpolation method. The repair of the missing value of the load curve is written as Equation 2.1:

$$P_t = \frac{\sum_{k=1}^{m_1} P_{t-k} + \sum_{i=1}^{m_2} P_{t+i}}{m_1 + m_2} \quad (2.1)$$

m_1 and m_2 refer to the number of forward periods and backward periods, and t stands for the time when the load data is missing. After data cleaning, X is recorded as $(n - \lambda) \times 24$ th order effective load curve matrix composed of $(n - \lambda)$ effective daily load curves.

2.1.3. Data dimension reduction. Electricity load data is easily affected by many factors such as income, price policy, and temperature. The results caused by these influences cannot be fully reflected through the distance, and the similarity of the shape or contour of the time series cannot be fully guaranteed. In order to fully reflect the similarity between loads while taking into account the operational efficiency, six commonly used daily load characteristic indexes are selected to comprehensively reflect the electricity consumption characteristics [11].

All-day load rate reflects all-day load variation:

$$a_1 = \frac{P_{av}}{P_{max}} \quad (2.2)$$

The maximum hourly utilization rate of the whole day reflects the time utilization efficiency:

$$a_2 = \frac{P_{sum}}{24P_{max}} \quad (2.3)$$

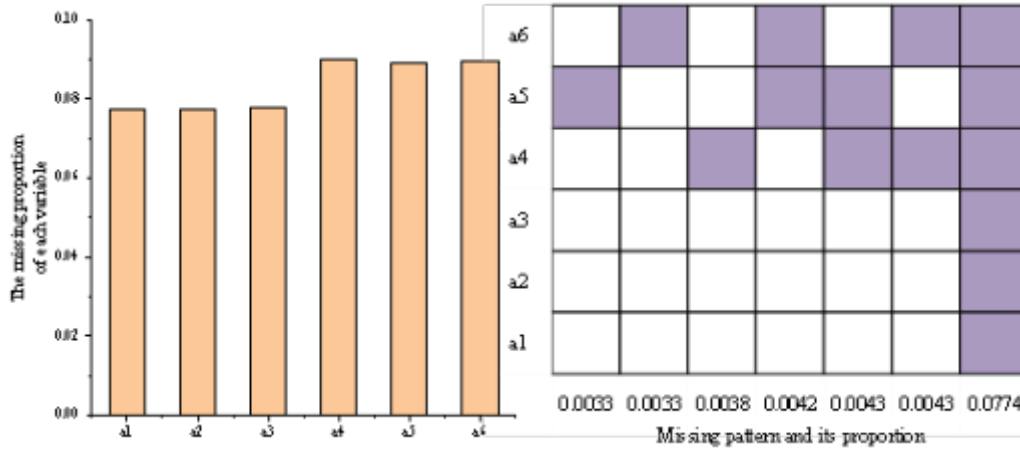


Fig. 2.1: Data missing pattern after dimensionality reduction of the daily load curve

The daily peak-valley difference throughout the day reflects the capacity of the peak regulating of the power grid:

$$a_3 = \frac{P_{max} - P_{min}}{P_{max}} \tag{2.4}$$

Peak load rate reflects peak load variation:

$$a_4 = \frac{P_{av.peak}}{P_{av}} \tag{2.5}$$

Normal load rate reflects the change of normal load:

$$a_5 = \frac{P_{av.sh}}{P_{av}} \tag{2.6}$$

The load rate in the valley period reflects the load change in the valley period:

$$a_6 = \frac{P_{av.val}}{P_{av}} \tag{2.7}$$

By using the load characteristic index to reduce the characteristic dimension of the effective load curve matrix, the $(n\lambda) \times 6$ th order characteristic dimension reduction matrix is obtained, which is recorded as Y. Through visualization processing of the overall situation of data after dimensionality reduction, the result is expressed in Figure 2.1:

Figure 2.1 signifies the missing pattern of the daily load curve. After dimensionality reduction, nearly 90% of the samples do not miss any information, and the 6 features of 7.7% of the samples are Not a Number (NaN), indicating that the value does not exist, which proves that there is no electricity or the account has been canceled. These users will be deleted. The remaining six missing patterns are caused by only a small amount of load during one part of the day and no power during the other. Then divided by the daily average load, it is judged by the computer as 0 / 0 type, the value does not exist, and it is displayed as a missing value, so these six types of users are listed as suspicious users.

2.2. Data anomaly detection model based on the iForest algorithm. The daily load curves of 5972 users in S City on July 18, 2021, were taken as the research object. The daily load curves were mainly selected for small and medium-sized special transformer users and three general industrial and commercial users. The sampling interval of samples was 15 minutes, and there were 96 measurement points in total. After data dimension reduction and cleaning, a total of 4872 effective daily load curves were obtained. There were 63 abnormal users, accounting for 1.29%.

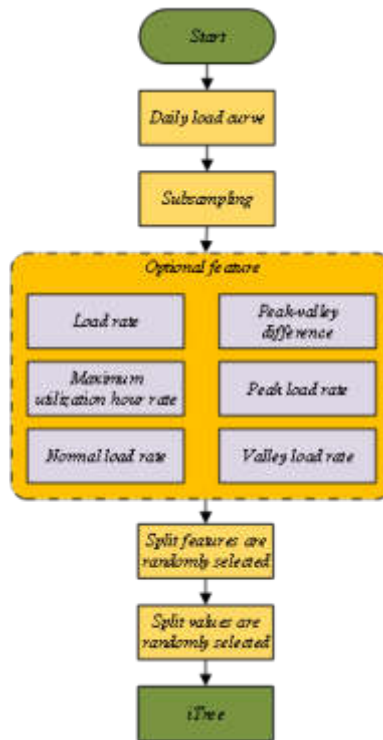


Fig. 2.2: The flow chart of the iTree construction

2.2.1. The construction of the Isolation Tree (iTree). The iForest is mainly composed of the iTree, which refers to a random binary tree in which each node contains two child nodes or leaf nodes [12]. The flow chart of the iTree construction is displayed in Figure 2.2:

Among them, the features in the data set of the daily load curve are all continuous variables. The construction steps of iTree are as follows:

- Step 1: A feature is randomly selected among the 6 daily load characteristic indexes;
- Step 2: A value k of the characteristics selected in step 1 is randomly selected;
- Step 3: According to the characteristics, records are classified every day, and records with characteristics smaller than k are placed in the left branch, and records with characteristics greater than or equal to k are placed in the right branch;
- Step 4: Then the left and right branches are constructed recursively until the following two conditions are met; There are only multiple identical records or one record in the incoming data set; The height of the tree reaches the specified height.

2.2.2. The construction of the iForest. The construction of iForest is similar to the method of random forest, both of which are carried out by random sampling. Each tree needs to be constructed through part of the data set to ensure that each tree has certain differences. The construction process of the iForest is revealed in Figure 2.3:

In the process of constructing iForest, the sampling size should be limited on the one hand, and the maximum depth should be set for each iTree on the other hand. Finally, it is necessary to calculate the power value of the tested user. In the process of evaluating the tested user, iForest can only evaluate a single customer at a time. Meanwhile, during the process of evaluation, each iTree needs to be traversed, the statistical query object is in the position of the leaf node, and then the average path length is employed to calculate the abnormal score. Finally, the user is evaluated by the value of the abnormal score, and then the user type is judged.

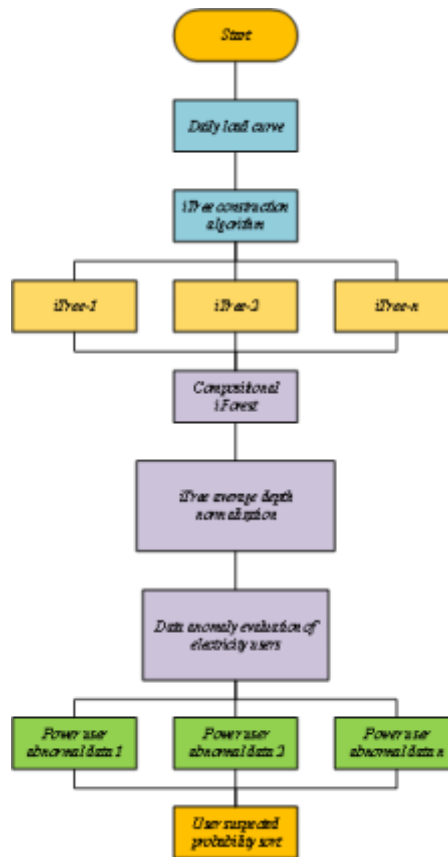


Fig. 2.3: The construction process of the iForest

2.3. Abnormal recognition of electricity consumption data based on the DT algorithm. The iForest algorithm is adopted to model the daily load curve of power users for abnormal detection, which is helpful to automatically screen users with abnormal data suspicion and realize the preliminary screening of abnormal users. Because the detection only by daily load curve is easy to cause misjudgment, the further analysis combined with other electrical variables of suspicious users can effectively improve the accuracy of detection.

The steps for the construction of the anomaly identification model of electricity power data based on the DT algorithm are demonstrated in Figure 2.4 [13].

The construction of the training set is mainly to sort out the date, meter number, and voltage data of the day in the EIAS, and sort out the transformer ratio and connection mode in the meter. Finally, the above data are combined to construct the training set of DT.

The DT algorithm is used to process the training set. Firstly, the training set is sorted, then it is divided by the threshold value of each data and the information gain is calculated. Furthermore, the threshold value is selected according to the maximum gain and the training set is divided.

The generation of DT: The root node and leaf node of DT correspond to a classification rule and synthesize all paths into a rule set, which is stored in a two-dimensional array.

Check the rationality of DT: the classification rules of DT are checked to see if there is any wrong decision. If so, the training set is adjusted until the classification is correct.

2.4. Update method for data exception. In actual work, the fault phase voltage of the data of the power acquisition system and the field measured data during the failure period is not 0, but eventually stabilizes

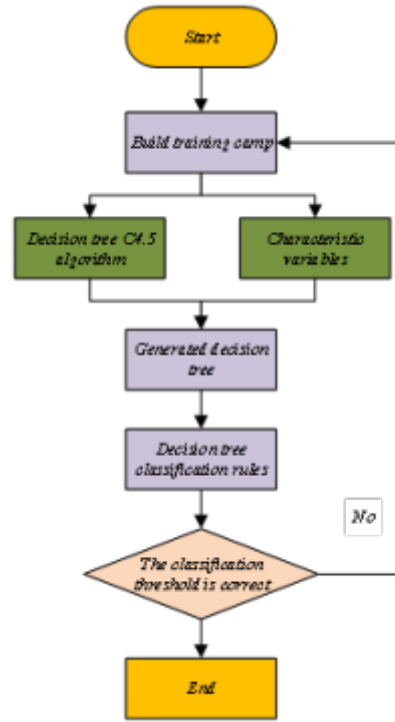


Fig. 2.4: The construction process of anomaly identification model for electricity consumption data

at a value not 0 with the change of time, which is called residual voltage [14, 15]. The residual voltage makes the amount of electricity measured by the energy meter during the voltage loss period contain the fault phase element and the amount of electricity measured by the non-fault phase element under the residual voltage. Thereupon, the calculation method of correction coefficient is determined by calculating the value of γ_{ab} , γ_{cb} according to the actual situation.

$$K = \frac{P_T}{P_F} = \frac{\sqrt{3}UI\cos\varphi}{\gamma_{ab}UI\cos(30^\circ + \varphi) + \gamma_{cb}UI\cos(30^\circ - \varphi)} \quad (2.8)$$

The determination of γ_{ab} and γ_{cb} is related to the metering principle of the intelligent energy meter, and the measured electric energy is illustrated in Equation 2.9:

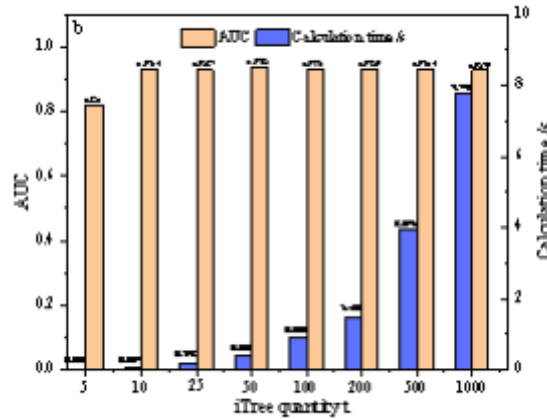
$$P = \frac{1}{T} \int_0^T u(t) \cdot i(t) dt \quad (2.9)$$

T represents the cycle of Alternating Current (AC) voltage and current.

By taking Δt as the sampling interval for voltage and current, the left discretization of Equation 2.9 is as follows:

$$P = \frac{1}{T} \sum_{k=1}^N u(k) \cdot i(k), T = N\Delta t \quad (2.10)$$

It can be seen that the determination of the correction factor is related to the current and voltage data during the failure. Through the EIAS, the frozen AC data of the Potential Transformer (PT) during the failure of voltage breakdown and loss can be known. Since the fusing on the PT side is independent of the current,



(b) Computation time and AUC

Fig. 3.1: Different iTree quantities

the calculation of the correction factor only needs to consider the voltage. The voltage curve can be obtained by freezing the voltage data of the electricity acquisition system. Through linear regression or piecewise linear regression on the voltage change curve, the average value γU of voltage during fault is obtained.

3. Results and Discussion.

3.1. Model parameter analysis. The iForest algorithm is mainly based on the thought of ensemble learning. There are two extremely important parameters in the anomaly detection model of electro-data, iTree sampling scale Ψ and the integration scale t . The simulation experiment is carried out on the system based on a Central processing unit (CPU) of dual-core 2.3GHz with 8GB memory, and the program is written through R language.

3.1.1. Quantity t of the iTrees . The iForest algorithm forms iForest by generating a certain number of itrees. It is mainly by means of random sampling to extract Ψ subset and construct iTree, and guarantee the diversity of iTree. Therefore, the number of iTree determines the size of the ensemble learning of the model. The Receiver operating characteristic curve (ROC) of the iTree with different numbers is portrayed in Figure 3.1.

Figure 3.1a portrays that the ROC curve is very close. The Area Under Curve (AUC) is obtained by calculating different numbers of iTrees respectively. According to the data in Figure 3.1b, the length of the path can be well covered when the number of iTrees reaches 100. After that, increasing the number of iTrees does not significantly improve the AUC, which is about 0.93 at this time. The cumulative recall ratio curve and Precision-Recall (P-R) curve of the iForest algorithm under different iTree numbers are shown in Figure 3.2.

In Figure 3.2a, when the number of iTrees is greater than 100, the gap between curves is very small. When the detection rate is less than 0.03, the curve has a very large upward trend, and when the detection rate is greater than 0.03, the curve tends to be flat. In the phase where the detection rate is less than 0.03, only the top 3% of users need to be detected to detect about 80% of abnormal users. At the stage where the detection rate is greater than 0.03, only 20% of abnormal users can be detected by detecting the remaining 97% of users. Thereby, the research focus of cumulative recall ratio is the stage where the detection rate is less than 0.03. As can be seen from the P-R curve in Figure 3.2b, when the iTree reaches more than 100, the precision ratio can exceed 80% when the recall ratio is 70%. Combined with Figure 3.2a, it can be found that 80% of abnormal users can be detected by detecting the top 2.5% of users for the abnormal score. When the detected users reach 3.5%, the precision ratio decreases significantly, only about 40%.

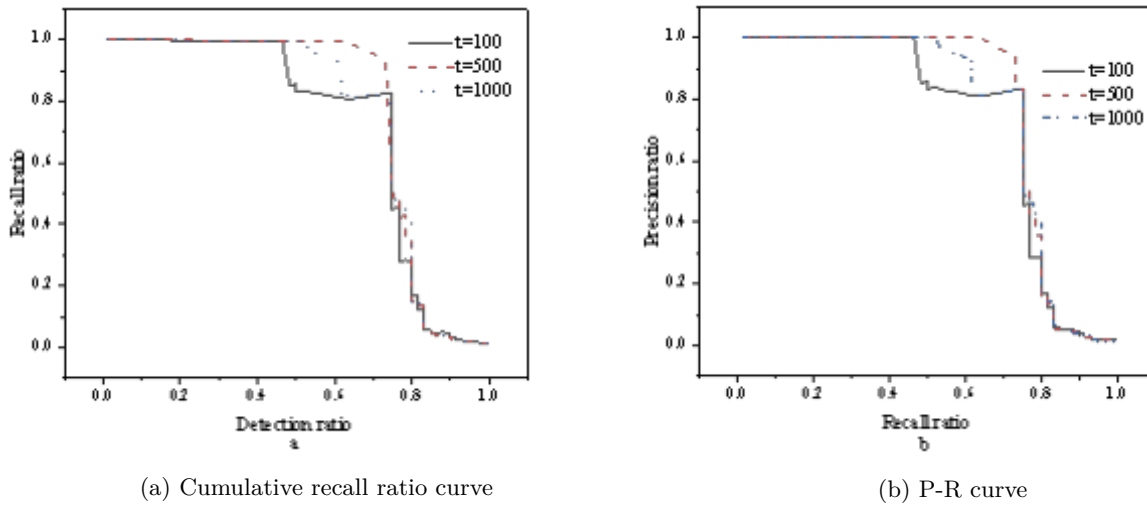


Fig. 3.2: The iForest algorithm with a diverse number of iTrees

3.1.2. The iTrees sample numbers Ψ . For any object in the data set, different sample numbers will affect the user's abnormal score and affect the final output of the model. Thus, it is important to study the sensitivity of model parameters. The relationship between the ROC curve and the iTrees sampling number is indicated in Figure 3.3.

In Figure 3.3a, when the collection number of iTrees is small, the performance of the model is poor, but when the number of iTrees reaches a certain value, the ROC curve will be very close. It can be seen from the data in Figure 3.3b that as the number of iTrees samples increases, the calculation time continues to increase. The area AUC under the ROC curve is not normal with the change of parameters, on the contrary, it decreases a little. The P-R and cumulative recall ratio curves of the iForest algorithm with diverse iTrees sampling numbers are implied in Figure 3.4.

The cumulative recall ratio curve in Figure 3.4a shows that when the sampling number is 100, 65% of abnormal users can be detected by detecting 2% of users. When the number of samples is much larger than 100, only 42% of abnormal users can be detected by detecting the top 2% of users. The P-R curve in Figure 3.4b can be more significantly found that when the sample numbers of iTrees is too large, the performance of the model is poor. The reason for the above phenomenon is primarily that the purpose of iTrees sampling is to better separate normal users from abnormal users, and the more sampled data, the worse the ability of the iForest algorithm to identify anomalies.

3.2. DT of voltage classification for power users. For the sake of explanation, the training set mainly covers 10kV three-phase measurement points. The model is constructed by using R language, and the DT of voltage classification is obtained, as plotted in Figure 3.5.

The output results in Figure 3.5 are voltage classification, where N, H, and L represent normal voltage, high voltage (HV), and low voltage (LV), respectively. The judgment condition of DT is the value between each node, and the whole DT mainly contains 10 decision points. Taking the left-most root node as an example, when the voltage value is less than 51V, the phase sequence is B and the connection mode is three-phase and three-wire, the voltage is normal; If the phase sequence is B and the connection mode is three-phase and four-wire, it is LV. If the phase sequence is A or C, it is LV. The confusion matrix of performance evaluation of the voltage classification model is expressed in Table 3.1.

Table 3.1 exhibits that there are 1166 data in the DT training set of voltage classification, and the classification accuracy is 100%. In general, the classification model is an overfitting phenomenon when the success rate

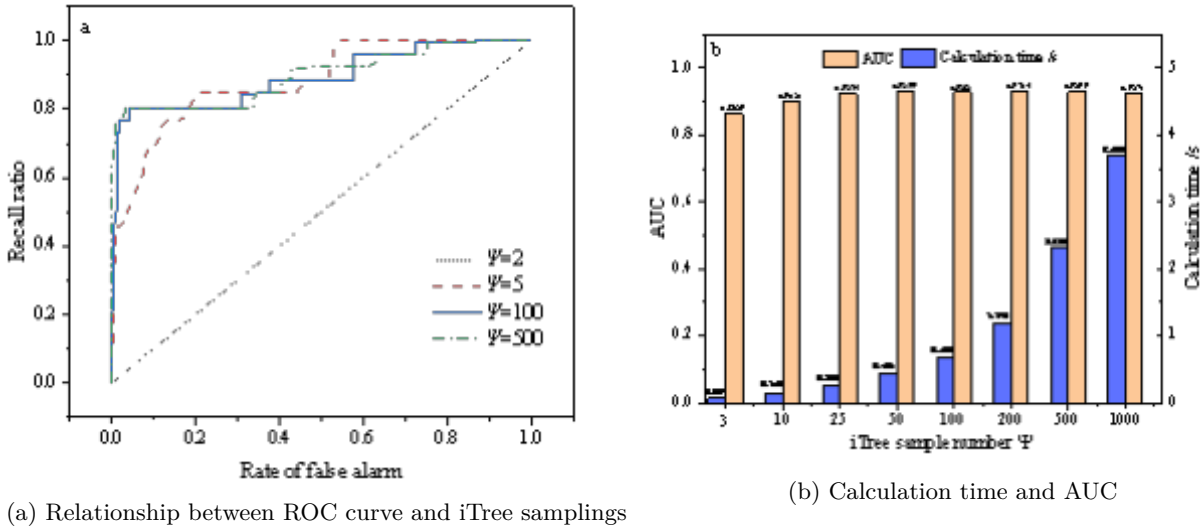


Fig. 3.3: The iForest algorithm with different iTree sampling numbers

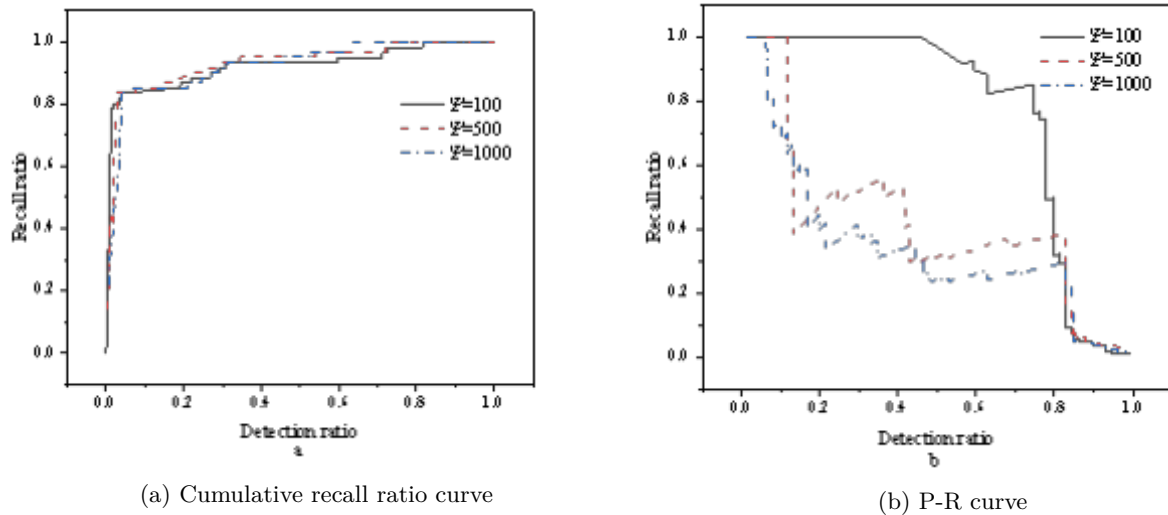


Fig. 3.4: The iForest algorithm with various iTree sampling numbers

of the classification reaches 100%, which is the characteristic of DT. The fitting phenomenon is more beneficial to the accurate classification of voltage between different measurement points, and the DT is more sensitive to the training set, which is conducive to the dynamic adjustment of voltage judgment rules.

3.3. Data anomaly detection of power users. Through iForest algorithm, data anomalies of users of special transformers in S City are detected. After preliminary screening, 179 suspected abnormal users are detected. According to the energy meter number of the abnormal user, the date and 96-point voltage in the EIAS of the user are sorted out. The voltage transformer ratio and connection mode of the meter in SG186 are also sorted out, and the above two parts are combined. Then DT algorithm is used to identify the voltage anomaly of the metering device of the suspected abnormal users, and 65 abnormal users are output. Finally,

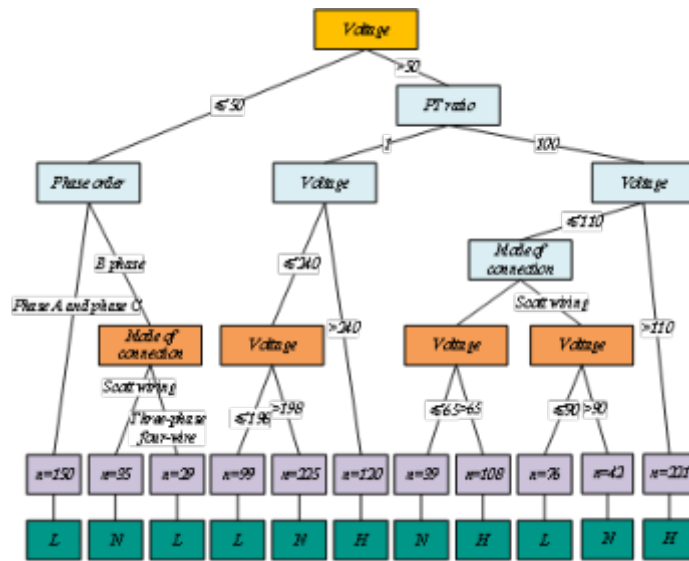


Fig. 3.5: The DT of voltage classification for users

Table 3.1: The confusion matrix of performance evaluation of the voltage classification model

Prediction	Normal voltage N	HV H	LV L
Normal voltage N	339	0	0
HV H	0	451	0
LV L	0	0	354

45 users with the highest degree of suspicion are sent out operation and maintenance work orders after manual review. After on-site verification, 37 abnormal users are identified. Figure 3.6 demonstrates the statistics of the feedback results of abnormal troubleshooting for users of special transformers:

3.4. Update of abnormal electricity consumption data for energy meters. Among the anomalies determined by the energy meter, the fuse failure of the voltage transformer is broken the most. Taking a case of fuse burn-out of a voltage transformer verified in S City as an example, it is found that there has an LV phenomenon through model recognition. The 96 points of voltage and current are obtained and sorted out from the EIAS. From A certain time period, the C-phase voltage has a large jump, the A-phase voltage is normal, and the current of the A and C phases is normal, indicating that the power supply of the user is normal, but the electric energy metering device is abnormal. The user data and file information obtained through the EIAS and SG186 are signified in Table 3.3:

According to the 96-point voltage data of the abnormal user, after the fuse on the primary side of C-phase PT is blown, the voltage between C and B phases on the secondary side quickly drops to about 14V and remains stable. The voltage data collected by the EIAS every 15 minutes during the fault period are statistically analyzed. It can be seen that the average voltage between C and B phases during the failure period is 14V, namely $\gamma_{cb} = 0.14$. It means that the power factor of the user is relatively stable, and the average value is $\cos\varphi = 0.85$, obtained after conversion $\varphi = 31.79^\circ$.

4. Conclusion. With the swift growth of the power industry, a variety of power equipment terminals have appeared. Based on the analysis and research status of electro-data anomalies, the data anomalies of energy meter terminals are studied. Firstly, the data of the daily load curve is preprocessed and the iForest algorithm is selected to construct the abnormal detection model of electricity data. Additionally, the accuracy

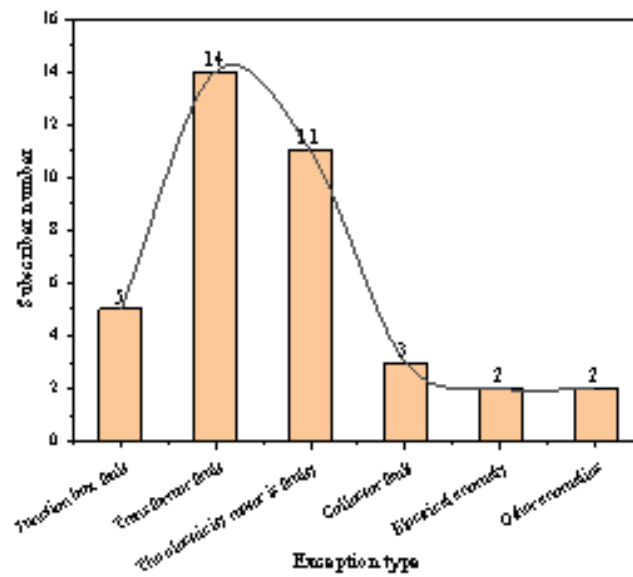


Fig. 3.6: Abnormal feedback results of electric energy metering device

and calculation time of the model are explored by analyzing the different values of the sample numbers of iTrees. Secondly, on account of the voltage and current data of suspected abnormal users initially screened, the anomaly identification model of electricity consumption data of the DT algorithm is implemented. Besides, the voltage at different metering points is classified and judged in combination with current data to identify abnormal energy metering devices. Finally, the case of an abnormal user in S City is analyzed, and the 96-point voltage data of the EIAS during the fault period are studied. Furthermore, the correction coefficient is adjusted on the basis of the participating voltage of the fault phase. The results manifest that the anomaly detection model of electricity power data constructed by the iForest algorithm has high validity and computational efficiency. The traditional recharge method does not consider the residual voltage, but the recharge method proposed here makes the recharge more accurate by determining the residual voltage and ensuring the accuracy of the charge. However, since only electrical variables such as voltage and current are considered in the analysis of anomalies in electricity data, the accuracy of identification needs to be improved. It is hoped that more in-depth exploration can be carried out in the subsequent research, and more variables can be introduced to improve the accuracy of recognition.

REFERENCES

- [1] Avancini D. B., Rodrigues J. J. P. C., Rabêlo R. A. L., et al. (2021) A new IoT-based smart energy meter for smart grids[J]. *International Journal of Energy Research*, 45(1), 189-202.
- [2] Santhosh C., Kumer S. V. A., Krishna J G., et al. (2021) IoT based smart energy meter using GSM[J]. *Materials Today: Proceedings*, 46, 4122-4124.
- [3] Butt O. M., Zulqarnain M., Butt T. M. (2021) Recent advancement in smart grid technology: Future prospects in the electrical power network[J]. *Ain Shams Engineering Journal*, 12(1): 687-695.
- [4] Zheng K., Chen Q., Wang Y., et al. (2018) A novel combined data-driven approach for electricity theft detection[J]. *IEEE Transactions on Industrial Informatics*, 15(3), 1809-1819.
- [5] Hasan M. N., Toma R. N., Nahid A. A., et al. (2019) Electricity theft detection in smart grid systems: A CNN-LSTM based approach[J]. *Energies*, 12(17), 3310.
- [6] Ji X., Yin Z., Zhang Y., et al. (2021) Real-time robust forecasting-aided state estimation of power system based on data-driven models[J]. *International Journal of Electrical Power & Energy Systems*, 125, 106412.
- [7] Liu X., Ding Y., Tang H., et al. (2021) A data mining-based framework for the identification of daily electricity usage patterns and anomaly detection in building electricity consumption data[J]. *Energy and Buildings*, 231, 110601.
- [8] Yan Z., Wen H. (2021) Electricity theft detection base on extreme gradient boosting in AMI[J]. *IEEE Transactions on*

- Instrumentation and Measurement, 70, 1-9.
- [9] Wang Z., Fu Y., Song C., et al. (2019) Power system anomaly detection based on OCSVM optimized by improved particle swarm optimization[J]. IEEE Access, 7, 181580-181588.
 - [10] Chapaloglou S., Nesiadis A., Iliadis P, et al. (2019) Smart energy management algorithm for load smoothing and peak shaving based on load forecasting of an island's power system[J]. Applied energy, 238, 627-642.
 - [11] Shi Y., Yu T., Liu Q., et al. (2020) An approach of electrical load profile analysis based on time series data mining[J]. IEEE Access, 8, 209915-209925.
 - [12] Zhang Y., Zhang J., Yao G., et al. (2020) Method for clustering daily load curve based on SVD-KICIC[J]. Energies, 13(17), 4476.
 - [13] Jiang J., Li T., Chang C., et al. (2022) Fault diagnosis method for lithium-ion batteries in electric vehicles based on isolated forest algorithm[J]. Journal of Energy Storage, 50, 104177.
 - [14] Charbuty B., Abdulazeez A. (2021) Classification based on decision tree algorithm for machine learning[J]. Journal of Applied Science and Technology Trends, 2(01), 20-28.
 - [15] Abd Rahman F. A., Ab Kadir M. Z. A., Ungku Amirulddin U. A., et al. (2021) Computation of energy absorption and residual voltage in a fourth rail LRT station arresters in EMTP-RV: A comparative study[J]. Urban Rail Transit, 7(2): 71-83.

Edited by: B. Nagaraj M.E

Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing

Received: Dec 27, 2023

Accepted: Mar 18, 2024



PROCESS TESTING AND ALGORITHM DETECTION ANALYSIS OF MECHANICAL STRENGTH OF ELECTROMECHANICAL COUPLING IN THE MAIN DRIVE OF ROLLING MILL

DONGBAO ZENG* AND HAI YAO[†]

Abstract. In order to deeply analyze the compressive strength of the electromechanical coupling system of the main drive of the rolling mill, this work takes the electromechanical coupling system of the main drive of the R2 rolling mill as a subject to study and analyze. First, experiments are designed to test the strength and some mechanical parameters of the main drive electromechanical coupling system of the R2 rolling mill, including the introduction of the main drive system of the R2 reversible rolling mill, test content, test scheme, test system framework and test materials. Next, this work analyzes the strength of the main drive shaft in the electromechanical coupling system of the main drive of the R2 rolling mill, and theoretically checks the fatigue strength of the main drive shaft, including the static strength analysis of the main drive shaft. Finally, this work analyzes the experimental data. The research shows that: (1) the peak value of the lower torque of the main drive coupling electromechanical of the R2 rolling mill is mainly distributed in the range of 650~1550 mk N. The average value of the upper torque is mainly distributed in the range of 350~1100 mk N. (2) The peak value of principal stress at 1[#] measuring point is within the range of 10-35Mpa, and the mean value of principal stress is within the range of -24-24Mpa. The peak value of principal stress at the 2[#] measuring point is in the range of 5-25Mpa, and the mean value of principal stress is in the range of -16-16Mpa. (3) When the upper drive shaft of the rolling mill runs at a working angle of 1.02°, the maximum Von Mises Stress of the drive shaft of the rolling mill is 95.10MPa. When the drive shaft of the rolling mill operates at a working angle of 3.14°, the maximum Von Mises Stress of the drive shaft of the rolling mill is 95.15MPa. This work aims to provide a theoretical reference for enhancing the compressive strength of the electromechanical coupling system of the main drive of the rolling mill.

Key words: electromechanical coupling system of the main drive of the rolling mill, compressive strength, main drive shaft, principal stress, fatigue strength

1. Introduction. The main characteristics of modern rolling mills are large, heavy load, high speed, continuous automation, precision and poor working conditions. These characteristics require the long service life of the main parts of the rolling mill. In order to achieve this goal, it is the most effective and convenient means to use modern mechanical methods and fully use the idea of the finite element method (FEM) and its generated software to study classical mechanical problems. Meanwhile, it is essential to apply advanced test technology for equipment diagnosis to provide conditions for analyzing the mechanical behavior of equipment [1]. In the past decade, the rapid development of computer hardware technology and the rapid decline of computer costs have greatly driven the development of finite element calculation software, and the function of display programs has also been increasingly enhanced. It provides a broad platform for the use of FEM to solve practical engineering problems [2]. Besides, as early as the 1960s, foreign steel companies noticed the impact of torsional vibration on the main drive system of the rolling mill, and have conducted massive experimental studies. At that time, the research level was limited to experimental measurement and the establishment of simple mechanical models, and there was no more in-depth theoretical exploration [3]. For example, in the 1960s, the Bethlehem Steel Corp of the United States conducted field tests and theoretical analysis on the rolling mill, and analyzed the resonance problem caused by the coupling of the frequency of the rolling mill drive system and the natural vibration frequency of the motor riser [4]. In 1973, the Association of Iron & Steel Engineers in America and Jones&Lauhlin Iron and Steel Company studied the torsional vibration of a rolling mill. They reached the following conclusions: the head of the hot top is conducive to reducing part of the mechanical coefficient; it is important to eliminate the clearance in the transmission link to reduce the value; shear parts should not be used as mechanical insurance parts [5].

*Jiangxi Technical College of Manufacturing, Nanchang, Jiangxi, 330095, China (DongbaoZeng20163.com)

[†]Mechanical College, Shanghai Dianji University, Shanghai, 201306, China. (Corresponding author, HaiYao17@126.com)

Table 2.1: Main technical parameters of R2 reversible rolling mill

Parameter name	Specific value
Working roll diameter	1200/1100 mm
Support roll diameter	1600/1440 mm
Roller length	2250 mm
Rolling force	45000 kN
Motor power	2×7500 KW
Motor speed	0~45/100 rpm
Balance coefficient of balance cylinder	1.1 (test condition), 1.3 (damage condition)

It was not until the 1970s that the torsional vibration of the rolling mill was paid attention to in China. In particular, several serious equipment accidents related to system torsional vibration have occurred in several domestic steel rolling plants in the past decade, which has caused people to study the impact of torsional vibration load on the system [6]. For example, the University of Science and Technology Beijing has studied the main drive system of the blooming mill of Baotou Iron and Steel Company and the second blooming mill of Anshan Iron and Steel Company. It is found that the self-excited vibration caused by roll slipping is the main cause of the main drive torsional vibration damage [7]. Professors at the University of Science and Technology Beijing adopted the rolling mill's shafting parameter optimization design method to predict and reduce the torque amplification factor in the design [8]. Wuhan University of Science and Technology studied the vibration of the arc joint shaft of the 1700 cold chain mill drive system of Wuhan Iron and Steel Company. It is found that the rolling speed greatly impacts the vibration of the shaft. The misalignment of the arc joint shaft causes polarization and gearbox meshing excitation [9].

In the rolling system, due to the manufacturing process, assembly and other factors, there is often a certain gap at the system's connection, which will make the stiffness of the main transmission system behave as nonlinear. Hence, the system cannot be simply described by a linear model. Avoiding the resonance synchronization phenomenon in the working process of the system, the reasonable control of the main drive system of the rolling mill, the establishment of the vibration simulation mechanical model and the torsional vibration of the system in case of slipping. Based on the above contents, for the frequent fracture accidents and fatigue damage of the main drive system of the rolling mill, this work takes the R2 four-roll reversing roughing mill of a certain unit as the research object. Then, based on the field test, combined with the worldwide torsional vibration research of the main drive system of the rolling mill, the dynamic characteristics analysis of the main drive system of the R2 rolling mill is conducted. Moreover, the FEM analysis software ANSYS is adopted to carry out three-dimensional (3D) FEM analysis on the spider universal joint of the main drive system of the R2 rolling mill. Moreover, the strength and fatigue check of the drive shaft in the main drive system of the R2 rolling mill are analyzed in detail. This work aims to design relevant simulation experiments to study the vibration of a rolling mill and analyze its strength to make a beneficial combination and supplement between theory and practice. It has crucial engineering practical significance and theoretical research significance.

2. Design of process test and simulation model.

2.1. Parameters test of mechanical mechanics in main drive electromechanical coupling system of R2 rolling mill .

(1) Main drive system of R2 reversible rolling mill.

1. Schematic diagram of the main drive system of the R2 reversing rolling mill

The main drive system of the R2 reversible rolling mill adopts two synchronous motors to directly drive the upper and lower rolls. Figure 2.1 is the structure diagram.

2. Main technical parameters of R2 reversible rolling mill

Table 2.1 displays the main technical parameters of the R2 reversible rolling mill.

(2) *Field test content.* The test content of this experiment mainly includes the torque M1 and M2 of the upper and lower drive shafts of the R2 reversible rolling mill, the stress σ of the lower connecting shaft, and rolling force P.

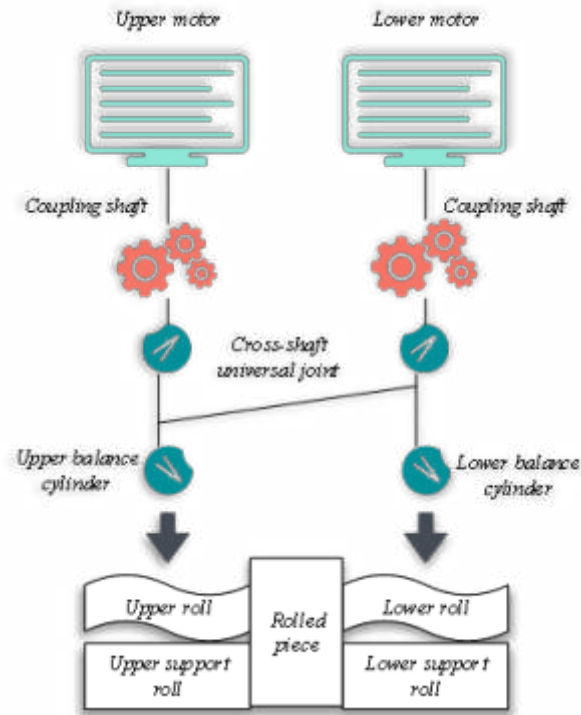


Fig. 2.1: Structural diagram of the main drive system of the R2 rolling reversible mill

(3) Test plan.

1. Measuring point position and testing method of main drive torque of R2 rolling mill:

The test positions of the upper drive shaft torque and the lower drive shaft torque are as follows.

The measuring point is located at the hollow shaft section with an outer diameter of $D=930\text{mm}$ and an inner diameter of $d=736\text{mm}$. On this shaft section, four resistance pieces are pasted in the direction of $\pm 45^\circ$ with the axis to form a full-bridge test circuit. The signal is input into the YD-28A dynamic strain gauge through the slip ring, and then recorded by the computer after A/D conversion. The calibration is carried out with a constant-strength graduated beam. Resistance strain gauges, R1+R3 and R2+R4, are respectively attached to the constant strength beam's upper (tension) and lower (compression) sides, forming a full-bridge test circuit. Then, it is calibrated with standard weight loading. The torque calibration value of the upper drive shaft is $K=24.237\text{MPa/V}$, and the torque calibration value of the lower drive shaft is $K=24.190\text{MPa/V}$.

2. R2 rolling force:

After the rolling force P of the R2 rolling mill is isolated through the isolator, it is led from the main electrical room to the test site through the shielded cable and connected to the computer. The calibration value of the rolling force is 450t/V (provided by a two-roll hot-rolling electrical workshop).

3. Location and test method of main drive stress measuring point of R2 rolling mill:

The stress is measured through resistance strain. Because the stress situation of the shaft is complex and the direction of the principal stress is unknown, the strain rosette is adopted to measure the stress at this point. The distance between the main drive shaft 1[#] stress measuring point and the center line of the main shaft arm is 793 mm. The signals corresponding to the strain gauges in the horizontal direction (0°), 45° and vertical direction (90°) are 3[#], 4[#] and 5[#], respectively. The distance between the main drive shaft 2[#] stress measuring point and the center line of the main shaft arm is 643mm. The

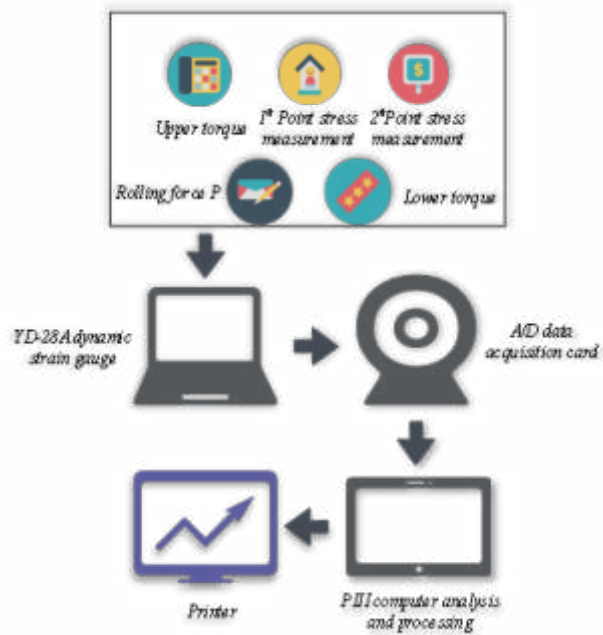


Fig. 2.2: Test system framework

signals corresponding to the strain gauges in the horizontal direction (0°), 45° and vertical direction (90°) are 2#, 7# and 6#, respectively. Considering that the temperature of the measuring point will change during the rolling process, in order to compensate for the test error caused by the temperature change, a small steel block (without stress) is placed near the measuring point as the temperature compensation block (the same as the temperature of the measuring point). Three resistance strain gauges are pasted on each temperature compensation block, and a half-bridge test circuit is formed with the three working pieces of the strain rosette at the measuring point.

The calibration is conducted with a constant-strength graduated beam. On the constant-strength graduated beam, a resistance strain gauge R1 is pasted and a temperature compensation block is placed near the measuring point. The resistance strain gauge R2 is attached on the temperature compensation block, and R1 and R2 form a half-bridge test circuit. The stress calibration value of each measuring point can be obtained by using standard weight loading for calibration. The calibration value of 3# is $K=97.087\text{MPa/V}$, that of 4# is $K=97.276\text{MPa/V}$, that of 5# is $K=96.712\text{MPa/V}$, that of 2# is $K=96.900\text{MPa/V}$, that of 7# is $K=97.466\text{MPa/V}$, and that of 6# is $K=97.087\text{MPa/V}$.

(4) *Test system block diagram.* The stress signal of each measuring point is connected to the dynamic strain gauge and recorded by the computer after A/D conversion. The sampling frequency of each channel is 100Hz. Figure 2.2 displays the system test block diagram.

(5) *Test materials.* The force and energy parameters of the main drive system during the rolling process of 174 steel billets of 11 rolling varieties produced by a certain unit were tested. Table 2.2 shows the specification and quantity of the test billet.

2.2. Research on strength and fatigue strength check of the main drive shaft of the R2 rolling mill.

(1) *Mechanical analysis of the main drive shaft of the rolling mill.*

1. FEM software

With the continuous progress of computer technology, the theoretical technology of Computer Aided Engineering (CAE) is also maturing [10]. Many kinds of FEM software can be used in the project,

Table 2.2: Specification and quantity of test billet

Material and model	Product specification (thick \times wide \times Length) (mm)	Slab quantity (block)
DC01	250 \times 1300 \times 8600	2
Q235A	230 \times 1550 \times 8600	2
RCL380	230 \times 1550 \times 8000	5
Q235A	230 \times 1550 \times 10800	3
St52-3	230 \times 1700 \times 10000	3
Q345A	230 \times 1597 \times 10500	1
Q235A	230 \times 1550 \times 10800	8
Q235B	230 \times 1550 \times 9950	14
SAE1008	230 \times 1600 \times 9950	22
SPHC	230 \times 1550 \times 10800	5
08AL	230 \times 1550 \times 10800	4
RCL380	230 \times 1550 \times 10800	6

such as Hyper Mesh, ANSYS, and ANSYS Workbench [11,12,13]. Here, Hyper Mesh and ANSYS are adopted to analyze the main drive shaft of the rolling mill under dangerous working conditions.

FEM software ANSYS is an excellent numerical analysis and calculation software. In the field of linear computing, the position of ANSYS is quite stable [14]. ANSYS is a multi-physics field analysis package that integrates structure, fluid, electromagnetism and acoustics. Its common applications include industrial manufacturing, nuclear industry, household appliances, aerospace, and biomedicine. It has the most users in the world and is also a very successful FEM software [15].

Hyper Mesh software is a general FEM analysis software developed by Altair. It can be applied in the fields of modeling, visualization, automation and manufacturing, and is recognized and applied by the world industry. The shape of the drive shaft and bearing seat in the main drive device of the rolling mill is not too regular. Rough grids have low credibility and are likely to lead to incorrect calculation results. Since the quality of finite element mesh will directly affect the accuracy of final calculation results, finite element mesh is the key to analysis and calculation. The mesh quality greatly impacts the calculation results, and Hyper Mesh has the characteristics of a high-quality mesh division. It has a complete interactive two-dimensional and 3D cell division tool [16]. The user can adjust the mesh parameters of each face during the division process, such as cell density, cell offset gradient and mesh division algorithm. The 3D cell generation method provided by Hyper Mesh is employed to build high-quality tetrahedral and hexahedral meshes and Computational Fluid Dynamics meshes [17].

2. FEM model creation

The R2 rolling mill used in this section has a motor power of 7500kw and a motor speed of 45~90rpm. According to the calculation equation, the maximum torque output at the motor end of the mill drive shaft is 1591000N \cdot m. According to the functional relationship and parameters between the balance force on the drive shaft of the rolling mill and the driving force of the hydraulic cylinder, it is calculated that the balance force provided by the bearing seat is 348520N under the dangerous angle of the upper drive shaft of the rolling mill. The balance force provided by the bearing seat of the lower drive shaft of the rolling mill at a dangerous angle is 350100N. The material of the drive shaft in the main drive device of the rolling mill is 42Cr Mo. According to the manual, the elastic modulus of the drive shaft material in the main drive device of the rolling mill is 2.06×10^{11} Pa, Poisson's ratio is 0.3, the maximum tensile strength of the material is 1080MPa, and the yield limit of the material is 930MPa. Other properties of materials can be calculated by empirical equation, such as torsion fatigue limit 301.5MPa, and fatigue limit 542.7MPa.

According to the calculated dangerous working condition angle of the rolling mill drive shaft, in the 3D software Solid Works, the drive shaft and bearing seat are virtually assembled. Then, it is saved to the file in .iges format to facilitate the reading of FEM software Hyper Mesh.

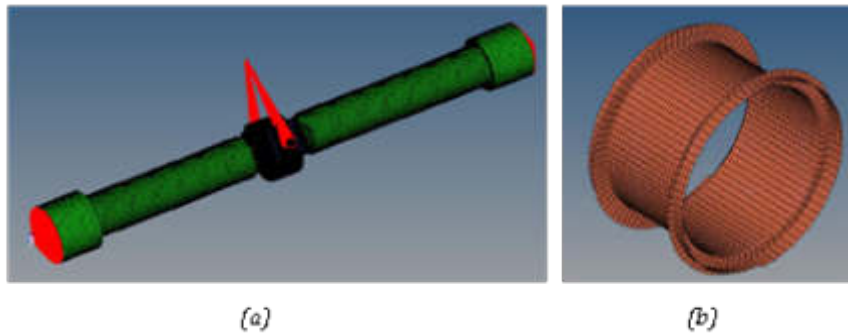


Fig. 2.3: Mesh element model and contact element (a) Mesh element model; (b) Contact element

FEM software Hyper Mesh is started. Because the solution and calculation phase need to be carried out in ANSYS, the ANSYS panel is selected in the Load User Profile. When importing a model, the iMPortgeometry option is selected to import the .iges format file previously saved through Solid Works into Hyper Mesh. The mesh of the rolling mill's imported main drive shaft and the bearing seat model are divided. In order to make the divided element mesh neat and regular, the area where the balance force is added in the bearing seat is coupled with the bearing seat by establishing a contact surface, as shown in Figure 3(a). The cell type selected is SOLID185 cell. There is contact between the drive shaft and the bearing seat, so it is essential to establish contact between the bearing and the drive shaft, as shown in Figure 3(b). The contact pairs are CONTA173 and TARGE170.

Then, the dangerous angle of the upper drive shaft and the lower drive shaft of the rolling mill can be judged according to the theoretical calculation results. Then, through rotation in Hyper Mesh, the data file of dangerous working conditions of the main drive shaft of the rolling mill is obtained. In order to add the constraints of the two dangerous angles of the main drive shaft of the rolling mill correctly, it is necessary to establish a local coordinate system at both ends of the drive shaft. The number of the local coordinate system at the motor end and at the roll section is set to 20 and 21, respectively, to prevent errors in the simulation. The constraint added at the motor end is to limit the three degrees of freedom of displacement of the motor end face. The constraint added at the roller end is to limit the three degrees of freedom of displacement of the roller end and the degrees of freedom of rotation in the Z-axis direction (axial direction).

3. Static strength simulation analysis

After the preprocessing of Hyper Mesh, it is essential to open the saved data file in ANSYS and solve it. After the solution is completed, the stress change nephogram of the main drive shaft of the rolling mill can be obtained under the dangerous working angle of the drive shaft in the main drive device of the rolling mill. The material selected is 42Cr Mo, which is plastic material. The fourth strength theory (Von Mises) is used as the criterion for strength judgment. It is believed that the density of distortion energy or the specific energy of shape change is the main reason for material yield failure. Von Mises Stress is output in the post-processing [18].

(2) *Fatigue strength check of the main drive shaft of the R2 rolling mill.* First, the final element calculation of the drive shaft at a dangerous angle is conducted to determine whether the design strength of the drive shaft meets the requirements. Then, through the analysis of the working state of the drive shaft, the bending moment of the drive shaft of the rolling mill is the alternating stress. As the rolling mill is reversible, the torque of the drive shaft is also alternating stress. The fatigue strength of the drive shaft at the dangerous angle of the mill drive shaft is checked.

According to the fatigue limit σ_{-1} of the material under symmetrical cyclic alternating stress, the fatigue

limit $(\sigma_{-1})_G$ and allowable stress $[\sigma_{-1}]$ of the drive shaft can be calculated. Their values are:

$$(\sigma_{-1})_G = \frac{\varepsilon_\sigma \beta}{K_\sigma} \sigma_{-1} \quad (2.1)$$

$$[\sigma_{-1}] = \frac{(\sigma_{-1})_G}{n} \quad (2.2)$$

n represents the allowable safety factor. Generally, according to the regulations, when the material quality is uniform and the calculation accuracy is high, $n=1.3\sim 1.5$. According to the engineering calculation, the allowable safety factor is 1.5. ε_σ is the size factor of the component, with a value of 0.53. β is the surface processing coefficient, and the value is 1.

When the rolling mill is at a dangerous working angle, the bending moment of sections A-A and B-B are calculated. Since the direction of gravity is not perpendicular to the axis direction, the gravity component perpendicular to the axis direction is adopted to calculate the bending moments of the transmission shaft sections A-A and B-B when calculating the bending moments of the above sections. Finally, according to the theoretical data, the maximum working stress of the rolling mill drive shaft at the dangerous section A-A and B-B can be calculated.

In order to calculate the safety factor closer to the real situation, the maximum working stress and maximum shear stress of the dangerous section are extracted from the post-processing of ANSYS. The post-processing interface of ANSYS is opened. In the command flow window of ANSYS, RSYS,20 are input to make the stress distribution of the drive shaft along the axial direction. It is convenient to extract the stress distribution diagram of the rolling mill drive shaft in the dangerous area to complete the fatigue strength check of the rolling mill drive shaft.

3. Simulation experimental results.

3.1. Test results of mechanical mechanics parameters of main drive electromechanical coupling system of R2 rolling mill .

(1) *Distribution probability of peak and average rolling force.* Figure 3.1 presents the distribution probability results of the peak and average rolling force of the electromechanical coupling system of the main drive of the R2 rolling mill.

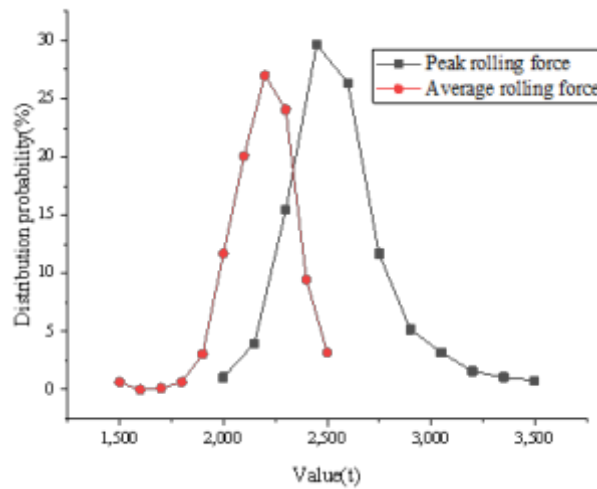


Fig. 3.1: Distribution probability diagram of peak and average rolling force

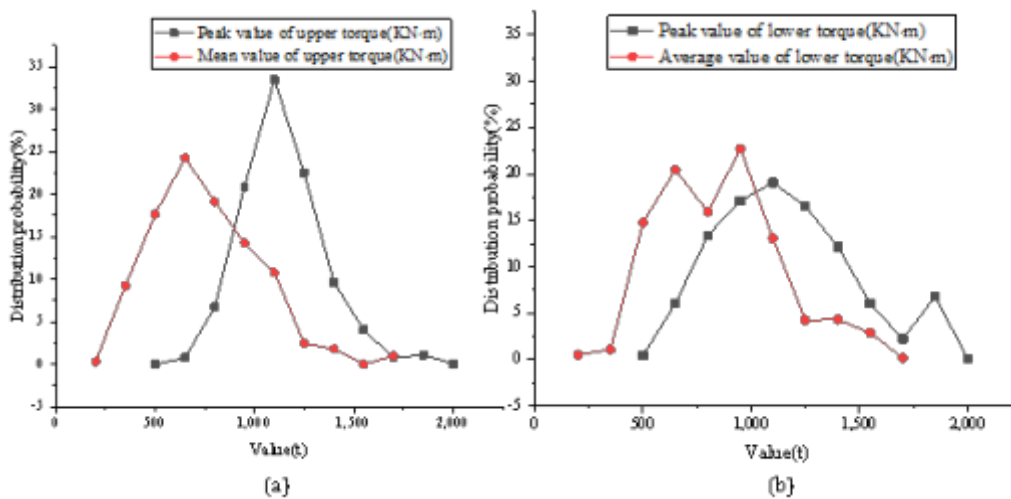


Fig. 3.2: Distribution probability of torque peak value and average value (a) Distribution probability of peak value and the average value of upper torque; (b) Distribution probability of the peak and the average value of the lower torque

Figure 3.1 reveals that the measured average value of rolling force is in the range of 1230~2660t, and the peak value is in the range of 1400~3620t. The results of the probability distribution of the peak rolling force show that the peak rolling force is mainly distributed in the range of 2300~2750t. The average distribution probability of rolling force suggests that the average distribution of rolling force is mainly in the range of 2000~2400t.

(2) *Distribution probability of torque peak and average.* Figure 3.2 displays the probability results of the peak and average distribution of the upper torque and lower torque of the electromechanical coupling system of the main drive of the R2 rolling mill.

Figure 3.2 illustrates that the peak value of the upper torque is mainly distributed in the range of 800~1400

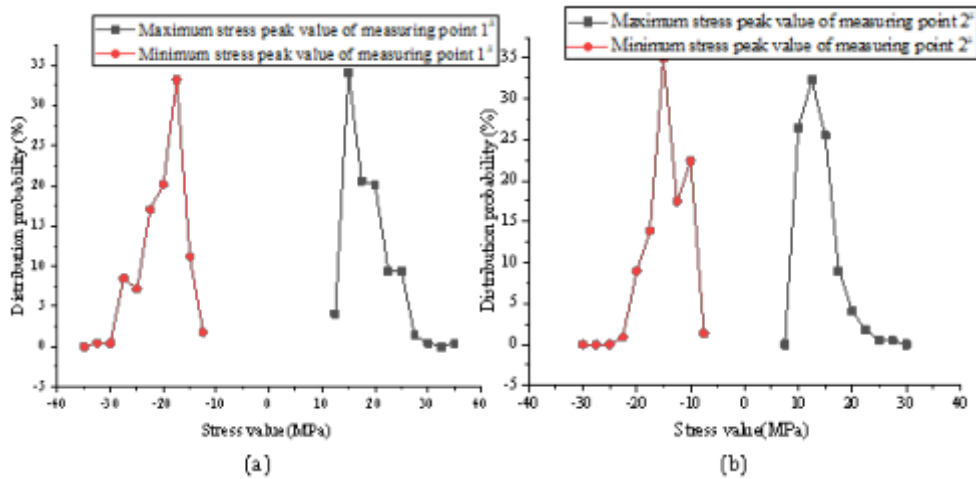


Fig. 3.3: Distribution probability of stress peak at different measuring points (a) Distribution probability of stress peak at measuring point 1[#]; (b) Probability of stress peak distribution at measuring point 2[#]

mk N, and the average value of the upper torque is mainly distributed in the range of 350~1100 mk N. The peak value of the lower torque is mainly distributed in the range of 650~1550 mk N, and the average value of the lower torque is mainly distributed in the range of 500~1100 mk N.

(3) *Distribution probability of stress peak.* The fatigue of metal structure is closely related to the main cycle characteristics of structural stress, and the small fluctuation of stress has little effect on fatigue life. Hence, the change of stress amplitude at each measuring point during the rolling process is recorded, and the stress equivalent curve of each measuring point is made.

Figure 3.3 displays the distribution probability of the maximum principal stress peak at different measuring points (near the drive side) after testing.

Figure 3.3 suggests that the maximum principal stress peak value of 1[#] measuring point is within the range of 10-35Mpa, and the minimum principal stress peak value of 1[#] measuring point is within the range of 10-30Mpa. The maximum principal stress average value is 24Mpa, and the minimum principal stress average value is -24Mpa. The maximum principal stress peak value of 2[#] measuring point is in the range of 5-25Mpa, the minimum principal stress peak value of 2[#] measuring point is in the range of 8-22Mpa, the maximum principal stress average value is 16Mpa, and the minimum principal stress average value is -16Mpa. During idling, the maximum and minimum principal stress values of 1[#] and 2[#] measuring points are 8Mpa.

(4) *Torsional vibration power change of main drive system of R2 rolling mill.* The first natural frequency of the upper and lower main drive systems can be obtained by power spectrum analysis of the measured torsional vibration waveforms of the upper and lower main drive systems, as shown in Figure 3.4.

Figure 3.4 suggests that the torsional vibration power spectrum of the upper and lower main drive systems of the R2 rolling mill is basically similar, and both reach a peak at about 17.5Hz.

3.2. Strength simulation analysis results of the main drive shaft of the R2 rolling mill. Figure 3.5 presents the stress distribution nephogram of the drive shaft of the rolling mill when the upper drive shaft is at different working angles.

Figure 3.5 suggests that when the upper drive shaft of the rolling mill is at a working angle of 1.02°, the maximum Von Mises Stress of the drive shaft of the rolling mill is 95.10MPa. It appears in the transition shaft shoulder’s fillet area where the main drive shaft of the rolling mill contacts the bearing seat. When the lower drive shaft of the rolling mill is at a working angle of 3.14°, the maximum Von Mises Stress of the rolling mill drive shaft is 95.15MPa. It appears in the transition shaft shoulder’s fillet area where the main drive shaft of the rolling mill contacts the bearing seat. The analysis reveals that the Von Mises Stress value of the rolling mill drive shaft is far less than the yield limit of 930MPa.

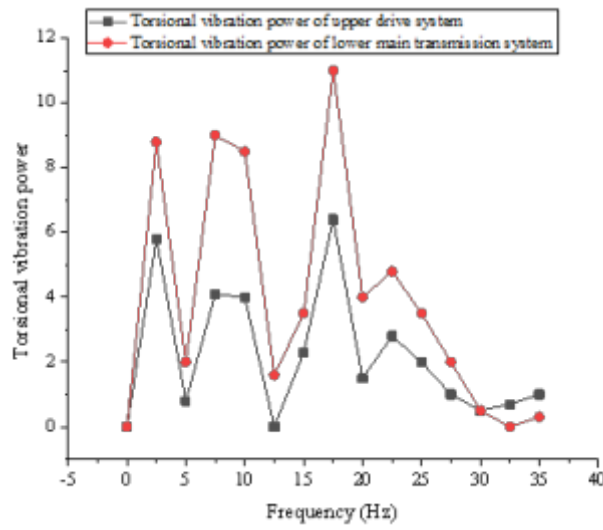


Fig. 3.4: Torsional vibration power spectrum of upper and lower main drive systems of R2 rolling mill

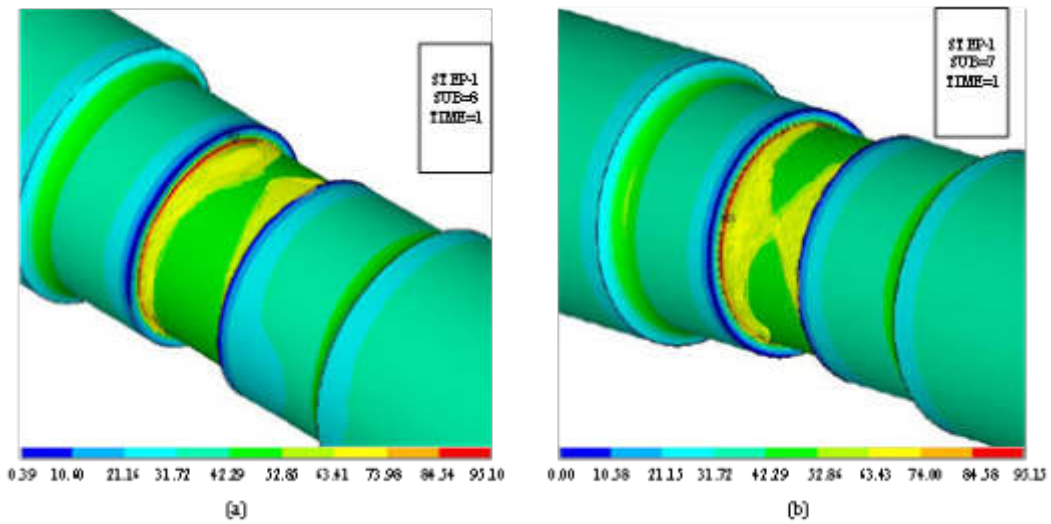


Fig. 3.5: Stress distribution nephogram of rolling mill drive shaft under different working angles (a) Stress distribution nephogram of rolling mill drive shaft at 1.02° working angle; (b) Stress distribution nephogram of the lower drive shaft in the main drive device of the rolling mill at a working angle of 3.14°

4. Conclusion. This work further analyzes and tests the compressive strength of the electromechanical system of the main drive coupling of the R2 rolling mill through testing and simulation experiments based on previous scholars' research on the electromechanical system of the main drive coupling of the rolling mill. Besides, the strength of the main drive shaft in the main drive coupling electromechanical system of the R2 rolling mill is analyzed. The research results show that the peak value of the lower torque of the main drive coupling electromechanical of the R2 rolling mill is mainly distributed in the range of 650~1550 mk N, and the average value of the upper torque is mainly distributed in the range of 350~1100 mk N. When the upper drive shaft of the rolling mill runs at a working angle of 1.02°, the maximum Von Mises Stress of the drive shaft of

the rolling mill is 95.10MPa. When the drive shaft of the rolling mill operates at a working angle of 3.14° , the maximum Von Mises Stress of the drive shaft of the rolling mill is 95.15MPa. The research deficiency is that only the whole part of the main drive coupling electromechanical system of the R2 rolling mill and the strength of the rotating shaft have been studied, while the strength of other structures has not been studied. Later, it will be studied and analyzed to provide some reference for improving the strength of the main drive coupling electromechanical system of the rolling mill.

REFERENCES

- [1] Sanjari R. K. S. S. (2021) Semi-quantitative health risk assessment of exposure to chemicals in an aluminum rolling mill[J]. *International Journal of Occupational Safety and Ergonomics*, 27(2), 22-23.
- [2] Ramirez-Tamayo D., Soulamy A., Gupta V., et al. (2021) A complex-variable cohesive finite element subroutine to enable efficient determination of interfacial cohesive material parameters[J]. *Engineering Fracture Mechanics*, 247(14), 107638.
- [3] Nazaretov A. A., Yaitkov I. A., Chukarin A. N. (2021) The Derivation of the Noise Level Dependences and Vibration Velocities of Elements and Units of the Drive System of the Rail Grinding Machines[J]. *IOP Conference Series: Earth and Environmental Science*, 720(1), 012014.
- [4] Zhang Z., Ji W., Yang B., et al. (2022) Dynamic analysis and vibration reduction of mechanical-hydraulic coupled tunnel boring machine (TBM) main drive system:[J]. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 236(1), 115-125.
- [5] Zhang G., Bao J., Li W., et al. (2021) Coupled Vibration Characteristics Analysis of Hot Rolling Mill with Structural Gap[J]. *Shock and Vibration*, 2021(3), 1-10.
- [6] Liu Z., Zhang X., Zhu Z., et al. (2021) Study on Dynamic Amplification Factor of UHV Pillar Equipment[J]. *IOP Conference Series Earth and Environmental Science*, 791(1), 012141.
- [7] Thomas G., Campbell O., Nichols N., et al. (2021) Formulating and Deploying Strength Amplification Controllers for Lower-Body Walking Exoskeletons. [J]. *Frontiers in robotics and AI*, 8(3), 720231.
- [8] Wang Q. H., Liu X., Wang D. N. (2021) Ultra-sensitive gas pressure sensor based on vernier effect with controllable amplification factor[J]. *Optical Fiber Technology*, 61(13), 102404.
- [9] Fanaie N., Nadalipour Z., Sarkhosh O. S., et al. (2021) Elastic drift amplification factor in steel moment frames with double reduced beam section (DRBS) connections[J]. *Journal of Building Engineering*, 43(43), 1-22.
- [10] Rsel G. (2021) Strength-based design of a fertilizer spreader chassis using computer aided engineering and experimental validation[J]. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 235(12), 095440622199384.
- [11] Tai J., Li H., Guan Y., et al. (2021) Simulation of a maize ear picking device with a longitudinal horizontal roller based on hypermesh modeling[J]. *Bioresources*, 16(1), 1394-1410.
- [12] Lin Y., Xie X., Yan L., et al. (2022) Research on the collapse of Tacoma narrows bridge under the finite element application ANSYS of computational mechanics[J]. *Journal of Physics: Conference Series*, 2230(1), 012018.
- [13] Li J., Pan M., Sun K., et al. (2022) Mechanical Characteristics Analysis of Grinding Plate of Food Waste Grinding Mill Based on ANSYS Workbench[J]. *Journal of Physics: Conference Series*, 2152(1), 012021.
- [14] Adhikari N., Alexeenko A. (2021) Development and Verification of Nonequilibrium Reacting Airflow Modeling in ANSYS Fluent[J]. *Journal of Thermophysics and Heat Transfer*, 2021(1):1-11.
- [15] Khan N. B., Ibrahim Z. B., Ali M. A., et al. (2021) Numerical simulation of flow with large eddy simulation at $Re = 3900$ [J]. *International Journal of Numerical Methods for Heat & Fluid Flow*, 30(5), 2397-2409.
- [16] Wang M., Dong M., Lu S., et al. (2021) Modal analysis of excavator toolbox based on hypermesh[J]. *Journal of Physics Conference Series*, 1965(1), 012037.
- [17] Zhang T. (2021) Lightweight Analysis Based on the Hypermesh Frame[J]. *Mechanical Engineering and Technology*, 10(3), 407-418.
- [18] Ji D., Hu X., Zhao Z., et al. (2022) Stress Rupture Life Prediction Method for Notched Specimens Based on Minimum Average Von Mises Equivalent Stress[J]. *Metals*, 12(1), 68.

Edited by: B. Nagaraj M.E

Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing

Received: Dec 28, 2023

Accepted: Mar 20, 2024



RESEARCH ON POWER LINE COMMUNICATION BASED ON DEEP LEARNING FOR ELECTROMECHANICAL EQUIPMENT ELECTRICITY ACQUISITION TERMINALS

CHENGFEI QI^{*}, XIAOBO YANG[†], XIAOKUN YANG[‡], CHAORAN BI[§] AND WENWEN LI[¶]

Abstract. The purpose is to use power line communication technology to re-optimize the power acquisition terminals of electromechanical equipment and improve the efficiency of information acquisition and management of the system grid. This paper analyzes the power and signal composite modulation mode of power line data communication in distributed power grids. Combined with the topology of the communication network in the power transmission process, the management mode of integration of power lines and wireless communication equipment is redesigned. Firstly, the characteristics of power line communication and wireless channel are analyzed. Aiming at the problem that most communication network operators use the fixed relay for communication, the main network communication mode is selected for optimization. Then, the information fusion method is adopted to integrate the network structure of the enterprise, network, and physical layer. The management of wireless communication equipment is carried out by reasonably allocating power resources. Additionally, the structure of the power acquisition terminal model is designed based on strict standards and practice. Finally, the communication fusion method is used for experimental simulation. The results show that when the input current is 1A, the experimental, theoretical value of the system is 3A, and the actual instrument output is close to 5A. When the input current is 6A, the instrument output of the system is 7.5A. Therefore, loading a reasonable load impedance value in the system can optimize the current output value of the model. The paper has important reference value for optimizing electromechanical equipment and power acquisition terminal.

Key words: power line communication, electromechanical equipment, power acquisition terminal, management efficiency, communication integration

1. Introduction. WWith the development of power line communication technology and the Internet of Things (IoT), the management and collection of energy resources have begun to develop in the direction of intelligence [1]. Based on the re-optimization of wireless networks and power infrastructure, the efficiency of information exchange and resource exchange of power, transportation, and the IoT can be greatly improved. At present, the existing power communication network structure in China is complex and widely distributed, which brings great inconvenience to the use of individual users and the collection of power consumption information. Therefore, optimizing electromechanical equipment management strategy using power line communication is the key to solving the power problem.

Power line communication technology has the advantages of small investment, strong flexibility, and wide coverage. It is widely used in communication and data transmission between the power grid and the IoT [2]. However, due to the power grid's serious electromagnetic interference and channel attenuation, the communication quality is relatively poor. In order to solve the problems such as coverage and reliability of the power line communication network of the distribution grid, power line communication can be combined with the collection of electrical information from mechanical and electrical equipment. By designing the corresponding power line communication protocol, networking process, and route reconstruction strategy, the automatic networking and dynamic maintenance of the network can be realized, which is of great significance in improving the reliability of the power line communication network [3].

This paper makes an intelligent evaluation of the power line communication system based on relevant literature. Electromechanical equipment is used to optimize the relevant structure of the acquisition terminal.

^{*}State Grid Jibei Electric Power Supply Company Meterology Center, Beijing, 102208, China (Corresponding author, ChengfeiQi5@163.com)

[†]State Grid Jibei Electric Power Supply Company Meterology Center, Beijing, 102208, China (XiaoboYang3@126.com)

[‡]State Grid Jibei Electric Power Supply Company Meterology Center, Beijing, 102208, China (XiaokunYang3@163.com)

[§]State Grid Jibei Electric Power Supply Company Meterology Center, Beijing, 102208, China (ChaoranBi@126.com)

[¶]State Grid Jibei Electric Power Supply Company Meterology Center, Beijing, 102208, China (WenwenLi55@163.com)

According to the power grid intelligent sensing technology requirements, the combined network distribution strategy of power equipment and intelligent power acquisition terminal is optimized. Section 1 introduces the background of power line communication technology and the development of IoT networks. Section 2 sorts out the literature on power line communication technology and power acquisition terminals of mechanical and electrical equipment and finds the practical application scenarios of power consumption monitoring in large power system networks. Section 3 analyzes the characteristics of power line communication and wireless channel and uses information fusion to design a power load management system and data acquisition terminal. Section 4 compares and discusses the simulation results of the electricity acquisition terminal. Section 5 analyzes and discusses the experimental results and draws experimental conclusions. This paper has practical reference value for promoting the digital and intelligent transformation of electromechanical equipment in the power grid.

2. Recent related work.

2.1. Relevant research on power line communication technology. Oliveira et al. (2018) [4] studied the access control protocol of power line communication media. They studied the design of new requirements related to the media access control layer and physical network system through the analysis of the time-varying behavior of load in the power system and high-power impulse noise. The results show that the development of power line communication technology can solve the problem of network resource sharing by comparing the access control protocols of smart grids and multimedia devices. Ghasempour (2019) [5] studied the architecture and application of the IoT in the smart grid and built a dynamic global network based on Internet entities with network services. The results show that IoT devices have limitations in computing and storage. Therefore, it is necessary to design or use security solutions so that IoT devices can safely conduct power communication transmission. Matheus et al. (2019) [6] studied the concept, application, and challenges of visible light communication. They explored the security architecture of power line communication by improving mobile communication architecture and wireless service infrastructure. Based on the redesign of the IoT communication architecture, the results show that the proposed model can promote the optimization of wireless networks.

Kolade et al. (2020) [7] studied the indoor amplification, forwarding power line, and visible light communication channel model. With the help of software-defined radios, they obtain measurements in an indoor testbed from different locations between powerline communication transmitters and receivers. The results show that the error-free operation distribution and the error probability of the measurements in the proposed mixed channel are far lower than those in other models. Therefore, the effectiveness of the experimental results is verified. Yu et al. (2021) [8] studied the power and signaled composite modulation mode of power line data communication in the distributed grid. They are based on a single generator and load design of the converter, using a power spectral density approach to demodulate useful data for transmission lines. Hardware experiments verify the effectiveness of the proposed signal composite modulation strategy. Koshkouei et al. (2022) [9] evaluated the in situ power line communication system of lithium-ion batteries and combined it with the real-world dynamic drive configuration experimental files. The results show that an increase in the quality and quantity of battery characteristic data available at runtime can improve the accuracy of the system's assessment of Li-ion battery health. Fei et al. (2022) [10] used the missing value pattern and non-technical structure loss of neural structure search to conduct research and adopted a detection algorithm and supervised learning technology to process distribution network data. The results show that the analysis of the relationship between various power theft techniques and data loss can use the missing value location information to enhance the model's performance further.

In order to improve the effectiveness and reliability of power line communication, the existing research generally starts from two aspects improving the point-to-point communication quality of power line communication physical layer and data link layer and network layer performance. Therefore, in order to improve the communication speed and reliability of the power line communication network, the paper can optimize the power resource acquisition efficiency of the power acquisition terminal from the network layer.

2.2. Recent research on power acquisition terminal of electromechanical equipment. In the study of mechanical and electrical equipment and electrical acquisition terminals, Shi et al. (2018) [11] studied the data acquisition system of high-precision electrical impedance tomography. They used a high-precision digital synthesis method and digital demodulation technology with high immunity to eliminate random errors,

focusing on the main problems encountered in electrical impedance tomography data acquisition. The conclusion has practical application value for promoting the intelligent development of data acquisition systems. Zhang et al. (2020) [12] studied the electrical equipment identification system based on K-means clustering in intelligent buildings. They analyzed the load characteristics of the system and extracted the electrical characteristics for equipment identification from the collected data. The successful application of the system verifies the effectiveness of the proposed identification method. Budiman et al. (2021) [13] studied the monitoring and data acquisition information system. The experiment used Codeigniter for framework construction. System tests show that the design functionally generates 79.74% value, meets good standards, and is feasible. Herman et al. (2021) [14] studied the improved method of medical mask pressure for automatic data acquisition and analysis measurement. They created a simple pressure device with a three-dimensional (3D) printing model method. They used Python and MatLab scripts to acquire real-time pressure drop data and analyze multiple samples or batches. This paper has important reference value for improving the efficiency of medical data collection and processing.

In addition, in the research on intelligent manufacturing and the use of transmission lines in the power transmission process, Parveen et al. (2021) [15] explored the possibility of enhancing power transmission and multi-terminal power systems. They proposed to increase the use of existing transmission lines in addition to the independent control of AC alternating current/ direct current (DC) power flows. The results show that their proposed scheme can also meet the increase in electricity demand by increasing the use of existing transmission lines. Jiang et al. (2021) [16] studied the current compensation control strategy of electricity collection terminals for mechanical and electrical equipment in rectifier terminals. They proposed a current DC bus compensation control scheme and applied it to the rectifier terminal. In this scheme, the DC bus impedance gain at the rectifier end is reduced to balance the impedance at both ends and improve stability and controllability. Zhang et al. (2022) [17] used the energy-saving production architecture for the intelligent manufacturing process to describe the configuration of the data acquisition network. This approach enables the combination of social manufacturing and real-time energy profiling. Additionally, the energy consumption characteristics provide the decision-making basis for the energy-saving control of intelligent manufacturing workshops. Cao et al. (2022) [18] studied the wind farm data acquisition and management system based on edge computing. Firstly, the principles and advantages of edge computing technology are introduced. Then, they proposed to apply the technology to the wind farm data acquisition so that the data acquisition and management work could be carried out normally. Finally, the application of edge computing technology in wind farm data acquisition and management is simulated. The results show that the application of edge computing in wind farm data acquisition and management can improve the data transmission difficulties of electromechanical equipment during data acquisition.

According to the power grid smart sensing technology requirements, the current electromechanical equipment uses the power grid smart sensing terminal based on the power line and wireless communication integration technology. The designed terminal carries out power consumption monitoring and energy efficiency analysis in the large-scale power system network. Based on typical application scenarios such as power resource dispatching, power equipment, and intelligent power acquisition terminals have been widely promoted and applied.

3. Design of power load management system and data acquisition terminal based on power line communication.

3.1. Characteristic analysis of power line communication and wireless channel. At present, the topology of the power line communication network is complex and changeable. The network topology affects the transmission of communication channels in the current network, which changes with the distribution area [19]. Generally, this is a hybrid tree topology. Suppose the power communication network's distance between the source node and the target node is too large. In that case, reliable signal transmission between different nodes in the power network cannot be guaranteed. The power line communication network mainly includes a communication channel, an edge server, and a mobile base station. The structural framework of the power line communication network is shown in Figure 3.1.

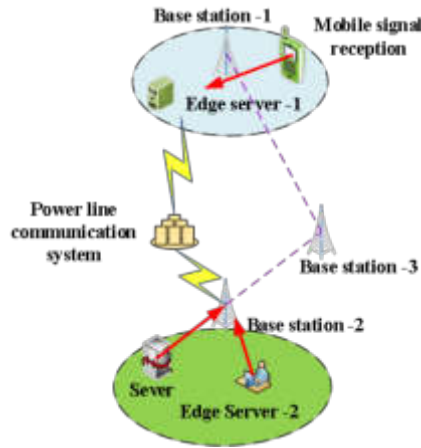


Fig. 3.1: Structural Framework of Power Line Communication Network

3.2. Management mode of power line and wireless communication equipment using information fusion. Enterprise, network, and physical layers are different levels of power line and wireless communication management structure. The existing research on power line communication combined with electromechanical equipment usually involves only one or two aspects, network or physical layers. The layer integration research on the power line and wireless network communication management model has not been fully launched. In order to improve the understanding of the complexity and reliability of network information, low-voltage line and microelectronic radio communication are deeply integrated by using multi-level and combined network and distribution strategies [20]. This is achieved by combining the concept of unified communication protocol, adopting the strategy of reasonable allocation of power resources, and understanding the network information of communication objects. For the management mode of wireless communication equipment, the structure is shown in Figure 3.2:

3.3. Structural design and research of power acquisition terminal model. Technically, the power acquisition terminal model is a centralized, spatially evenly distributed real-time computer control system. One of the technical bases of centralized monitoring of power load is the data communication between computers. A communication disk is established between the main system of the operating company and the locally distributed user data terminals to realize the centralized equipment management during the data exchange between the headquarters and the end users. Strict standards and practices must be followed; the system must receive data correctly before the timer locks. If the output power consumption data is inaccurate or lost, the timer will specify a timeout, and the system administrator will resend the power consumption data. By transmitting communication protocols between the application, data link, and physical layers, the structure of the power acquisition terminal model can be optimized. The structure of the power acquisition terminal model in the system network is shown in Figure 3.3.

3.4. Design of simulation experiment. The electromechanical power line communication equipment includes a timer and an external interrupt service program. Based on the timer interrupt service program, the data type conversion needs to be started when the timer stops in the model. A fixed number of network nodes in the weekly signal wave can be uniformly sampled. The power acquisition terminal module adopts a special power measurement chip. The power consumption can be measured by capturing the voltage and current of the integrated three-phase voltage sensor and external current sensor. The parameters of power, power supply, reactive power, and reactive voltage factor can be calculated. In the dynamic input range, the nonlinear measurement error of the chip is less than 0.1%.

In order to facilitate the verification of the communication fusion method proposed, the simulation experiment is based on the following assumption. Each node can calculate the communication rate, delay, and bit

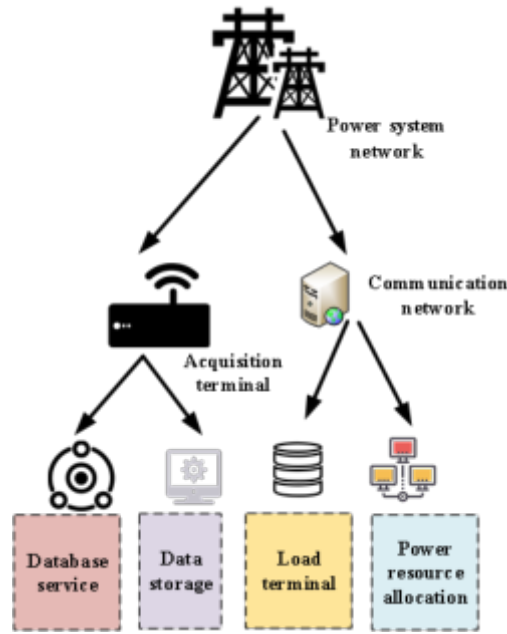


Fig. 3.2: Management Mode Structure of Wireless Communication Equipment

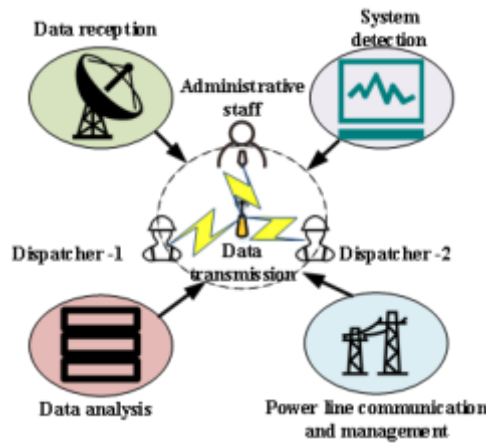


Fig. 3.3: Network structure of power acquisition terminal model

error rate between it and the previous node after receiving the data packet and storing it in the corresponding routing table. In order to research the power acquisition terminal of electromechanical equipment, Visual C++ is used as the experimental simulation environment. The modular design method is adopted, which is divided into power information acquisition, Bluetooth wireless communication, and system debugging modules. The main function is to display and verify the AC sampling value and detect other power data. The microprocessor system realizes the communication with the wireless Bluetooth HC-05 debugging communication module or serial port through the RS-232 serial interface, mainly realizing the parameter setting of the intelligent perception terminal. In addition, in order to compare the performance of power acquisition terminals, the system performance is tested under different current values, voltage values, and load impedance, and the results are analyzed and discussed. The scene structure of the simulation experiment application is designed, and the

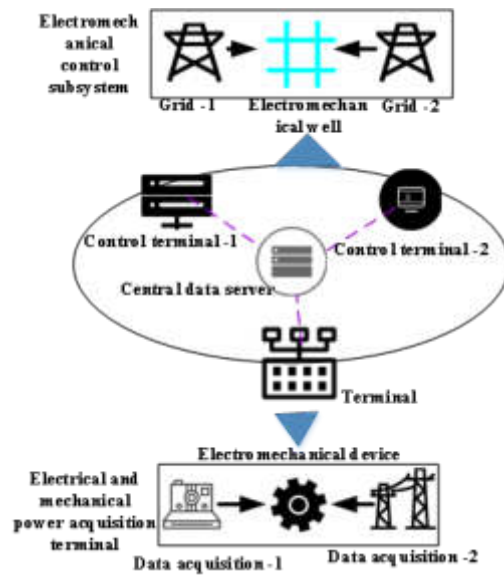


Fig. 3.4: Scene structure of simulation experiment application

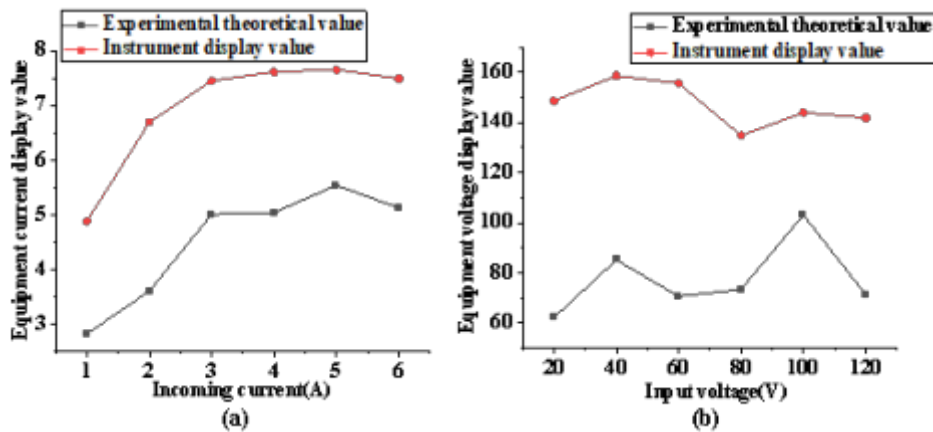


Fig. 4.1: Current and voltage value change curve of electric power acquisition terminal under different input values (a. comparison of terminal output current values under different current input values of the system; b. comparison of terminal output voltage values under different voltage input values of the system)

structure is shown in Figure 3.4.

4. Results and Discussion.

4.1. Comparison between current and voltage value of power acquisition terminal. In order to test the performance of electromechanical equipment combined with the power line communication proposed, it is necessary to compare the current and voltage value of the power acquisition terminal under different input values. Compare the results of the current and voltage values of the power acquisition terminal under different inputs, which are shown in Figure 4.1. In addition, in order to analyze the instrument power change of the power acquisition terminal, the instrument power change data of the power acquisition terminal under different current and voltage values are sorted out. The results are shown in Figure 4.2.

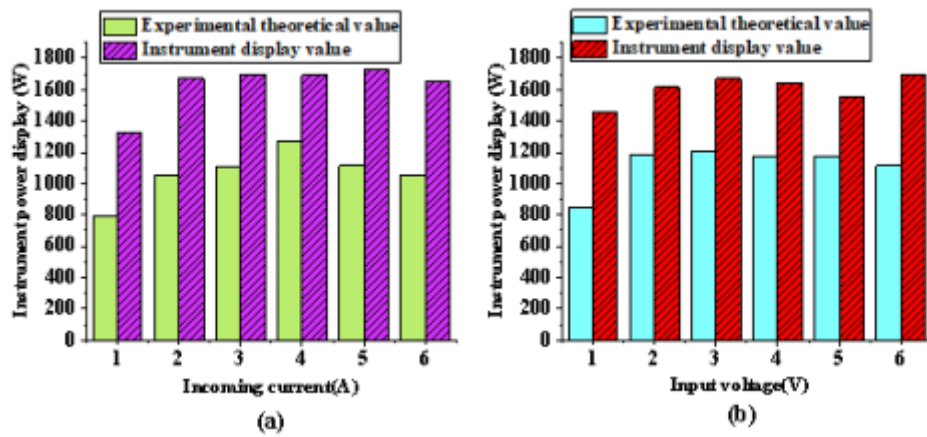


Fig. 4.2: Instrument power change curve of power acquisition terminal under different current and voltage values (a. comparison of instrument power change results of power acquisition terminal under different current input values of the system; b. comparison of instrument power change results of power acquisition terminal under different voltage input values of the system)

In Figure 4.1, when the input current value of the system is increasing, the data output results of the experimental theoretical and the instrument display value show an increasing trend. When the input current is 1A, the experimental, theoretical value of the system is 3A, and the actual instrument output is close to 5A. When the input current is 6A, the instrument output of the system is 7.5A. In addition, when the system input voltage is 100V, the theoretical output value is 110V, and the actual instrument output value is 140V. Therefore, the power acquisition terminal designed by the research can reasonably dispatch and allocate power resources.

In Figure 4.2, the theoretical output power value of the acquisition terminal and the actual instrument output value will fluctuate under different current and voltage values. When the system input current is 1A, the theoretical output value of the system is 800W, and the actual output value is 1300W. In addition, when the system input voltage is 1V, the theoretical power output value of the terminal is 850W, and the actual instrument output value is 1480W. The influence of the voltage input value on the system output power is analyzed. When the voltage input is 6V, the theoretical power output value of the terminal is 1100W, and the actual instrument output value is 1680W. Therefore, the results show that the system's output power can be effectively adjusted with the help of experimental instruments.

4.2. System performance analysis under different load impedance. In order to further analyze the system performance of the electrical acquisition terminal, the gap between the experimental theoretical value of the system and the instrument display value is compared under different load impedances. The data change curve is shown in Figure 4.3. The change curve of output power and output voltage value of the system under different input power is shown in Figure 4.4. In addition, the data receiving accuracy and data detection error rate change data of the power acquisition terminal are sorted out. The data results are shown in Figure 4.5.

In Figure 4.3, the difference between the experimental theoretical and the instrument display value under different load impedances will cause certain fluctuations. When the load impedance of the system is 20, the theoretical current of the instrument of the electric data acquisition terminal is 0.58, and the current value of the terminal data acquisition instrument is displayed as 0.98. In addition, when the load impedance of the system is 180, the current value of the terminal data acquisition instrument is displayed as 0.44. When the capacitance is removed, the load impedance range of the system changes from 100 to 600. When the load impedance is 200, the theoretical current of the instrument at the electric data acquisition terminal is 2.5, and the actual current of the terminal data acquisition instrument is 4.25. Therefore, the reasonable load impedance value loaded in the system can optimize the model's current output value and improve the resource utilization rate.

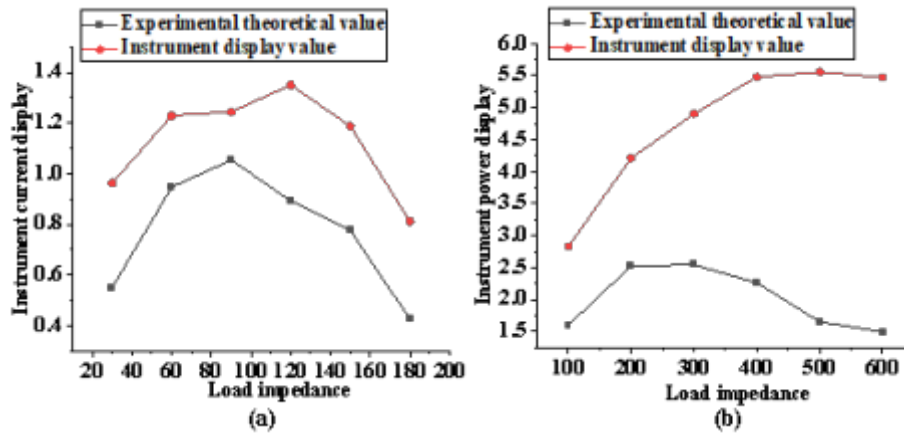


Fig. 4.3: Comparison of the gap between the theoretical, experimental value and the instrument display value under different load impedances of the system (a. the curve of the data changes between the theoretical, experimental value, and the instrument display value under different load impedances of the system after adding capacitance; b. the curve of the change of the data between the theoretical, experimental value, and the instrument display value under different load impedances of the system after removing capacitance)

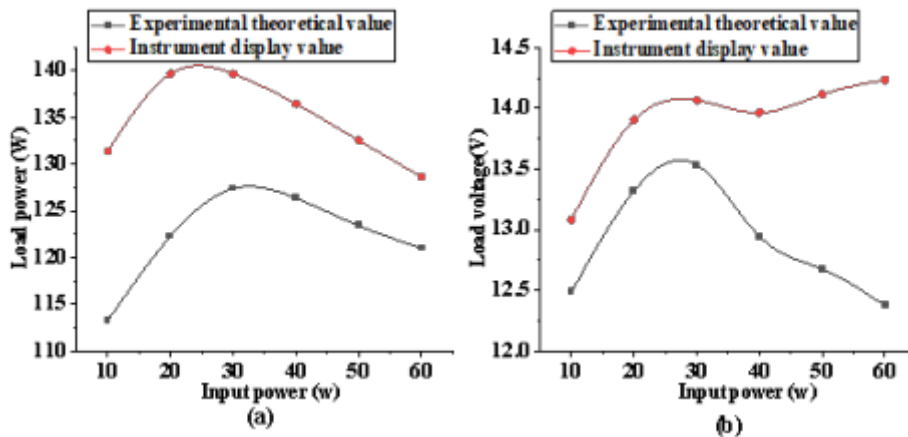


Fig. 4.4: Comparison of the system's output power and output voltage under different input power of the terminal (a. the change curve of the output power value of the system under different input power of the terminal; b. the change curve of the output voltage value of the system under different input power of the terminal)

In Figure 4.4, under the different input power of the electricity acquisition terminal, the system's output power and voltage value vary greatly. When the terminal input power is 10W, the theoretical output power of the system is 112.5W, and the actual output value is 133W. Currently, the theoretical output voltage of the system is 12.5V, and the actual model output voltage is 13.2V. When the terminal input power is 60W, the output power of the actual model is 130W, and the output voltage of the actual model is 14.25V. Therefore, reasonable allocation of grid resources can improve the working efficiency of power acquisition terminals and promote the rapid transformation of power resources.

In Figure 4.5, under the increasing input of the power acquisition terminal, the system's data-receiving accuracy is constantly improving, and the error results of data detection are constantly decreasing. When the

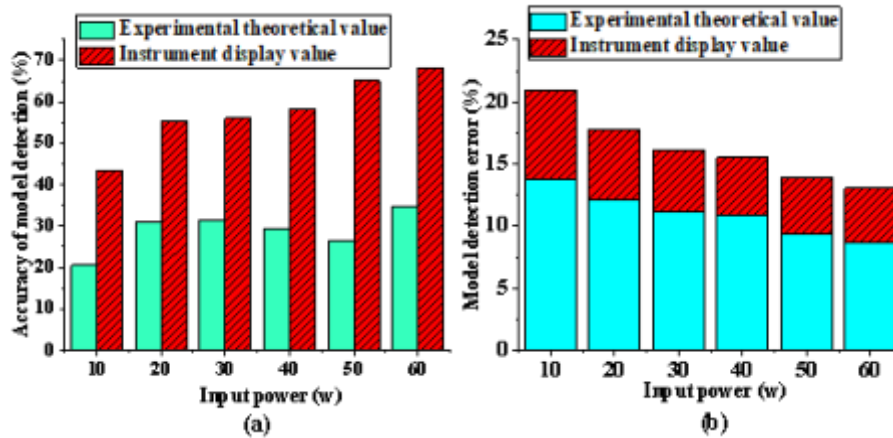


Fig. 4.5: Comparison of data receiving accuracy and data detection error of electric power acquisition terminal (a. change curve of data receiving accuracy of electric power acquisition terminal; b. change curve of data detection error of electric power acquisition terminal)

input power of the system is 10W, the accuracy of the experimental, theoretical output data is 20.5%, and the accuracy of the actual output data is 43.3%. At this time, the error rate of theoretical detection data is 13.7%, and the detection error rate of the actual data acquisition terminal can be reduced to 7.2%. In addition, when the system input power is 60W, the theoretical error detection rate of the data acquisition terminal of the system is 8.7%. The error detection rate of the actual terminal can be reduced to 4.3%, which is much lower than the previous data detection error rate. Therefore, using power line communication technology to optimize the power acquisition terminal of electromechanical equipment is conducive to improving the accuracy of power resource data transmission.

5. Conclusion. With the progress of the times and the rapid development of power supply technology, users have put forward higher requirements for power load resource management and power resource transmission. The paper adopts power line communication technology to analyze the transmission efficiency of resources in the current network. Based on the optimization of network topology, the main and fixed relay communication modes are selected to transmit power resources in the network. The electromechanical equipment type of power line communication is studied with the form of information fusion and structure optimization of power acquisition terminal model. The results show that the difference between the experimental theoretical and the instrument display value under different load impedances will cause certain fluctuations. When the load impedance of the system is 20, the theoretical current of the instrument of the electric data acquisition terminal is 0.58, and the current value of the terminal data acquisition instrument is 0.98. In addition, when the load impedance is 200, the theoretical current of the instrument of the electrical acquisition terminal is 2.5. The actual current of the terminal data acquisition instrument is 4.25. This paper has practical application value for promoting the deep integration of electromechanical equipment and power line communication technology. However, there are some deficiencies in the research. The main disadvantage is that in the actual power system network, the actual distribution network impedance is affected by many factors. Here, only a specific length of the power cable is simulated. In future research, more powerful system network data should be combined to optimize the model management strategy.

REFERENCES

- [1] Ma, Z., Xiao, M., Xiao, Y., Pang, Z., Poor, H. V., & Vucetic, B. (2019). High-reliability and low-latency wireless communication for internet of things: Challenges, fundamentals, and enabling technologies. *IEEE Internet of Things Journal*, 6(5), 7946-7970.

- [2] Zhang, H., Li, R., & Shi, C. (2022). Deep learning technology of Internet of Things Blockchain in distribution network faults. *Journal of Intelligent Systems*, 31(1), 965-978.
- [3] Lin, H., Xu, X., Zhao, J., & Wang, X. (2020). Dynamic service migration in ultra-dense multi-access edge computing network for high-mobility scenarios. *EURASIP Journal on Wireless Communications and Networking*, 2020(1), 1-18.
- [4] Oliveira, R. M., Vieira, A. B., Latchman, H. A., & Ribeiro, M. V. (2018). Medium access control protocols for power line communication: A survey. *IEEE Communications Surveys & Tutorials*, 21(1), 920-939.
- [5] Ghasempour, A. (2019). Internet of things in smart grid: Architecture, applications, services, key technologies, and challenges. *Inventions*, 4(1), 22.
- [6] Matheus, L. E. M., Vieira, A. B., Vieira, L. F., Vieira, M. A., & Gnawali, O. (2019). Visible light communication: concepts, applications and challenges. *IEEE Communications Surveys & Tutorials*, 21(4), 3204-3237.
- [7] Kolade, O., Familua, A. D., & Cheng, L. (2020). Indoor amplify-and-forward powerline and visible light communication channel model based on a semi-hidden Markov model. *AEU-International Journal of Electronics and Communications*, 124, 153108.
- [8] Yu, D., Li, K., Yu, S., Trinh, H., Zhang, P., Oo, A. M., & Hu, Y. (2021). A novel power and signal composite modulation approach to powerline data communication for SRM in distributed power grids. *IEEE Transactions on Power Electronics*, 36(9), 10436-10446.
- [9] Koshkouei, M. J., Kampert, E., Moore, A. D., & Higgins, M. D. (2022). Evaluation of an in situ QAM-based Power Line Communication system for lithium-ion batteries. *IET Electrical Systems in Transportation*, 12(1), 15-25.
- [10] Fei, K., Li, Q., & Zhu, C. (2022). Non-technical losses detection using missing values' pattern and neural architecture search. *International Journal of Electrical Power & Energy Systems*, 134, 107410.
- [11] Shi, X., Li, W., You, F., Huo, X., Xu, C., Ji, Z., Dong, X. (2018). High-precision electrical impedance tomography data acquisition system for brain imaging. *IEEE Sensors Journal*, 18(14), 5974-5984.
- [12] Zhang, G., Li, Y., & Deng, X. (2020). K-means clustering-based electrical equipment identification for smart building application. *Information*, 11(1), 27.
- [13] Budiman, A., Sunariyo, S., & Jupriyadi, J. (2021). Sistem Informasi Monitoring dan Pemeliharaan Penggunaan SCADA (Supervisory Control and Data Acquisition). *Jurnal Tekno Kompak*, 15(2), 168-179.
- [14] Herman, A., Porter, D., Rottach, D., & Guha, S. (2021). A Modified Method for Measuring Pressure Drop in Non-medical Face Masks with Automated Data Acquisition and Analysis. *Journal of the International Society for Respiratory Protection*, 38(2), 42-55.
- [15] Parveen, S., Hameed, S., Rahman, H., Rahman, K., Tariq, M., Alamri, B., & Ahmad, A. (2021). The Possibility of Enhanced Power Transfer in a Multi-Terminal Power System through Simultaneous AC-DC Power Transmission. *Electronics*, 11(1), 108.
- [16] Jiang, Y., Tian, Y., Li, Y., & Wang, F. (2021). DC-side current compensation control in the rectifier terminal for power variations in back-to-back converters. *IET Renewable Power Generation*, 15(13), 3025-3037.
- [17] Zhang, C., Zhang, J., Ji, W., & Peng, W. (2022). Data Acquisition Network Configuration and Real-Time Energy Consumption Characteristic Analysis in Intelligent Workshops for Social Manufacturing. *Machines*, 10(10), 923.
- [18] Cao, X., Xu, Y., Wu, Z., Qin, X., & Ye, F. (2022). Data acquisition and management of wind farm using edge computing. *International Journal of Grid and Utility Computing*, 13(2-3), 249-255.
- [19] Tumash, L., Olmi, S., & Schöll, E. (2019). Stability and control of power grids with diluted network topology. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(12), 123105.
- [20] Donders, K. (2019). Public service media beyond the digital hype: distribution strategies in a platform era. *Media, Culture & Society*, 41(7), 1011-1028.

Edited by: B. Nagaraj M.E

Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing

Received: Dec 28, 2023

Accepted: Mar 20, 2024



A FAULT MONITORING SYSTEM FOR MECHANICAL AND ELECTRICAL EQUIPMENT OF SUBWAY VEHICLES BASED ON BIG DATA ALGORITHMS

GENG LI^{*}, YA LI[†] AND HONGXUE BI[‡]

Abstract. This paper uses big data technology to extract the electromechanical fault characteristics of metro vehicles and analyze the current situation under different fault conditions to ensure the operation quality and safety of metro operations. It also establishes a simulation model to simulate the current waveform of metro vehicles under different fault conditions and analyze the fault phenomenon. The simulation test results show that: (1) The current waveform of a single transistor with the hard fault is compared with the simulated current waveform under a normal state. The upper part of the A phase current waveform is lost when T1 fails. When T2 fails, the current waveform in the lower half of the C phase current is lost. When T3 fails, the upper half of the B phase current waveform is lost. (2) The current waveform in the hard fault's upper and lower bridge arms will have phase loss. In the T25 fault, the C phase current is completely lost. In the T14 fault, the phase A current waveform is completely lost. In the T36 fault, the phase B current is completely lost. (3) The current waveform of a single transistor with a soft fault is complete, but the overall current amplitude is reduced. When a T1 fails, the A phase current tends to rise first and then fall. Compared to normal, the amplitude of the current decreases, and the peak decreases slightly. (4) The current values of phase B and phase C of the two transistors on the same bridge above and below the soft fault are mostly the same. The phase A current output value decreases in both the positive and negative half cycles. This paper aims to improve the monitoring ability of the monitoring system of electromechanical equipment of metro vehicles, which plays a specific role in maintaining the safety of subway operations and improving the quality of subway operations.

Key words: Big data technology, metro vehicles, electromechanical equipment, monitoring systems, simulation

1. Introduction. With the development of urbanization, the role of urban rail transit in cities is becoming increasingly important. The metro is the backbone of public transportation in modern large cities and the backbone of passenger transportation within the city [1,2]. Metro vehicle Electromechanical Equipment (EE) is a general term for electrical and mechanical equipment that converts electricity and other energy. It is indispensable equipment for maintaining metro operations. The failure of electrical equipment in a metro train will directly affect the operational safety of the city's rail vehicles. Therefore, fault monitoring of electrical equipment in metro trains is critical. The converter is the core component of controlling multiple components in the metro electrical equipment. Monitoring the condition of the converter can prevent possible electromechanical failures during metro operation.

At the end of the last century, big data technology began to develop. Big data has a wide range of applications, for example, in healthcare. Analyzing big data through machine learning can help evaluate large amounts of complex healthcare data to improve medical diagnosis and disease classification [3]. During the 13th Five-Year Plan period, big data technology was widely used in urban rail transit. Through big data analysis, South Korea could review the operation of express trains and detect the trend of train traffic [4]. Big data analysis can plan metro travel routes [5]. Advanced big data technology can detect defects in metro tunnels [6]. In terms of metro operation, based on big data technology, an automatic toll collection system with stable system operation, relatively small memory, fast response speed, and low delay can be designed to improve the convenience of residents' lives and accelerate the process of urban construction [7]. Using big data technology can realize the intelligent management of metro operations, which plays a vital role in alleviating traffic congestion, reducing traffic accidents, and reducing energy consumption [8,9]. In addition, in terms of electromechanical fault detection, the detection system can be highly adaptable based on big data technology.

^{*}Zhengzhou CSCEC Shenzhen Railway Rail Transit Co., Ltd, Zhengzhou, 450000, China (Corresponding author, GengLi2@126.com)

[†]Zhengzhou CSCEC Shenzhen Railway Rail Transit Co., Ltd, Zhengzhou, 450000, China (YaLi381@163.com)

[‡]Zhengzhou Railway Vocational & Technical College, Zhengzhou, 450000, China (HongxueBi@126.com)



Fig. 2.1: Characteristics of big data technology

It can predict the early electromechanical failure of the aircraft and improve the fault diagnosis performance of the electromechanical system [10,11]. When dealing with complex classification problems, the use of big data technology can realize real-time fault detection of key components of the electromechanical traction system of high-speed trains, which is of great significance for improving the reliability of train motors and reducing the cost of support [12,13]. Electromechanical faults can be monitored using big data technology to establish simulation models.

Firstly, this paper uses big data technology to extract the electromechanical fault characteristics of metro vehicles and analyze the current situation under different fault conditions to ensure the operation quality and safety of metro operation and improve the monitoring of metro EE. Secondly, a simulation model is established to simulate the current waveform of metro vehicles under different fault conditions and analyze the fault phenomenon. The innovation point is to combine big data technology with the electromechanical fault monitoring system of metro vehicles and analyze the current waveform of the converter under different faults through simulation models. The monitoring of EE of metro vehicles has been improved, which has played a specific role in maintaining the safety of metro operations.

2. Establishment of the electromechanical fault simulation model for metro vehicles.

2.1. Big Data Technology. Big data technology refers to the application technology of big data, including various big data platforms, big data index systems, and big data application technology. As one of the leading development directions in the information field, big data technology can be applied to data mining, data analysis, and data sharing in massive data. It leverages the potential value of data to create substantial economic benefits. Big data can optimize resource utilization while making informed decisions [14,15]. In the oil and gas industry, big data technology can be used to analyze seismic and microseismic data, reduce drilling time and improve drilling safety, optimize the performance of production pumps, improve asset management, and improve transportation safety [16,17]. Big data technology can promote students' interest in learning and improve their concentration on learning in college education. Big data management in big data technology can organize and analyze mining data well. Big data technology is one of the promising technologies that could reshape the entire mining landscape.

The characteristics of big data technology are shown in Figure 2.1.

Big data technology has many data processing methods, such as Bayesian networks, random forests, decision trees, Principal Component Analysis (PCA), Gaussian mixture models, and regression analysis models. Here, the PCA and the Weibull distribution method are mainly used for data preprocessing.

PCA is a statistical method that can comprehensively analyze the problem, save calculation time and cost by reducing the dimensionality of the data set, and improve the accuracy of the calculation. As a basic mathematical analysis method, PCA has applications in the fields of demographics, quantitative geography, mathematical modeling, and mathematical analysis. It is a commonly used multivariate analysis method. PCA has three main functions: reducing the dimensionality of the data space, analyzing the relationship between variables, and using graphics to represent multidimensional data.

Suppose $X_1, X_2, \dots, X_i (i = 1, 2, \dots, j)$ is the eigenvector corresponding to the eigenvalues of the covariance matrix of M . $M_1, M_2, \dots, M_i (i = 1, 2, \dots, j)$ is the normalized value of the original variable. Then, the following relationship exists when normalizing.

$$F = X_1 * M_1 + X_2 * M_2 + \dots + X_i * M_i \quad (2.1)$$

The Weibull distribution method is the theoretical basis of reliability analysis. It is suitable for inferring the wear of electromechanical products, industrial manufacturing, predicting the weather, predicting technological changes, and modeling the received clutter signal. The Weibull distribution method was proposed in 1927 and explained in detail by the Swedish engineer and mathematician Waloddi Weibull in 1951. According to the form, it can be divided into three types: the one-parameter Weibull distribution, the two-parameter Weibull distribution, and the three-parameter Weibull distribution or mixed Weibull distribution.

The Weibull distribution is a continuous probability distribution. Suppose x is a random variable, $y > 0$ is the scale parameter, and $z > 0$ is the shape parameter. When $z=1$, the probability density function is an exponential distribution; when $z=2$, it is a Rayleigh distribution. The probability density is calculated as follows.

$$f(x, y, z) = \begin{cases} \frac{z}{y} \left(\frac{x}{y}\right)^{z-1}, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad (2.2)$$

γ is the gamma function, and the mean is calculated according to Equation 2.3.

$$E = y\gamma\left(1 + \frac{1}{z}\right) \quad (2.3)$$

If the time is t , m is the standard parameter, and s is the positional parameter. Then, the continuous distribution function $f(x)$ is:

$$f(t) = \frac{z}{m} \left(\frac{1-s}{z}\right)^{z-1} E^{-\left(\frac{t-s}{m}\right)^z} \quad (2.4)$$

2.2. Metro EE. Metro EE is mainly divided into ventilation and air conditioning systems, water supply and drainage systems, power lighting systems, elevator and screen door systems, automatic ticket vending systems, power supply systems, communication signals and other weak current systems, civil air defense projects, and subway vehicles [18,19]. Metro trains mainly use automation equipment with electronic computer processing technology as the core. The role of this equipment is to replace manual labor with mechanized and electrified systems to ensure the operation and safety of trains. Metro trains mainly include car bodies, power bogies and non-power bogies, traction buffer connection devices, brake devices, current receiver devices, internal vehicle equipment, and electrical systems. Among them, the electrical system is divided into the main, auxiliary, and electronic and control circuits. Converters are auxiliary circuit systems.

The characteristics of EE failures are the characteristics of the occurrence, development, and change of faults in a complete life cycle of EE from use to no longer use. Figure 2.2 shows the classification of fault characteristics of EE.

The auxiliary power supply system of the metro train receives the current of the high-voltage power supply network. These currents are transmitted by a pantograph or a third rail catenary. The controller controls the inverter circuit. The voltage on the catenary of the auxiliary inverter is changed to a medium-pressure flow.

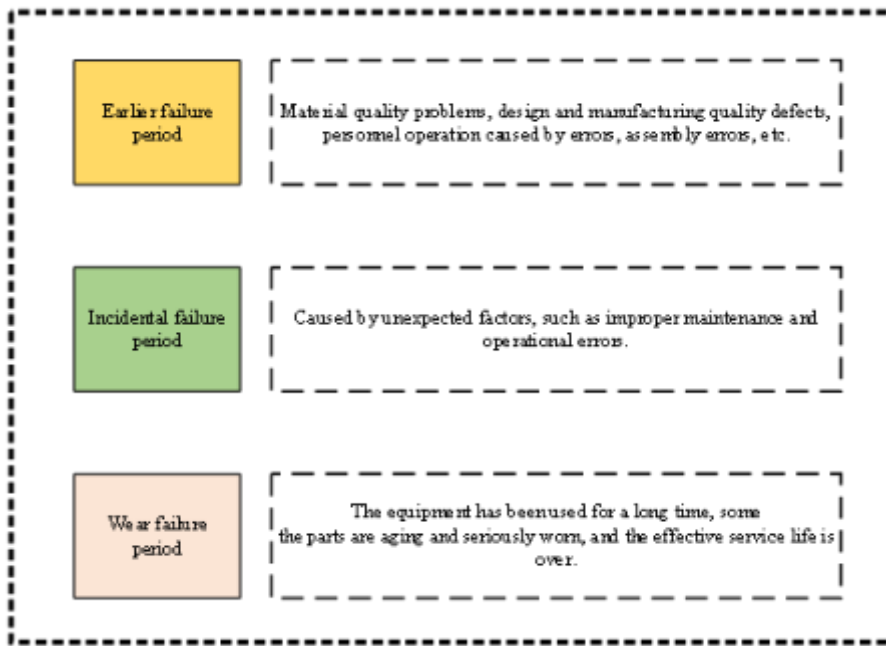


Fig. 2.2: Classification of electromechanical fault characteristics

The auxiliary power supply system is divided into the grid-connected, cross-power, and extended power supply. A bus bar connects the grid-connected power supply to the whole vehicle. There is no contactor in the middle, and the two auxiliary inverters work together. The cross-power supply is the direct power supply from the busbar to the trunks that have been grouped. Under normal circumstances, the vehicle alternating current load is divided into two groups on average according to the equipment's power. Two auxiliary inverters supply power to each of the two sets of equipment. The extended power supply is a more complex way of supplying power to two independent units using a bus bar. There is a contactor between the two auxiliary inverters. When the contactor is turned off, the malfunctioning auxiliary inverter stops working, and another auxiliary inverter is responsible for the entire vehicle.

Among them, the power supply switching mode of cross-power and grid-connected power supply is the simplest, and no additional hardware is required. When switching the expansion circuit, it is necessary to set up the expansion contactor, and the auxiliary inverter must also stop working.

The specific power supply system load is presented in Figure 2.3.

The fault characteristics extraction of the metro electrical system can be carried out through three aspects: time domain, frequency domain, and time-frequency domain. The time domain mainly describes the relationship of mathematical functions or physical signals to time. Time is the independent variable, and signal change is the dependent variable. Time domain analysis uses the time axis as the coordinate to represent dynamic signal changes. The frequency domain is a coordinate system used to describe the frequency characteristics of a signal. The independent variable is the frequency, and the dependent variable is the change amplitude of the frequency signal. It is also a spectrogram, which describes the frequency structure of the signal and the relationship between the frequency and the amplitude of the frequency signal, mainly used in electronics, control system engineering, and other aspects. The time-frequency domain combines time and frequency to express information about frequency over time. The time domain process is straightforward, and the frequency domain process needs to be transformed. The time-frequency domain process is more complex, but the resulting signal quality is higher. Therefore, the fault characteristics of the metro electrical system are extracted from the perspectives of time-frequency and time domains.

Time-domain fault feature extraction is to count the different feature manifestations of signals in different

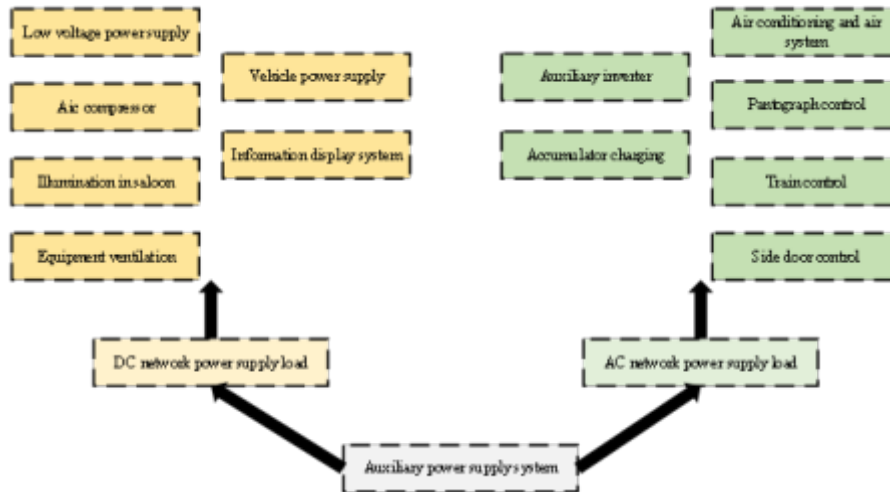


Fig. 2.3: Auxiliary power supply system load

states in the time domain. There are two main indicators of time-domain signals: dimensioned and dimensionless parameters. Dimensioned parameters have unit values and are easily affected by environmental interference. Dimensionless parameters are values without units and are not affected by interference factors.

Dimensioned parameters are calculated according to Equation 2.5-Equation 2.7.

$$A = \int_{-\infty}^{\infty} AP(A)dx \tag{2.5}$$

$$\alpha = \int_{-\infty}^{\infty} A^3 P(A)dx \tag{2.6}$$

$$\beta = \int_{-\infty}^{\infty} A^4 P(A)dx \tag{2.7}$$

In Equation 2.5-2.7, P(A) is the probability density function, A is the mean of any sample, α is the degree of skewness, and β is the steepness.

Dimensionless parameters are calculated according to Equation 2.8-Equation 2.10.

$$X = \frac{A_{RMS}}{|\bar{A}|} \tag{2.8}$$

$$L = \frac{A_{MAX}}{A_{RMS}} \tag{2.9}$$

$$K = \frac{e[(A - U)^4]}{\sigma^4} \tag{2.10}$$

In Equation 2.8-Equation 2.10, U is the mean, U^4 is the fourth-order center distance of the random variable A, e is the expectation, A_{RMS} is the square root of the mean, σ is the standard deviation, X is the waveform index, L is the peak indicator, and K is the steepness factor.

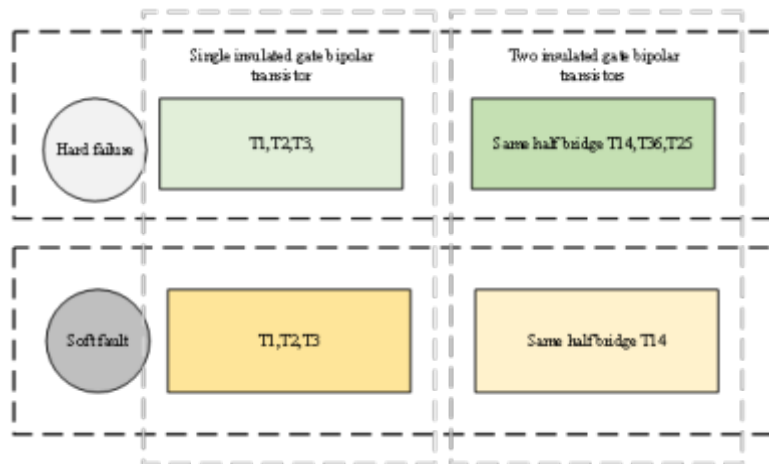


Fig. 2.4: Electrical system fault classification

Time-frequency domain fault feature extraction uses Short-time Fourier Transform (STFT), S transform, Empirical Mode Decomposition (EMD), and other methods to extract signals.

The STFT refers to a mathematical transformation related to the Fourier transform to determine the frequency and phase of a sine wave in the local area of a time-varying signal. The uncertainty criterion limits the STFT window function, and the area of the time-frequency window is not less than two. Therefore, the time and frequency resolution of the STFT window function cannot be optimized simultaneously.

The S transform replaces wavelet basis functions with Gaussian windows. It is also known as the "phase orthogonal" continuous wavelet transform. The signal is divided into many small time intervals, and each time interval is analyzed with the Fourier transform to determine the frequency at which the time interval exists. The feature quantity extracted by the S transform is not sensitive to noise.

EMD is a novel adaptive signal processing method for nonlinear and nonstationary signals. It decomposes the signal based on the timescale characteristics of the data itself. It does not require pre-setting. Compared with the previous two methods, the data of EMD is more intuitive.

Figure 2.4 demonstrates the fault classification of the metro electrical system.

2.3. Simulation model establishment. Matrix Laboratory (MATLAB) simulation tool is a commercial mathematical software produced by MathWorks in the United States. It is mainly used in data analysis, deep learning, control system design and simulation, data image processing, and data signal processing. MATLAB, along with Mathematica and Maple, is known as the three major mathematical software. MATLAB integrates matrix calculations, numerical analysis, modeling and simulation of nonlinear dynamic systems, and scientific data visualization in an easy-to-use windowed environment, largely freeing itself from the editing mode of traditional non-interactive programming languages.

The advantage of MATLAB is its efficient numerical and symbolic computing capabilities. It can handle complex and tedious mathematical operations. It has complete graphics processing functions to visualize calculation results. MATLAB also has a feature-rich toolkit that simplifies data processing.

MATLAB has five major system structures: development environment, mathematical library, language, graphics processing system, and application programming interface. The specific system structure is shown in Figure 2.5.

With the help of the Simulink module in the MATLAB tool, simulation models of auxiliary inverter circuits can be built. The simulation model is divided into two inverter modules: the inverter control module and the inverter output module.

Pulse-width modulation control is implemented in the inverter control module using the pulse-width modulation generator in the module library. The input of the pulse-width modulation generator is set with a sine

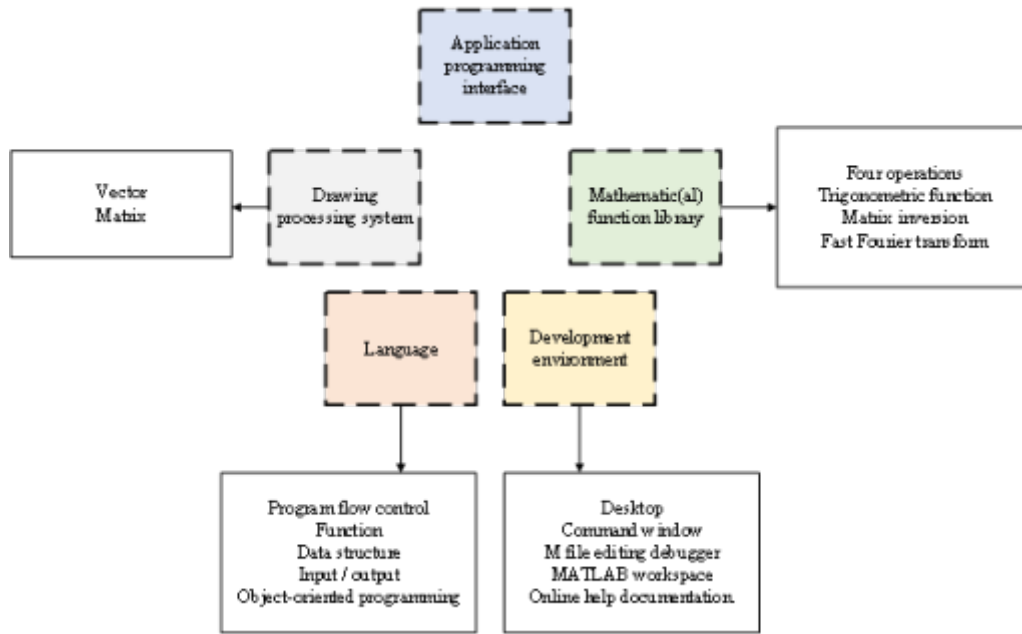


Fig. 2.5: MATLAB system structure

wave with a difference of 120° to control the Insulated Gate Bipolar Transistor (IGBT). The generator output is connected to the gate. It is necessary to prevent the reduction of harmonic content and power life problems of the inverter simulation output voltage caused by the excessive switching frequency of the components in the actual experiment. Therefore, the carrier frequency in the pulse generator is set to 0.6, the modulation wave frequency is set to 50, and the generation period mode is selected as the 3-arm bridge of six pulses.

In the inverter output module, the main circuit is set as a three-phase bridge inverter circuit. IGBT modules include gates, collectors, generators, module test terminals, and module monitoring terminals. The internal resistor size is set to 0.02Ω , and every six IGBT modules form a complete three-phase bridge inverter circuit.

A filter circuit that combines inductors and capacitors is built using modules in the component library. The input voltages of interfaces one, two, and three at the left end are set to the output voltage of the inverter circuit, and the output voltages of four, five, and six after the interface at the right end are set to sine waveforms.

In the simulation model, the grid voltage output of the auxiliary inverter system is set to 1,500V. The direct current voltage changes from 300V to 600V through the battery circuit and resonant converter. In the simulation experiment, the maximum voltage is selected as the input voltage value, set to a purely resistive load. The resistor size is 0.6, the total simulation time is 0.1 seconds, and waveforms are recorded every 0.01 seconds. The output of the simulation circuit is given in Figure 2.6.

3. Simulation results of electromechanical faults of metro vehicles.

3.1. Hard fault simulation results. The variation in the current waveform of a single insulated-gate bipolar transistor is shown in Figure 3.1.

From Figure 3.1, when the power tube T1 fails, the current fluctuation range of phase A is between 0 and -400A, the upper part is lost, and the changing trend will repeat every 0.02s. Phase B current fluctuates from 300 to -300A. The current change showed a trend of increasing first and then decreasing, repeating every 0.02s and five times within 0.1s. The C phase current range and repetition are consistent with the B phase current, but the trend is first falling and then rising. When the power tube T2 fails, the current change trend of phase A is first increasing and then decreasing, and it repeats every 0.02s. The current varies from 300 to -300A. The current change trend of phase B is first falling and then rising, which coincides with the phase A current about every 0.01s. The current varies from 300 to -400A. The C-phase current varies from 400 to 0A. The

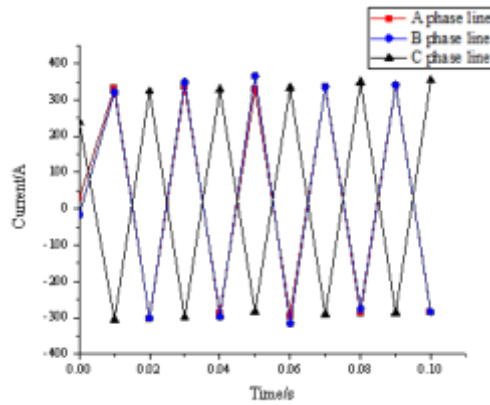


Fig. 2.6: Output of simulation circuit

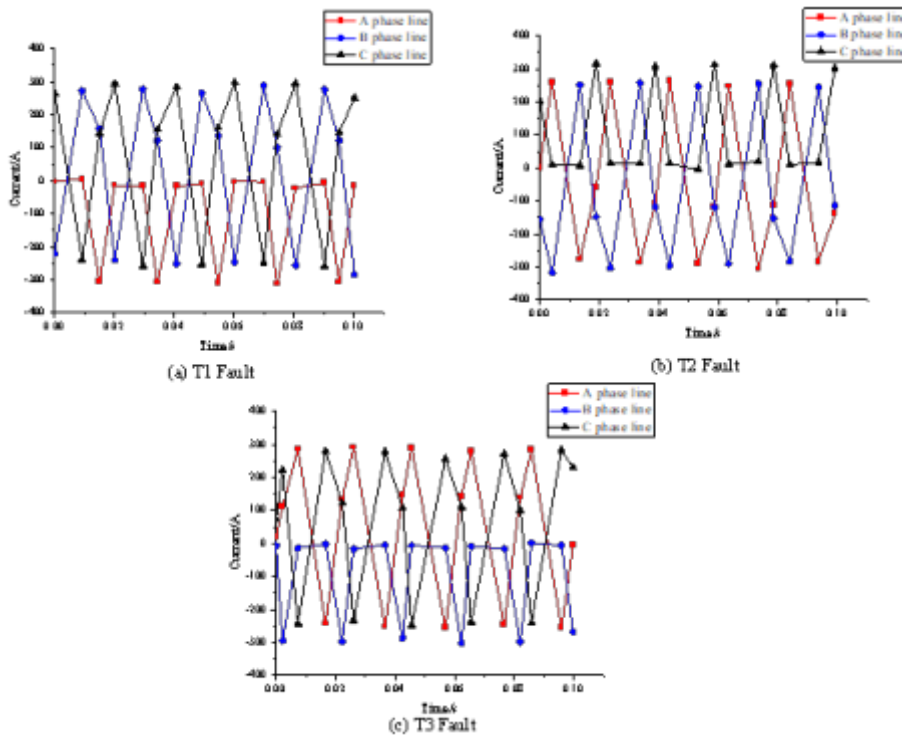


Fig. 3.1: Change in the current waveform of a hard-faulted single-insulated-gate bipolar transistor

current coincides every 0.02s, and the lower half of the current waveform is lost. When the T3 fault occurs, the current trend of phase A is consistent with that of the T2 fault. The specific current values are slightly different, but the overall range is the same. The waveform in the upper half of the phase B current is lost, and the current trend is first to fall and then rise. After rising, the current value remains 0.01s and drops again, and the consumption time of each process is about 0.02s. The trend of the current change in phase C is similar to the change in phase C at the T1 fault. In general, compared with the simulated current waveform under normal conditions, waveform loss of different phases will occur under different faults. The fault type can be

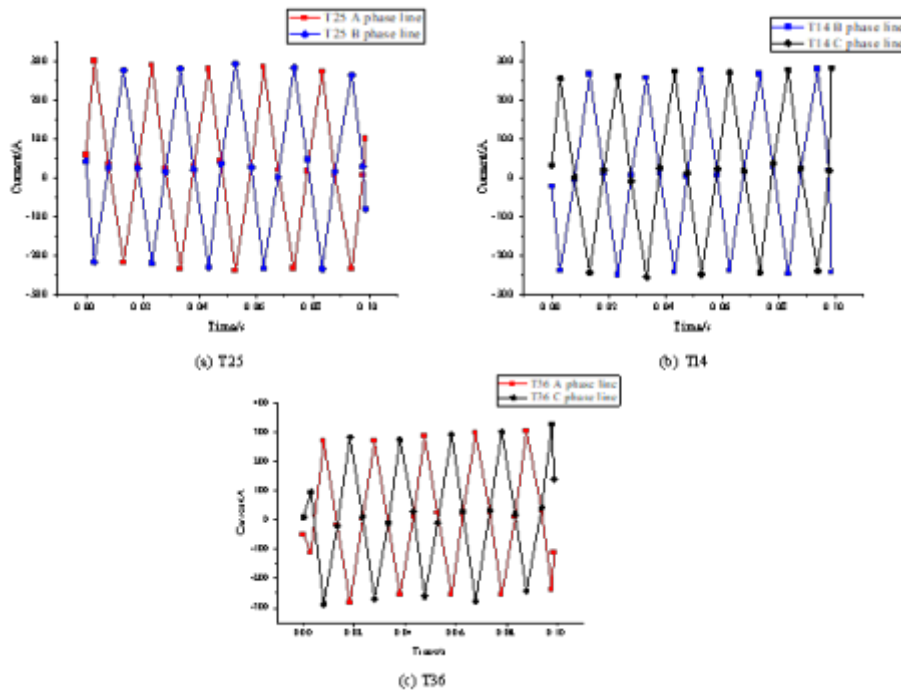


Fig. 3.2: Changes in current waveforms of two insulated-gate bipolar transistors on the upper and lower legs of a hard fault

reversed according to the waveform loss.

The current waveforms of the two insulated-gate bipolar transistors are plotted in Figure 3.2.

From Figure 3.2, in both transistor faults on the upper and lower bridge arms, the T25 fault shows only phase A current and phase B current, and the phase C current is completely lost. Phase A's current trend is first up and then down, and the current range is between 300 and -250A. There is no waveform loss. The current trend of phase B is completely opposite to the current trend of phase A. The current range of both is consistent, and the waveform is complete. In the T14 fault, only the B phase current and the C phase current are present, and the A phase current waveform is completely lost. The current trend of phase B is consistent with the current trend of phase B of the T25 fault, but the lower limit of the variation range is slightly lower than that of the T25 fault. The current trend of phase C is first rising and then decreasing. The overall current range is approximately 250 to -250A, and the waveform is complete. In the T36 fault, the A phase current and C phase current remain intact, and the B phase current is completely lost. Phase A's current change trend is to decline and then rise. The current variation range is about 300 to -300A, with large fluctuations in the middle. The current trend of phase C is first increasing and then decreasing, the lower limit of the current range exceeds -300A, and the waveform is complete. In general, the current waveform in the fault of the upper and lower bridge arms will have phase loss, and the phase loss of different power tubes will be different. In real circuits, this type of fault will cause a current short circuit.

3.2. Soft fault simulation results. The variation in the current waveform of a single insulated-gate bipolar transistor is revealed in Figure 3.3.

From Figure 3.3, in soft faults, resistance is mainly considered. At a thermally stable operating temperature of 30°C, the thermally stable resistance exceeds 0.004Ω. When a T1 fault occurs, the current in phase A tends to rise first and then fall. The current range is between 300 and -300A. Compared to normal, the amplitude of the current decreases, and the peak decreases slightly. The current trend of phase B is opposite to the current trend of phase A. The current range is consistent. The current amplitude is small compared to the simulated

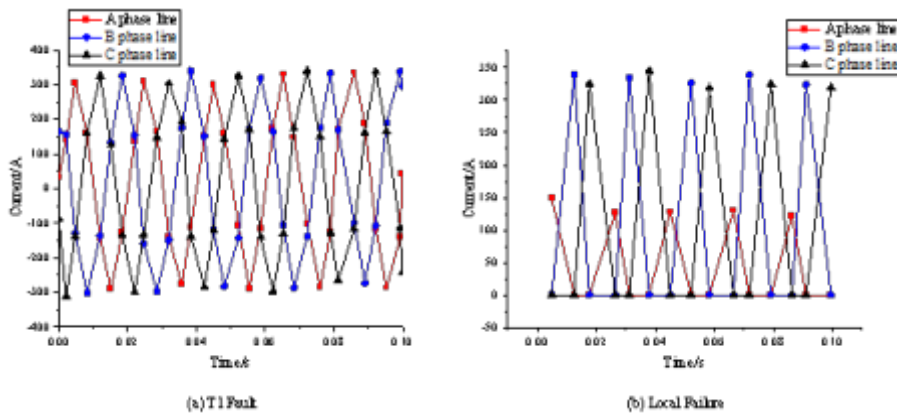


Fig. 3.3: Variation of the current waveform of a soft-faulted single-insulated-gate bipolar transistor

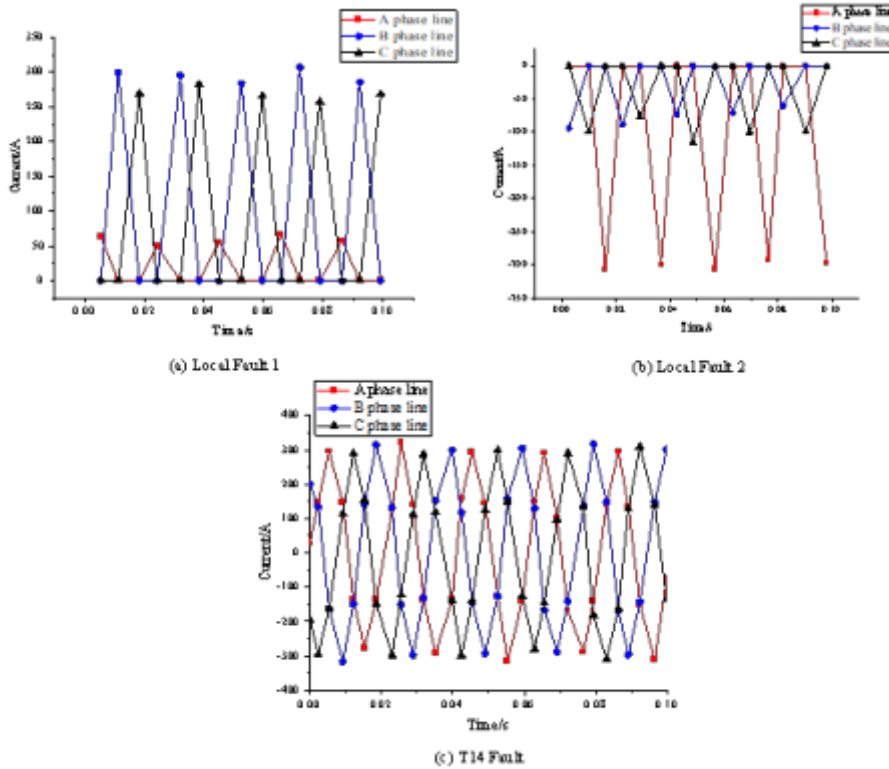


Fig. 3.4: Changes in current waveforms of two insulated-gate bipolar transistors with the same bridge above and below the soft fault

current waveform under normal conditions. The current variation trend of phase C is quite different from that of the simulation circuit under normal conditions, and the lowest value occurs when it is close to 0.02s. Overall, the soft-fault single transistor current waveform is intact, but the overall current amplitude is reduced.

The current waveforms of the two insulated-gate bipolar transistors are shown in Figure 3.4.

From Figure 3.4, when both transistors on the same bridge of the soft fault have a fault problem, the

current change trend of phase A increases first and then decreases, which is similar to the waveform of a single transistor with a soft fault[20]. Phase A currents range from 300 to -300A. Compared to the simulated current waveform under normal conditions, the current output value decreases in both the positive and negative half cycles. The current trend of phase B is completely opposite to the current trend in normal conditions. The current trend of the C phase is consistent with the case of a soft fault single transistor, and the current value at 0s is slightly different. On the whole, the current value of phase A of the two transistors with the same bridge above and below the soft fault will decrease, and the current value of phase B and phase C will not change much.

4. Conclusion. In this paper, the electromechanical fault characteristics of metro vehicles are extracted by big data technology, and the current situation under different fault conditions is analyzed. The simulation model is established. Additionally, the Simulink module in the MATLAB tool is used to simulate the current waveform of metro vehicles under different fault conditions and analyze the hard and soft fault phenomena. The simulation results show that: (1) The current waveform of a single transistor with the hard fault is compared with the simulated current waveform under a normal state. Waveform loss of different phases occurs under different fault conditions. The fault type can be reversed according to the waveform loss. When power tube T1 fails, the upper half of the phase A current waveform is lost. When the power tube T2 fails, the C-phase current coincides every 0.02s, and the lower half of the current waveform is lost. When T3 fails, the waveform in the upper half of the B phase current is lost. The current trend is first to fall and then rise, and the consumption time of each process is about 0.02s. (2) The current waveform in the hard fault's upper and lower bridge arms will have phase loss, and the phase loss of different power tubes is different. At the T25 fault, only phase A current and phase B current are shown, and phase C current is completely lost. In the T14 fault, only the B phase current and the C phase current are present, and the A phase current waveform is completely lost. In the T36 fault, the A and C currents remain intact, and the B phase current is completely lost. (3) The current waveform of a single transistor with a soft fault is complete, but the overall current amplitude is reduced. When T1 fails, the current trend of phase A rises first and then falls. Compared to normal, the amplitude of the current decreases, and the peak decreases slightly. The current trend of phase B is opposite to the current trend of phase A. The current variation trend of phase C is quite different from that of the simulation circuit under normal conditions, and the lowest value occurs when it is close to 0.02s. (4) The current value of phase A of the two transistors on the same bridge of the soft fault will decrease, and the current value of phase B and phase C will not change much. The current trend of phase A is to increase first and then decrease. Compared to the simulated current waveform under normal conditions, the current output value decreases in both the positive and negative half cycles.

The disadvantage is that the simulation model only studies the current waveforms of the two IGBTs on the upper and lower bridge arms of the hard fault and the upper and lower bridges of the soft fault. It does not consider the current waveforms of the two IGBTs under the same half-bridge and the cross half-bridge. The research on converter faults still needs to be completed. Furthermore, there is the problem of short simulation experiment time. In subsequent research, converter faults under different conditions can be added, the simulation time can be extended, and a more comprehensive study can be carried out. The electrical system faults can be inferred from different current waveforms to enrich the fault types, thereby enhancing the monitoring ability of the electromechanical fault system. The fault model of the transformer of the electrical system of EE of metro vehicles established by big data technology improves the monitoring ability of transformer faults of the metro electrical system. It also plays a specific role in improving the performance of the monitoring system of EE of metro vehicles, enhances the safety of metro operation and the quality of metro operation, and contributes to the development of urban rail transit.

REFERENCES

- [1] Liu Z., Li Y., Zhao L., et al. (2021) Construction of intelligent query system for metro electromechanical equipment faults based on the knowledge graph[J]. *Journal of Intelligent & Fuzzy Systems*, 41(3), 4351-4368.
- [2] Li S., Wei X., Zhang Z., et al. (2019) Subway Station Capacity Maintained by Optimizing a Maintenance Schedule of Key Equipment[J]. *Applied Sciences*, 9(7), 1386.

- [3] Ngiam K. Y., Khor W. (2019) Big data and machine learning algorithms for health-care delivery[J]. *The Lancet Oncology*, 20(5), e262-e273.
- [4] Yonghyun L., Minhyuk K., Jisoo K., et al. (2021) An exploratory study on the effectiveness verification and alternatives to improve the congestion level of subway line 9[J]. *IJEMR*, 5(4), 1-5.
- [5] Park S. T., Kim Y. K. (2019) A study on deriving an optimal route for foreign tourists through the analysis of big data[J]. *Journal of Convergence for Information Technology*, 9(10), 56-63.
- [6] Wang A., Togo R., Ogawa T., et al. (2022) Defect detection of subway tunnels using advanced U-Net network[J]. *Sensors*, 22(6), 2330.
- [7] Huang K., Lu S., Li X., et al. (2022) Predicted Mean Vote of Subway Car Environment Based on Machine Learning[J]. *Big Data Mining and Analytics*, 6(1), 1-14.
- [8] Welch T. F., Widita A. (2019) Big data in public transportation: a review of sources and methods[J]. *Transport reviews*, 39(6), 795-818.
- [9] Yin Z., Hu N., Chen J., et al. (2022) A review of fault diagnosis, prognosis and health management for aircraft electromechanical actuators[J]. *IET Electric Power Applications*, 16(11), 1249-1272.
- [10] Dalla Vedova M. D. L., Germanà A., Berri P. C., et al. (2019) Model-based fault detection and identification for prognostics of electromechanical actuators using genetic algorithms[J]. *Aerospace*, 6(9), 94.
- [11] Liu Z. H., Lu B. L., Wei H. L., et al. (2019) Fault diagnosis for electromechanical drivetrains using a joint distribution optimal deep domain adaptation approach[J]. *IEEE Sensors Journal*, 19(24), 12261-12270.
- [12] Arellano-Espitia F., Delgado-Prieto M., Martinez-Viol V., et al. (2020) Deep-learning-based methodology for fault diagnosis in electromechanical systems[J]. *Sensors*, 20(14), 3949.
- [13] Li Y. (2022) Exploring real-time fault detection of high-speed train traction motor based on machine learning and wavelet analysis[J]. *Neural Computing and Applications*, 34(12), 9301-9314.
- [14] Kushwaha A. K., Kar A. K., Dwivedi Y. K. et al. (2021) Applications of big data in emerging management disciplines: A literature review using text mining[J]. *International Journal of Information Management Data Insights*, 1(2), 100017.
- [15] Wang J., Yang Y., Wang T., et al. (2020) Big data service architecture: a survey[J]. *Journal of Internet Technology*, 21(2), 393-405.
- [16] Price W. N., Cohen I. G. (2019) Privacy in the age of medical big data[J]. *Nature medicine*, 25(1), 37-43.
- [17] Mohammadpoor M., Torabi F. (2020) Big Data analytics in oil and gas industry: An emerging trend[J]. *Petroleum*, 6(4), 321-328.
- [18] Yang C., Huan S., Yang Y., et al. (2020) Application of big data technology in blended teaching of college students: a case study on rain classroom[J]. *International Journal of Emerging Technologies in Learning (iJET)*, 15(11), 4-16.
- [19] Qi C. (2020) Big data management in the mining industry[J]. *International Journal of Minerals, Metallurgy and Materials*, 27(2), 131-139.
- [20] Qi H., Chen G., Ma H., et al. (2022) A Subway Sliding Plug Door System Health State Adaptive Assessment Method Based on Interval Intelligent Recognition of Rotational Speed Operation Data Curve[J]. *Machines*, 10(11), 1075.

Edited by: B. Nagaraj M.E

Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing

Received: Dec 28, 2023

Accepted: Mar 20, 2024



LEVERAGING EMOTIONS IN STUDENT FEEDBACK TO IMPROVE COURSE CONTENT AND DELIVERY

ABID HUSSAIN WANI*

Abstract. Emotions play a vital role in almost all the activities we perform, including learning. In fact, the success of any learning system is largely dependent upon its ability to deliver the course content in such a form so as to meet the learning requirements of the target audience. Learning Systems can be tailored to effectively utilize the feedback from learners to improve the course content, and thus the feedback can prove to be a valuable asset. There is an increased demand for focusing on a learner-centric approach to content delivery. In this study we attempt at detecting different learning-relevant emotions from the feedback for a course so as to enable course designers to incorporate the type of content that matches a learners requirements. Rather than taking into account six basic emotions (sadness, happiness, fear, anger, surprise and disgust) we consider interest, engagement, confusion, frustration, disappointment, boredom, hopefulness and satisfaction emotions for the purpose of our study since they are more relevant in a learning setup. We employed a supervised algorithm, Support Vector Machine, for affect detection from the textual feedback in our experiments.

Key words: Emotion Detection, Support Vector Machine, Content delivery, Feedback Mining

1. Introduction. The rapid growth in information technology coupled with the huge benefits it offers, has increased its adoption by leaps and bounds in almost every sphere of our lives. In the field of education too, ICT-based solutions are playing a major role and providing overall value addition to the process of learning [1]. Specifically, eLearning Systems have a potential to reform the traditional teaching structure by incorporating technology-driven learning stuff and more importantly catering to the needs of individual learners who may not have a face-to-face interaction with the instructor [2]. Although much research has been conducted to understand the learning requirements of a group audience, there is a lot of scope for improving the content presented to the audience with varied emotional states. Since the target group of learners can be in different emotional states the content delivery can be tailored so as to match and satisfy their individual needs. The essence of these systems lies in providing individualized instruction by being able to cater and furnish to the varying knowledge grasping capacities and information needs of prospective learners. In a learning setup, we are more concerned about effective absorption of the content by the learner. Therefore, contemporary eLearning Systems are designed to be more and more learner-centric and close to the students self-learning needs as per the demand from the student community. A course can be tagged as learner centric only if the course is rich enough to include content for learners who can be in different emotional states. The fact that the human brain performs a blend of both cognitive and affective processing demands that the systems which are modeled after it not only process the information from a cognitive point of view but also integrate and assimilate an affect sensing and processing functionality into the system. Thus, the technology-driven learning environments should not only be intelligent enough but also take into consideration emotional aspects [3]. Having emotion processing capability will enable these systems to customize the contents of a course and its flow as per the needs of the learner so as to make the learning process more productive. True engagement of the prospective learners both intellectually and emotionally forms the hallmark of productive eLearning. The student feedback pertaining to a course can prove as an asset to structure the content of a course and decide the delivery plan by detecting the emotions expressed in the feedback. In this work, we propose a framework grounded on supervised learning that performs this affect detection from student feedback and provides generalized guidelines for the course designers to fine-tune the course contents and delivery plan for each target category of students according to their mood and learning pattern.

*South, Campus, University of Kashmir, Jammu and Kashmir, India (abid.wani@uok.edu.in).

2. Related Work. The utility of emotions expressed by students in eLearning environments has gained much attention with the advancements in machine learning during the recent years. The major focus in this area has been to detect fine-grained emotions to track down the emotional response of the learners in an automated and seamless manner with least additional investment in terms of tangible and intangible resource [4, 5]. The research in this area comes under a more general area in computing known as “Affective Computing”. A good amount of research has been conducted to detect learner’s affective state [6, 7]. Most of the studies that have been carried out in this area principally take into account models and theories of emotions and employ a number of techniques to computationally capture the emotions. Since the focus of the present work is on recognizing emotions only from textual data, this section will trace the studies made for recognizing emotions from text. Emotion recognition involves theories and concepts from various areas including psychology, linguistics, information retrieval, text mining and recently machine and deep learning techniques have gained a lot of attention as well. Based on how emotions are modeled computationally and what techniques are employed, researchers have been able to achieve results which are close to that of humans. To detect emotions in text a number of approaches have been employed in different studies. Principally these approaches can be categorized into three categories; pure emotion keyword based approach, rule-based, (machine and deep) learning or combination of these. An emotion-aware framework for elearning systems has been presented in [8] that employs supervised learning approach. Regardless of the approach used, almost all methods have their own strengths and weaknesses when it comes to proper identification of emotional affinity [9]. Most emotion-recognition models are driven by identification of syntactic features (e.g. Parts-Of-Speech tags, N-grams etc.) as well as semantic features (e.g. synonym sets). There is a lot of scope for building syntactic and semantic resources which will serve as good resources for affect detection and a number of works have already been completed in this direction [10, 11]. Affect detection has proved to be an important tool in the deciphering of the interaction with the student for eLearning Systems in recent years [12]. An automated affect sensitive intelligent tutoring system, introduced by D’Mello and Graesser, tries to simulate dialogue patterns of tutor-pupil of real-world [13]. Emotions other than the basic ones have been considered for the purpose of affect computing in only a few studies. Liew et al. [14] constructed a text corpus comprising 15,553 tweets and annotated with twenty eight different emotions for their work on fine grained emotion recognition. Among the various emotions taken into account by them, emotions like boredom, confidence, excitement etc. are of particular interest in our study as well. In [15], Abdul-Mageed and Ungar have tried to build a large corpus using twitter hashtags thereby effectively removing the major impediment of creating instances labeled with emotion categories. For the purpose of our study, however, our corpus comprises of student feedback in textual form and moreover we take into account eight emotions which are most relevant to a learning environment. Our primary focus is on the detection of emotions relevant to learning activity from student feedback to fine-tune the course contents and delivery plan.

3. Detecting Emotions and Tailoring Course Content and Flow. Although the design of the content and its flow is dictated by various technical input factors relevant to a particular course yet the feedback received can be utilized to fine-tune the design. Fig. 1 shows the overall setup of our framework for detecting and utilizing affect information from student feedback for a particular course. The affect perception from students is then aggregated to take into account only the dominant emotions expressed to remove the outlier effect.

3.1. Detecting Emotions From Student Feedback. The task of emotion detection from student feedback has been modeled as a text classification task. After initial phase of preprocessing which consists of tokenization, POS tagging, lemmatization, stop-word removal, we train our supervised classifier Support Vector Machine. Support Vector Machines (SVM) learn to recognize emotions in data by identifying an optimal hyperplane that separates different emotion categories in a feature space derived from textual data. Initially, textual data is transformed into numerical feature vectors, incorporating linguistic cues. Through training on labeled datasets where each text sample is associated with a specific emotion label, SVM adjusts its parameters to maximize the margin between different emotion classes while minimizing classification errors. This process results in the creation of a decision boundary that delineates regions corresponding to distinct emotional categories. Subsequently, SVM utilizes this decision boundary to classify new text samples by determining which side of the boundary they fall on, effectively assigning them to the appropriate emotion category. The

classification of emotion from the student feedback takes place in the following steps:

1. **Feature Extraction:** The process begins with preprocessing the textual feedback, which involves tasks such as tokenization, removing stopwords, and stemming or lemmatization to normalize the text. Then, features are extracted from the preprocessed text. These features include various linguistic cues such as word frequencies, sentiment scores, part-of-speech tags etc. Each feedback is represented as a numerical feature vector based on these extracted features.
2. **Training Phase:** In the training phase, SVMs learn to distinguish between different emotion categories by finding the optimal hyperplane that separates the feature vectors belonging to each emotion class in the feature space. The hyperplane is a decision boundary that maximizes the margin, which is the distance between the hyperplane and the nearest data points (support vectors) from each emotion class. SVM aims to find the hyperplane that minimizes classification errors while maximizing the margin.
3. **Optimization:** SVMs optimize their parameters during training to find the hyperplane that best separates the emotion classes. This optimization process involves solving a constrained optimization problem, typically using techniques like gradient descent or quadratic programming. The objective is to find the parameters (weights and bias) that define the hyperplane, ensuring that it separates the classes with the maximum margin.
4. **Kernel Trick:** SVMs employ the kernel trick to handle nonlinear relationships between features and emotions. By transforming the feature space into a higher-dimensional space, SVMs can find a hyperplane that effectively separates the data points even when the relationships are nonlinear.
5. **Feedback Tagging:** Once trained, SVMs classify new textual feedback samples by determining which side of the decision boundary they fall on in the feature space. The signed distance of a data point from the decision boundary is used to predict its emotion category. SVM assigns the sample to the emotion class corresponding to the side of the decision boundary it lies on.

In our study we used two datasets. One is the student feedback dataset from Menekse et al. [16]. This dataset comprises both the students' responses and the gold-standard summaries created by the teaching assistant. The other dataset again encompasses student feedback, which Oza et al [17] have analyzed using artificial neural networks. Since the proposed method will be based on supervised learning we need a training corpus consisting of student feedback where each record has been assigned a certain emotion label or marked as neutral. As discussed above we take into consideration only those emotions which are relevant for affect detection in student feedback; those include interest, engagement, confusion, frustration, disappointment, boredom, hopefulness, satisfaction. For the purpose of our study, four annotators labeled both the datasets to serve as training data for Support Vector Machine, employed for emotion classification. The job of an annotator is simply to select an appropriate emotion for each of the student responses presented to him. Since the student feedback normally will be in the form of a sentence or paragraph hence the proposed system will limit the analysis to sentence level.

The interpretation of affect in text being highly subjective [18], as such it is quite possible that the perception of one judge/annotator is different from the other one. In order to account for this subjectivity, rather than being annotated by a single judge we put each student's feedback for annotation by four annotators/judges. To get a quantitative measure of the inter-judge agreement statistic Cohen's kappa is employed. The pairwise agreement in emotional categories for student feedback is shown in Table 3.1.

As evident from the inter-annotator agreement study, the perception for an emotion is person-specific. However from the pair-wise results it is evident those human judges mostly agree on the instances of interest, confusion and satisfaction.

3.2. Tailoring Course Content and Flow. Once the dominant emotional response is known, the next step is to use this affective feedback to customize the course content. For example, if the emotion detected is frustration (learner got frustrated), there is a requirement to generate hints to help the learner in understanding a particular topic, and include certain simple and illustrative examples. If the emotion detected is "boredom", then it would be imperative to display content for the learner that will get the learner look for his/her own titles of interest, likewise if the student is confused with course contents, more elaborate and worked examples need to be incorporated into the course so that the student can understand the concept being presented [19]. The emotion-inspired improvements from student feedback can be serve as a guiding factor for the inclusion of various

Table 3.1: Pair-wise Agreement in Emotional Categories

Emotion	<i>Judge1 ↔ Judge2</i>	<i>Judge1 ↔ Judge3</i>	<i>Judge1 ↔ Judge4</i>	Average
<i>interest</i>	0.85	0.77	0.76	0.79
<i>engagement</i>	0.65	0.66	0.54	0.61
<i>confusion</i>	0.87	0.81	0.84	0.84
<i>frustration</i>	0.75	0.78	0.70	0.74
<i>disappointment</i>	0.65	0.43	0.55	0.54
<i>boredom</i>	0.55	0.59	0.71	0.61
<i>hopefulness</i>	0.67	0.66	0.66	0.66
<i>satisfaction</i>	0.79	0.88	0.81	0.82

Table 3.2: Course Improvement Suggestions

Emotion Detected in the Feedback	Course Improvement Suggestion
interest	Links to more advanced topics
engagement	More detailed content
confusion	Elaborate worked examples on the topic
frustration	The more simple and precise explanation
disappointment	Elementary Video/ animations
boredom	Display lively and funny examples
hopefulness	More detailed content
satisfaction	Links to related topics

Table 4.1: Evaluation Results of the Proposed Framework

Emotion Category	<i>Menekse dataset</i>			<i>Oza dataset</i>		
	Precision	Recall	<i>F1 – score</i>	Precision	Recall	<i>F1 – score</i>
<i>interest</i>	58.54	62.78	60.59	54.21	57.77	55.93
<i>engagement</i>	47.63	52.47	49.93	56.23	62.5	59.20
<i>confusion</i>	53.23	58.12	55.57	48.30	51.23	49.72
<i>frustration</i>	59.22	55.47	57.28	52.46	67.33	58.97
<i>disappointment</i>	54.65	61.45	57.85	64.55	61.25	62.86
<i>boredom</i>	68.66	76.98	52.58	71.23	65.85	68.43
<i>hopefulness</i>	56.30	62.45	59.22	48.56	55.75	51.91
<i>satisfaction</i>	46.25	51.22	48.61	45.67	62.33	52.72

value-additions like explanation, hints, worked examples to the course content as described in Table 3.2.

4. Evaluation and Results. The real essence of any machine learning-based framework lies in not only providing good results (classification results in our case) on the data on which it is trained but also on new and unseen data and under-fitting and overfitting of data is avoided. We validated our model using Leave-one-out cross validation, by training our model on all the instances of student feedback except one, “n” number of times and predicting Output emotion label for that one instance using Support Vector Machine. Table 4.1 depicts the classification performance results obtained for the above two datasets.

To measure the performance of the proposed framework, we compute the classification metrics including Precision, Recall and F1-score on Menekse’s dataset and Oza’s dataset. For Menekse’s dataset, best results were obtained for ‘interest’ emotion and that for Oza’s dataset ‘boredom’ emotion category was detected with highest F1-score. Classification metrics such as Precision, Recall, and F1-score are essential for evaluating the performance of machine learning models, particularly in tasks like emotion detection from text. Precision quantifies the proportion of emotions detected that are correctly reported as positives (true positives) from the

whole number of examples which are tagged as positive (true positives + false positives) by the model. For the purpose of task of emotion classification, precision presents ability of a model to accurately identify a particular emotion category without misclassifying unrelated emotions. A model with high precision score implies that when the it suggests an emotion, there is high probability for it to be correct, minimizing false positives and ensuring the relevance and specificity of the predictions. Recall measures the proportion of correctly predicted positive cases (true positives) out of all actual positive cases in the dataset (true positives + false negatives). Overall, precision, recall, and F1-score are critical metrics for evaluating the performance of emotion detection models, as they offer insights into different aspects of the model's performance, such as accuracy, relevance, and comprehensiveness.

5. Conclusion and Future Scope. This paper proposes a supervised learning based emotion detection mechanism for detecting emotions from student feedback to tailor and enrich the course contents. The classification results obtained on Menekse's dataset and Oza's dataset does not seem too good which is due to small size of the text corpus. In future, we aim to build a larger annotated text corpus so as to achieve better classification performance. The limitation of our work lies in the fact that we only take into account the textual feedback received from the learner to decide on the content. The course content and delivery can further be improvised by taking into account other modalities including facial expressions, speech and voice modulations and not just written textual utterances. We limited our study to text as we take into account only those systems which receive just the textual feedback from the learner. In future, this work can be extended to have multi-modal feedback so as to guide the content design, presentation and delivery. In our experiments, though we have not made appropriate substitution of slangs (e.g. "dunno" with "don't know") and orthographic features (e.g. "sooconffuuusing"), we intend to take these issues into account as well to better gauge the emotion expressed in the student feedback., so as to enable course designers get a better idea about what all improvements need to be made to the course content and flow. Moreover, it must be noted here that detecting learners' current mood is not the only solution but one factor to be considered; there is more research needed to identify other situational and technical input factors that should guide the design, content and delivery of a course.

REFERENCES

- [1] K. Meteshkin, O. Sokolov, O. Morozova, and N. Teplova, Integration of Higher and Secondary Education: Problems and ways of their solution on the basis of Information Technologies, *Journal of Education, Health and Sport* **6** (2016), no. 7, pp. 375–390.
- [2] I. Ibrahim and K. Al, Smart Learner-Centric Learning Systems Smart Learner-Centric Learning Systems, *Proceedings of the International Conference on Information Science and Applications ICISA 2017*, Barcelona, Spain, March 20, 2017, Springer, Singapore (2017), 577-584
- [3] M. Feidakis, T. Daradoumis, and S. Caballé, Endowing e-learning systems with emotion awareness, *Proceedings of the 3rd IEEE International Conference on Intelligent Networking and Collaborative Systems, INCoS 2011*, Fukuoka, Japan, Nov 30 –Dec 2, 2011, IEEE, Japan (2011), 68–75
- [4] M. Wong, Emotion Assessment in Evaluation of Affective Interfaces, *Neuron* **65** (2006), no. 3, p. 293.
- [5] G. Zimmermann, Beyond Usability – Measuring Aspects of User Experience, Ph.D. thesis, Swiss Federal Institute of Technology Zurich (2008).
- [6] S. K. D. Mello, S. D. Craig, A. Witherspoon, and B. Mcdaniel, Automatic detection of learner 's affect from conversational cues, *User Modeling and User-Adapted Interaction* (2008), no. 1, pp. 45–80.
- [7] Wang, Yang. "Affective state analysis during online learning based on learning behavior data." *Technology, Knowledge and Learning* 28.3 (2023): 1063-1078.
- [8] AH.Wani, R. Hashmy, An Emotion-aware Framework for eLearning Systems, *Proceedings of 1st International Interactive Information Retrieval Conference*, Kish, Iran, Feb 22–24, 2017, Tehran University, Tehran (2017), 25–30
- [9] Alslaity, Alaa, and Rita Orji. "Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions." *Behaviour & Information Technology* 43.1 (2024): 139–164.
- [10] S. Aman and S. Szpakowicz, Identifying Expressions of Emotion in Text, *Proceedings of the International Conference on Text, Speech and Dialogue*, Kuala Lumpur, Malaysia, April 3–5, 2009, IEEE, Malaysia (2009), 195–205
- [11] S. Afzal and P. Robinson, Designing for Automatic Affect Inference in Learning Environments, *Journal of Educational Technology & Society* **14** (2011), no. 4, pp. 21–34.
- [12] E. Bevacqua et al., Interacting with emotional virtual agents, *Proceedings of the International Conference on Intelligent Technologies for Interactive Entertainment*, Genova, Italy, May 24–27, 2011, Springer, Italy (2011), 243–245
- [13] S. K. D'Mello, B. Lehman, and A. Graesser, A Motivationally Supportive Affect-Sensitive AutoTuto. In *S. K. D'Mello (ed) New perspectives on affect and learning technologies*, Springer, New York (2011), 113–126.
- [14] J. Liew, S. Yan, H. R. Turtle, and E. D. Liddy, EmoTweet-28: A Fine-Grained Emotion Corpus for Sentiment Analysis,

- Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Portoroz, Slovenia, May 23–28, 2016, Springer, Slovenia (2016), 1149–1156
- [15] M. Abdul-Mageed and L. Ungar, EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks, *Proceedings of the 55th annual meeting of the association for computational linguistics*, Vancouver, Canada, July 03–Aug 4, 2017, Association for Computational Linguistics, Canada (2017), 718–728
- [16] M. Menekse, G. Stump, S. Krause, and M. Chi, The effectiveness of students' daily reflections on learning in engineering context, *Proceedings of the ASEE Annual Conference and Exposition*, Vancouver, Canada, June 26–29, 2011, Canada (2011), 1149–1156
- [17] K. S. Oza, R. K. Kamat, and P. G. Naik, Student Feedback Analysis: A Neural Network Approach, *Proceedings of the Tenth International Conference on Information and Communication Technology for Intelligent Systems*, Ahmedabad, India, March 25–26, 2017, Springer, Cham (2017), 342–348
- [18] H. Binali, C. Wu, and V. Potdar, Computational approaches for emotion detection in text, *Proceedings of 4th IEEE International Conference on Digital Ecosystems and Technologies*, Dubai, United Arab Emirates, April 13–16, 2010, IEEE, Cham (2010), 172–177
- [19] S. D'Mello and R. A. Calvo, Beyond the basic emotions: what should affective computing compute?, *Proceedings of CHI '13: CHI Conference on Human Factors in Computing Systems*, Paris, France, April 27– May 2, 2013, Association for Computing Machinery, United States (2013), 2287–2294

Edited by: Mudasir Mohd

Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions

Received: Jan 1, 2024

Accepted: Apr 12, 2024

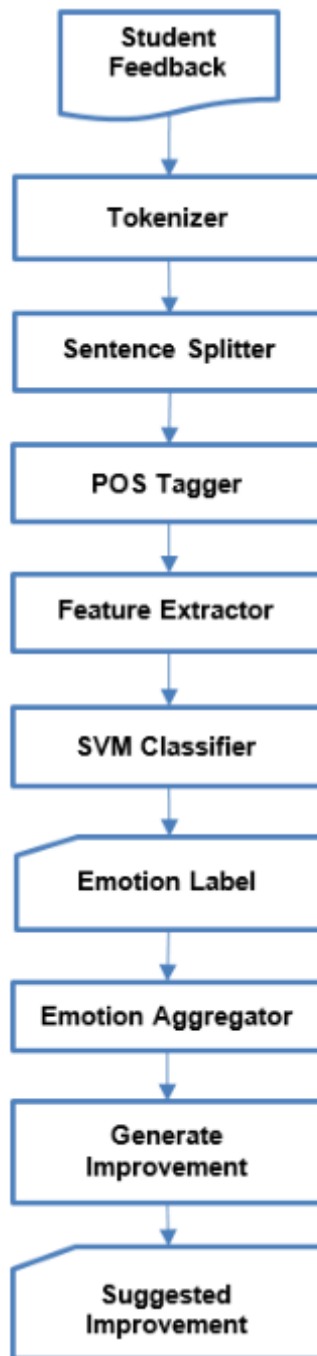


Fig. 3.1: Overall Framework for Course Improvement Using Student Feedback



FEATURE EXTRACTION OF GYMNASTICS IMAGES BASED ON MULTI-SCALE FEATURE FUSION ALGORITHM

KUN TIAN* AND QIONGHUA XIA[†]

Abstract. The feature extraction and analysis of gymnastics images is an essential foundation for estimating human posture. The primary step is to obtain various joint points of athletes based on basic information such as human texture and contour in the sports images. The reconstruction and analysis of the human skeleton is completed based on the feature data of the joint points. In this process, traditional algorithm models often have certain shortcomings in the accuracy of feature extraction for motion images. This paper combines multi-scale feature fusion algorithms to construct a gymnastics motion image feature extraction model, which can achieve more accurate and efficient analysis and research for the feature extraction process of motion images, further improving the detection accuracy of joint points in motion images; this lays an essential foundation for feature extraction of gymnastics images. At the same time, it also provides more methods for skeleton reconstruction based on image feature information during the motion process, improving the efficiency and accuracy of reconstruction.

Key words: Multi scale feature fusion; gymnastics; feature extraction; loss function

1. Introduction. With the continuous development and improvement of scientific and technological informatization, modern technology can support humans to obtain a large amount of image and video information from various channels. A common image processing technology is a technology that uses computers to process and analyse collected images and video information to meet human needs, covering various aspects of human clothing, food, housing, and transportation. The extraction of human skeleton features in images has always been an important research field in image processing and computer vision, and research in this field continues to drive the continuous updates and progress of modern audio and video technology [1-3]. The research and analysis of skeletal information during human motion lays an important foundation for further processing of images and videos, and can also help people analyse the behaviour and key actions of the target human body. For gymnasts, analysing and researching motion images is an extremely important process to improve and enhance their key movements. Therefore, using modern computer technology to extract and analyse the features of gymnastics images provides more means for this process. The human skeleton extraction algorithm divides the human skeleton into multiple joint points, such as the head, buttocks, shoulders, wrists, etc. Then, by analysing the position, direction, and motion of each joint point, the human skeleton information is obtained [4-5]. Further analysis of human posture and behaviour is carried out through the drawn human skeleton, in order to obtain the activity and motion information of the human body in the image.

The extraction of human skeletons in gymnastics images can be divided into two dimensions: two-dimensional human skeleton extraction and three-dimensional human skeleton extraction. 3D human skeleton extraction analyzes images obtained from 3D cameras such as Kinect to obtain the 3D shape or coordinates of human skeleton points. 2D human skeleton extraction mainly analyses 2D images obtained by ordinary image acquisition devices and obtains the 2D coordinates of human skeleton points. Compared to 2D skeleton extraction, 3D skeleton points are denser, modeling is more complex, and 3D cameras are more expensive [6-7]. In most cases, two-dimensional images are obtained for the analysis of gymnastics images. Therefore, this article mainly focuses on the feature extraction of two-dimensional gymnastics images.

The applications related to human pose estimation are all based on obtaining clear and accurate human skeletons in motion images. If the accuracy of feature extraction in motion images is insufficient, it will lead to

*School of Sports and Art, Guangzhou Institute of Physical Education, Guangzhou 510500, China

[†]Department of physical education, Guangdong University of Foreign Studies, Guangzhou 510420, China (Corresponding author, Qionghua_Xia23@outlook.com)

significant deviations in the analysis of human behaviour and movements. Therefore, improving the accuracy of feature detection and extraction in gymnastics images is of great significance. In recent years, the rapid development of the hardware field has led to the continuous enhancement of computer computing power, and more high-performance algorithms for extracting gymnastics motion images are emerging. The accuracy of human skeleton extraction is constantly improving, and the extraction technology, as the foundation of human posture recognition, will play an increasingly important role in more fields [8].

Early human pose estimation methods mainly relied on manually designed features and template matching, but this method was limited by the complexity and robustness of feature design. With the rise of deep learning, methods based on Convolutional Neural Networks (CNN) have gradually taken a dominant position. These methods train networks with a large amount of data to automatically learn the mapping from the original image to pose information, significantly improving the accuracy and robustness of estimation.

In recent years, researchers have begun to focus on more complex scenes and poses, such as multi-person pose estimation, 3D pose estimation, etc. These issues are more challenging as they require handling occlusion, changes in perspective, and changes in scale. To address these issues, researchers have proposed methods such as multi-scale feature fusion, multi-perspective fusion, and spatiotemporal modeling, further improving attitude estimation performance.

Feature extraction is a crucial step in computer vision tasks, aiming to extract helpful information from the original image for subsequent tasks. Selecting feature extraction techniques is crucial for the final performance in human pose estimation. Traditional feature extraction methods mainly rely on manually designed feature descriptors, such as SIFT, HOG, etc. Although these methods perform well in some simple scenarios, they are challenging to cope with complex and ever-changing human postures and environments. With the development of deep learning, CNN-based feature extraction methods have gradually become mainstream. CNN can automatically learn the mapping from the original image to high-level semantic features, effectively extracting helpful information for pose estimation. In addition, researchers have proposed methods such as multi-scale feature fusion and attention mechanisms to improve the accuracy and robustness of feature extraction.

Gymnastics image analysis faces many challenges. Firstly, gymnastics movements are complex and varied, requiring accurate capture and recognition of various subtle movements. Secondly, gymnastics competition scenes often have occlusion and changes in perspective, which puts higher demands on the robustness of image analysis algorithms. In addition, the differences in body shape and movement habits among different athletes also pose challenges to image analysis. In response to these challenges, researchers have begun to explore methods based on deep learning and multi-scale feature fusion. The mapping relationship from the original image to the gymnastics movements is automatically learned by training a deep neural network model. Meanwhile, multi-scale feature fusion technology can extract feature information at different scales to better cope with complex and ever-changing gymnastics movements.

The multi-scale feature fusion algorithm can comprehensively describe the characteristics and patterns of gymnastic movements by fusing feature information from different scales. Multi-scale feature fusion algorithms can extract feature information from different scales, such as athlete joint positions, movement trajectories, muscle morphology, etc., and combine temporal information to analyze actions continuously. In addition, multi-scale feature fusion algorithms can be combined with other advanced technologies, such as attention mechanisms, spatiotemporal modeling, etc., to improve the accuracy and robustness of gymnastics image analysis. For example, by introducing attention mechanisms, it is possible to automatically focus on feature regions that are more critical for recognizing gymnastics movements. Through spatiotemporal modeling techniques, it is possible to capture better and analyze the temporal changes and spatial relationships of gymnastics movements [9-10].

This article combines the current problems and related shortcomings in this field. It combines the unique characteristics of gymnastics images to construct a multi-scale fusion algorithm-based gymnastics image feature extraction model. In response to some joint points that are difficult to detect due to slight occlusion caused by environmental disturbances, multi-person interference, and human posture, this article gradually improves the performance of the constructed network model by enriching feature representation, enhancing the utilization of relevant features, effectively balancing, and fusing features of different scales. At the same time, combined with difficulty mining mechanisms, it improves the detection accuracy of challenging joint points in the feature

extraction process of moving images.

This study aims to develop a gymnastics image feature extraction model based on a multi-scale feature fusion algorithm to address the current accuracy and efficiency challenges in gymnastics image analysis. This model is expected to achieve more precise capture and comprehensive feature expression of gymnast movement details, providing strong technical support for subsequent gymnastics movement recognition, evaluation, and teaching. Athletes' movements are complex and varied in gymnastics, often containing rich spatial and temporal information. Therefore, for feature extraction of gymnastics images, it is necessary to consider feature information at different scales to capture the details and overall structure of athlete movements fully. The multi-scale feature fusion algorithm can effectively combine feature information from different scales, improving the accuracy and robustness of feature extraction. This study will first collect much gymnastics exercise image data and carry out preprocessing and annotation work. Then, design a multi-scale feature fusion algorithm that extracts and fuses image features of different scales to construct a comprehensive and effective feature representation. On this basis, further optimization of the algorithm parameters and structure is needed to improve the accuracy and efficiency of feature extraction.

2. Basic Theory of Multiscale Feature Fusion Algorithms.

2.1. Image Multi-scale Characteristics and Smoothing Models.

2.1.1. Image Multiscale Characteristics. For different scene images, observation can quickly and accurately identify objects of interest from the image, which is a visual characteristic that benefits from the observer's ability to adaptively adjust the distance between the human eye and objects according to different scenes to analyse image content. When the human eye observes an image at close range, even subtle changes in colour in local areas of the image are clearly visible, which is beneficial for the observer to obtain detailed information such as image edges and textures. However, image textures can worsen the consistency of colour distribution in the image area, and form a pseudo edge effect inside the area, thereby affecting the observer's understanding of the main contour of the image object [11-13]. As the distance between the human eye and the image continues to increase, observers can more easily capture the overall overview of the area where the semantic object is located and the contours between different objects. At the same time, the attention to subtle changes caused by the internal texture of the object is gradually decreasing.

In order to overcome the influence of image texture on feature extraction and combine the variation of image visual effect with observation distance, it is more important to establish a multi-scale edge-preserving smoothing model for images, which is used to obtain the smoothing components of the original image at different smoothing scales. As the smoothing scale continues to increase while protecting the edges of the image, the texture details inside the image area are gradually smoothed. By combining multi-scale edges and color distribution features of the image, accurate extraction of foreground objects in natural images is achieved [14]. To learn the global features of an image from multiple cascaded features of different scales, this section proposes an encoded Transformer network, which uses the Transformer to learn the segmented features of images at different scales, thereby obtaining the full-scale feature representation of the image. As shown in Figure 2.1, a schematic diagram of the network structure of the scale encoding Transformer model is provided.

2.1.2. Image Edge Preserving Smoothing Model. For gymnastics images, it can be seen as an organic combination of the object's main structure and texture details. The object's main structure is usually represented as a single or multiple areas with consistent brightness/colour in the image, while the image texture is usually represented as a periodic pixel fluctuation on the object's surface [15]. Therefore, image edge preserving smoothing can be understood as preserving important geometric attributes such as the main contour of the solid object, removing texture details on the surface of the solid object. For an input image $u^0(x, y)$, formula (2.1) is given:

$$u(x, y) = u^0(x, y) - t(x, y) \quad (2.1)$$

Among them, x represents the spatial position of image pixels in the horizontal direction; y represents the spatial position of image pixels in the vertical direction; $T(x, y)$ represents the texture component of image $u^0(x, y)$; The smoothing component $u(x, y)$ is a simplified approximation of the original image, mainly used to extract the overall overview features of each object in the image.

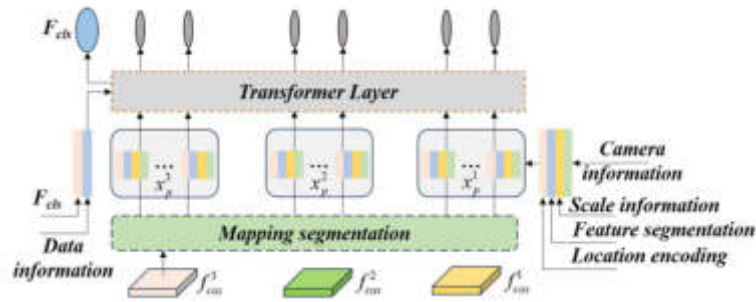


Fig. 2.1: Scale Encoding Transformer Network Architecture

Due to the fact that the smoothing component $u(x, y)$ discards the texture information in the original image, its pixel brightness or colour only changes at the edge of the image, while the degree of change inside the image area is relatively weak. Its mathematical representation is similar to a piecewise constant function. In mathematics, gradients are often used to describe the sudden changes in brightness or colour of image pixels [16]. Therefore, in order to measure the approximation degree between the smooth component and the piecewise constant function, a gradient function is used to calculate the overall change in the colour or brightness of $u(x, y)$ pixels in the smooth component, as shown in formula (2.2):

$$\Delta L = \int_{\Omega} f(|\nabla u|)d\Omega \tag{2.2}$$

Among them, Ω represents the domain of image theory; Represents a gradient function, $f(|\nabla u|)$ is also known as a diffusion function in the field of image processing; ∇u represents the gradient information of the smooth component $u(x, y)$, which satisfies the description given in formula (2.3):

$$\begin{cases} \nabla u = [u_x, u_y] \\ |\nabla u| = \sqrt{u_x^2 + u_y^2} \end{cases} \tag{2.3}$$

The texture components $t(x, y)$ describe slight pixel perturbation changes in the image area. Due to the limited range of values for image pixel brightness or color, the overall variation of texture components $t(x, y)$ satisfies the inequality (2.4):

$$0 \leq \frac{1}{|\Omega|} \int [u^0(x, y) - u(x, y)]^2 d\Omega = a_0^2 \leq C \tag{2.4}$$

Due to the equality constraints in the above model, the Lagrange multiplier method is first used to transform it into the following unconstrained functional optimization problem, as shown in formula (2.5):

$$M(u^0, u) = \int_{\Omega} f(|\nabla u|)d\Omega + \frac{\lambda}{2} \int_{\Omega} (u - u^0)^2 d\Omega \tag{2.5}$$

Among them, the first term is the regularized energy term, which suppresses image texture details by imposing constraints on the gradient amplitude of the image; The second item is the data fidelity term, which uses the difference measure between the original image u^0 and the smoothing component u to protect important detail features such as foreground contours; λ is a LaGrange multiplier used to balance the smoothness of the region and the strength of edge protection. Since equation (2.5) can be regarded as the independent variable u as the functional of the smooth component ∇u and its gradient image, as shown in equation (2.6):

$$F(u, \nabla u) = f(|\nabla u|) + \frac{\lambda}{2} (u - u^0)^2 \tag{2.6}$$

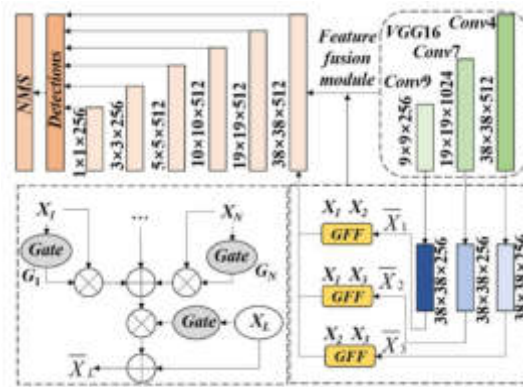


Fig. 2.2: Network Architecture for Multi-scale Feature Fusion

In addition, according to the Euler Lagrange equation, when there is an optimal solution to the above equation, formula (2.7) is satisfied:

$$F_u - \frac{\delta}{\delta_x} [F_{u_x}] - \frac{\delta}{\delta_y} [F_{u_y}] = 0 \tag{2.7}$$

2.2. Multi-scale Feature Fusion Network Architecture. The main challenge faced by the extraction and analysis of gymnastics image features is to analyze the ever-changing and complex situation of the movement process, which increases the difficulty of accurately detecting targets and is prone to missed detection. In addition, the contradiction between detecting and analyzing details of different body parts will become more prominent, resulting in low detection accuracy [17-18]. Therefore, fundamental research topics further enhance the model's ability to process details in gymnastics images, accurately extract features in complex environments, and accurately judge their status.

Traditional standard algorithms only perform independent output predictions on feature layers of different scales, with no connection between layers. Shallow feature maps are beneficial for target localization but lack sufficient semantic information. As convolutional neural networks deepen, feature maps can represent more semantic information, which is beneficial for target recognition but not conducive to target localization. Therefore, traditional networks need help solving the contradiction between target recognition and localization, and they cannot obtain helpful multi-layer information in images [19-21]. Based on the analysis of traditional network models, this paper combines multi-scale fusion algorithms to construct a network model with higher accuracy in feature extraction in gymnastics motion images. This model combines detailed features with global semantic features through a multi-feature layer fusion strategy and a lightweight and efficient feature fusion module. This can effectively alleviate the contradiction between target localization and recognition in traditional network target detectors. As shown in Figure 2.2, a network architecture for multi-scale feature fusion is presented.

This architecture adopts the idea of multi-scale feature fusion to strengthen the connections between various feature layers, combines the advantages of convolutional neural networks in high and low layers, and combines the useful feature information of high and low layer feature maps [22]. By increasing the semantic information of shallow feature maps and the positioning information of deep feature maps, the detection accuracy of targets is improved, and missed detections are reduced to achieve better performance.

3. A Feature Extraction Model for Gymnastics Images in Multiscale Feature Fusion Algorithms.

3.1. A Moving Image Enhancement Network Based on Multiscale Features.

3.1.1. Scale Selection and Loss Function. In the multi-scale smoothing process of gymnastics motion images, the texture area of the image gradually tends to become smooth as the smoothing scale increases,

which enhances the consistency of colour distribution in the image area. However, when the smoothing scale is too large, the image may have a false overlap effect between the front and background contours due to being too smooth, leading to poor foreground contour positioning accuracy. In addition, the execution time of the algorithm increases with the iteration process, and the larger the smoothing scale, the more time it consumes [23-24]. However, due to the unknown optimal scale for image edge smoothing, this article designs iteration termination conditions for the algorithm based on the feature extraction results of gymnastics motion images at different scales to ensure an appropriate scale for feature extraction to stop. Firstly, the Jaccard Distance is used to measure the similarity of the feature extraction results of the selected images at adjacent scales, and a similarity index is defined as shown in formula (3.1):

$$Js(i) = \frac{Card(T_F^i \cap T_F^{i-1})}{Card(T_F^i \cup T_F^{i-1})} \tag{3.1}$$

The similarity index $Js(i)$ uses the Intersection over the Union standard to quantify the similarity of adjacent scale foreground extraction results. The larger the value, the higher the similarity of adjacent scale foreground extraction results. The original gymnastics motion image generally contains many textures and colors. When the smoothing scale is low, the residual textures in the smoothing components can lead to poor parameter estimation accuracy. As the smoothing scale increases, the texture of moving images is gradually removed, the accuracy of parameter estimation continues to improve, and the accuracy of foreground extraction and similarity index $Js(i)$ continues to improve [25-27]. When the smoothing scale is too large, the image is excessively smoothed, resulting in a pseudo overlap between the image's front and background contour boundaries. The segmentation curve crosses the foreground boundary, decreasing the foreground extraction accuracy and similarity index $Js(i)$. Therefore, based on the similarity index $Js(i)$ and the trend of algorithm operation time with smoothing scale, the iteration termination condition of the algorithm in this paper is defined as shown in formula (3.2):

$$\xi [Js(i + 1)] 0 \tag{3.2}$$

Among them, the meaning of $\xi [\square]$ represents the operator of backward differentiation, as shown in formula (3.3):

$$\xi [Js(i + 1)] = Js(i + 1) - Js(i) \tag{3.3}$$

The multi-scale pyramid network based on attention mechanism uses thermal graph regression to conduct back error propagation, predict the Gaussian heat map of each joint point in the gymnastics motion image, and finally obtain the coordinate position of the joint point by finding the peak. Assuming we give a training set $\{T, J\}$, where T is the training set image set and J is the annotated joint point set, where the coordinates of the k-th joint point are (j_k^1, j_k^2) and the size of the thermal map is $H \times H$. B_i ($i=1, \dots, 5$) represents the i-th branch, and the thermal annotation process for each joint point is shown in formula (3.4):

$$g(j_k, x, y) = \frac{1}{\sqrt{2\pi\sigma_{B_i}^2}} e^{-\frac{(x-j_k^1)^2+(y-j_k^2)^2}{2\sigma_{B_i}^2}} \begin{cases} x = 1, 2, \dots, H \\ y = 1, 2, \dots, H \end{cases} \tag{3.4}$$

The multi-scale pyramid network based on attention mechanism will undergo error backpropagation in both stages 1 and 2. In stage 1, each branch will calculate a loss function, and then calculate their average value as the loss function for stage 1, as shown in equation (3.5):

$$L_1 = \frac{1}{k \times b} \sum_{i=1}^b \sum_{m=1}^k [y_{B_i,m} - g(j_m)]^2 \tag{3.5}$$

Calculate the loss function of all joint points, then classify the k joint points with more significant loss functions as the more challenging to detect joint points. At the same time, to prevent network oscillations caused by a significant difference between the maximum and minimum values of the loss function in a batch of

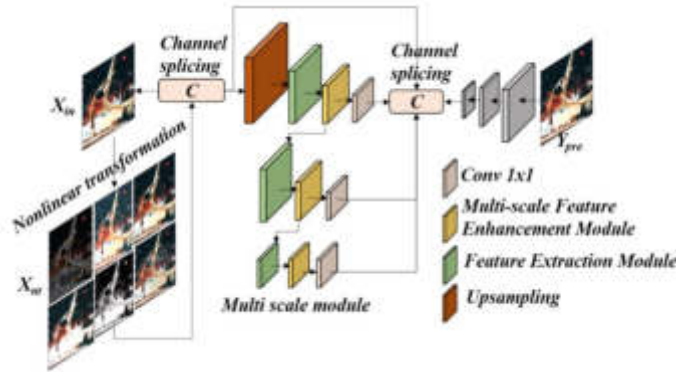


Fig. 3.1: A Model Framework for Multi-scale Feature Fusion Image Enhancement Network

gymnastics training images, the loss function in Stage 2 adopts an equalization strategy on a batch-by-batch basis. In summary, the definition of the loss function for stage two is shown in formula (3.6):

$$L_2 = \frac{1}{k \times n} \sum_{t=1}^n \sum_{m=1}^k \left[\hat{y}_{t,m} - g(j_m) \right]^2 \quad (3.6)$$

3.1.2. Multi Scale Feature Enhancement Module. Based on the above analysis and discussion, combined with the idea of end-to-end multi-scale feature fusion network, this paper constructs a multi-scale feature fusion algorithm to alleviate the problems of artifacts and noise in enhanced images. Firstly, non-linear transformation is performed on the gymnastics motion image to expand the low grayscale portion of the image and compress the high grayscale portion to enhance dark details. Then, channel fusion is performed with the original image before entering the network to enrich the original features; Subsequently, the fused features of the two images are fed into a multi-scale feature enhancement module, which includes an FEM, MFEM, and up sampling block. As the depth of the network layer increases, shallow features will decrease. Therefore, the network adopts a multi-scale feature fusion method to fuse low-level information with high-level information, reducing the amount of information lost due to functional loss [28]. The encoder in the enhancement module consists of convolutional kernels of size 3×3 , while the decoder consists of transposed convolutions and activation layers of the same size. Then fuse each extracted scale feature with the input image to output the final enhanced image. As shown in Figure 3, a model framework for multi-scale feature fusion image enhancement network is presented.

In the gymnastics motion image feature extraction model based on a multi-scale fusion algorithm, the feature extraction module consists of a convolution layer and a ReLU activation function. The convolution kernel size of the convolution layer is 3×3 , with a step size of 1. After each layer of convolution, there is a ReLU activation function to enhance the nonlinear representation of the network. After passing through the feature extraction module FEM at each scale, it will serve as the input of MFEM, and the output of the previous layer's enhancement module will serve as the input of the next layer's feature extraction module. In the multi-scale module, three feature extraction modules with different scales can extract image features of different scales.

The enhancement module in this article adopts an encoder decoder structure, where the encoding part consists of a convolutional kernel of size 5×5 and an activation function layer. The encoder, as a whole, presents a structure where the scale of the feature map gradually decreases, continuously reducing the resolution of the feature map to capture contextual semantic information. The decoder section consists of a deconvolution layer and an activation function layer, which also includes four stages. At each stage, after up sampling the input feature map, in order to reduce feature loss, it is concatenated with the corresponding proportion of the feature map in the encoder [29]. Unlike U-net, the decoder in this article does not use pooling layers, which can reduce data loss during the feature extraction process of gymnastics images.

3.2. Construction of a Feature Extraction Model for Gymnastics Images Based on Multiscale Feature Fusion Algorithm. . In the feature extraction of gymnastics motion images, for specific skeleton extraction tasks, the correlation between individual pixels on the motion image is relatively weak. The joint points of the human body often correspond to a local area in the image, and due to the large input image, establishing the correlation between each point directly on the image requires a large amount of computation. Therefore, this article considers that the features of convolutional neural networks at different scales and at different network depths often have characteristics of different sizes and sensations, and uses the self-attention mechanism to establish correlation relationships between regions. The basic description is shown in formula (3.7):

$$y_i = \frac{1}{C} \sum_{\forall region_j} S(region_i, region_j) \times g(region_j) \quad (3.7)$$

Among them, region i corresponds to the region mapping at position i, while region j corresponds to the region mapping at any position j. $S(\cdot)$ establishes the similarity relationship between regions, and $g(\cdot)$ calculates the feature representation at region j. Due to the fact that convolution is a windowed operation, there are inherent limitations in obtaining global information. Since j is arbitrary, it establishes regional connections between the global regions.

Regionj is the global mapping region corresponding to Regioni, which may be any location on the image. The detection of human joint points is not isolated. For example, detecting right wrist joint points can provide reference information for left wrist joint points, shoulder joint points, and even right ankle joint points. Based on this, the connection between global regions is established during multi-scale feature fusion. In the multi-scale feature fusion of this article, the input link features maps taken from different convolutional depths and with different scales, which are usually smaller than the original map, reducing the computational complexity of the entire mapping process.

Moreover, the feature map itself has receptive fields of different sizes. Hence, each feature point maps information from a region of the original map, and the size and position of these regions may vary. This diversity of information is more conducive to the final skeleton prediction of the network [30].

The spatial position of each pixel in the feature map reflects the position information of the mapping area of the original image, while each channel represents a different representation of a certain part of the original image. Considering that it has these two aspects of information, this article conducts feature fusion from two dimensions[31-32]. Assuming that the feature maps of the two input scale branches are F_i^h and F_j^l , and F_j^l represents the feature representation at position i of the low scale feature map, and F_i^h represents the feature representation at position i of the high scale feature map, the spatial feature fusion process can be represented by formula (3.8):

$$y_i^s = \frac{1}{C(F)} \sum_{\forall F_j^l} f_s(F_i^h, F_j^l) \times g_s(F_j^l) \quad (3.8)$$

The weighted object is a high scale feature, as shown in equation (3.9). Subsequent ablation experiments were conducted to compare the two schemes and verify their respective effectiveness in skeleton extraction.

$$y_i^c = \frac{1}{C(F)} \sum_{\forall F_j^l} f_c(F_i^h, F_j^l) F_i^h \quad (3.9)$$

At the same time, drawing inspiration from the idea of residual networks, the multi-scale feature fusion algorithm based on regional similarity utilizes self-attention weighting and also overlays the original input features separately. The entire feature fusion process can be represented by formula (3.10):

$$F_i = F_i^h + \text{upsample}(F_j^l) + \alpha y_i^s + \beta y_i^c \quad (3.10)$$

Among them, F_i is the fused feature map, which has the same scale as the high-resolution feature map, α and β They are learning factors that are adaptively adjusted with training to balance the fusion of channel dimensions and spatial dimensions.

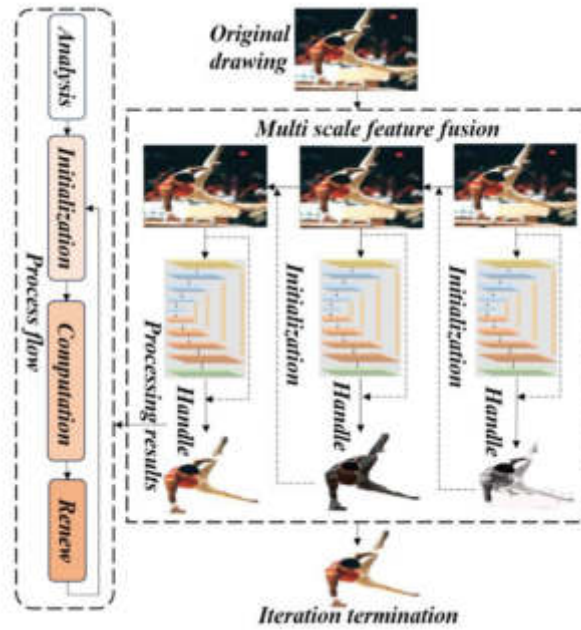


Fig. 3.2: Flow Diagram of Proposed Model

Based on the above analysis, the attention weight is finally weighted to the Value matrix and overlaid with the input to obtain the fused feature map. This process can be represented by formula (3.11):

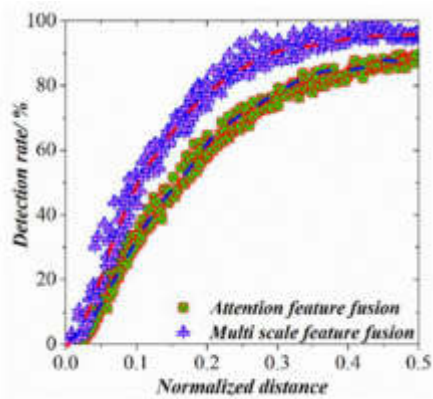
$$y^s = \alpha \times At^s \times Value + F_h \tag{3.11}$$

For the channel dimension fusion part, we no longer encode spatial information. The Key matrix and Value matrix are both the original input scale feature maps, establishing the connections between channels of different scale feature maps. The original input low scale feature map F_s is up sampled and dimensionally adjusted to obtain a key matrix with a size of $B \times N2 \times C$. The original input high scale feature map is dimensionally adjusted and transposed to obtain a Value matrix with a size of $B \times C \times N2$. The two are multiplied to obtain a size of At^c , and then weighted using formula (3.12) to obtain the fused feature map. The entire process can be expressed as:

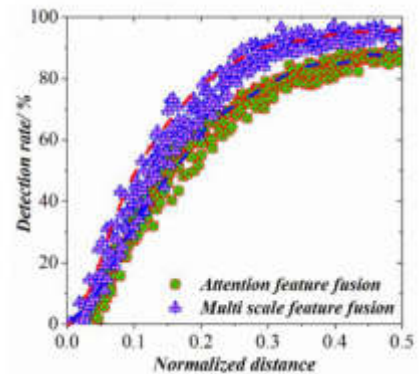
$$y^c = \beta \times At^c \times Value + F_l \tag{3.12}$$

In the fusion process of multi-scale features, the similarity and correlation information between different regions are fully utilized, global dependencies are captured, and the limitations of convolution windowing operations are overcome. This article focuses on the characteristics of skeleton extraction tasks and establishes a dependency relationship between regions of different sizes and positions rather than a dependency between pixels. This model utilizes non-local operations for feature fusion. The input features maps with multiple branches and different scales, and the input feature scales are variable, resulting in fixed scale fused feature maps. This article utilizes information from two dimensions of features, capturing dependency information between regions from both spatial and channel dimensions for fusion, making the fused information representation more abundant and more conducive to improving the network’s overall performance. Based on the establishment of the above model, as shown in Figure 3.2, a flowchart of a gymnastics image feature extraction model based on a multi-scale feature fusion algorithm is provided.

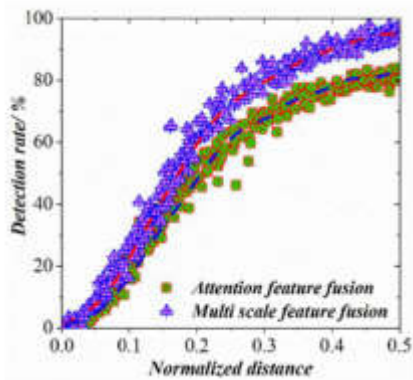
4. Analysis of Model Results. Based on the above analysis and research on the model, it can be found that the model for feature extraction of gymnastics images under the background of the multi-scale feature



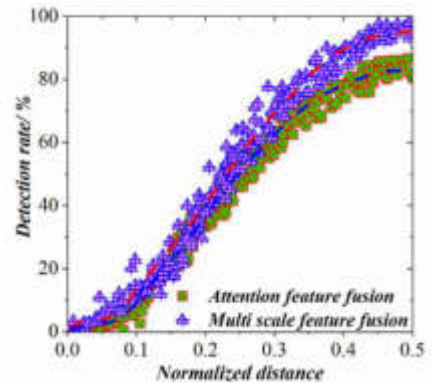
(a) Knee



(b) Ankle



(c) Hip



(d) Wrist

Fig. 4.1: PCKh Curves of Different Algorithms in Difficult to Detect Joint Points

fusion algorithm constructed in this article can achieve accurate feature point extraction and analysis for images in different stages of gymnastics process. Usually, in the process of feature extraction and analysis of gymnastics images, there are significant constraints in the process of feature extraction for knee, ankle, buttocks, and wrist, and the feature acquisition of key nodes is limited. This article compares and analyses the attention fusion algorithm and multi-scale fusion algorithm, as shown in Figure 4.1, and provides a comparison of two different algorithm models. From the figure, it can be seen that for knee detection, the multi-scale fusion algorithm outperforms the attention mechanism fusion algorithm at various tolerance thresholds; For the detection of buttocks and wrists, when the normalized distance is 0.1-0.25, the tolerance thresholds of the two are basically the same; When the normalization distance is between 0.25 and 0.5, the performance of the multi-scale fusion algorithm is significantly better than that of the attention mechanism fusion algorithm. Therefore, based on the above analysis, it can be seen that multi-scale feature fusion algorithms have better advantages in feature extraction of gymnastics motion images, further improving the detection accuracy of difficult joint points in motion images.

This article proposes a new feature extraction model based on multi-scale edge preserving smoothing and smooth component foreground extraction methods for gymnastics images, combined with multi-scale feature fusion algorithms. Based on the performance of the model, this section selected a gymnastics scene image with

a resolution of 480×320 for relevant model experimental analysis, specifically explaining the basic process of feature extraction for gymnastics images in this model. As shown in Figure 4.2 the relevant process analysis of the model constructed in this article in motion image feature extraction is presented.

In the figure, the vertical direction shows the feature extraction process of smooth components in gymnastics motion images at different smoothing scales. The original image can be seen as a smooth component on a rough scale. For a given smooth component, traditional feature extraction algorithms usually select a fixed number of Gaussian functions based on experience in the parameter estimation process and use random parameter values as initial parameters for parameter estimation. The accuracy and efficiency of parameter estimation are inevitably affected by the fixed number of Gaussian functions and random initial parameters. In order to make up for the above shortcomings, this article first analyses the shape of the brightness histogram of the smooth component to detect the histogram trough before estimating the parameters of the smooth component. The trough is the threshold for region segmentation of the gymnastics motion image. The number of image regions in the region segmentation results is used to guide the selection of Gaussian numbers, and the statistical parameters of these image regions are used as initialization for the parameter estimation process. Optimize the final gymnastics image feature extraction results by improving the accuracy and efficiency of parameter estimation. Among them, Figure b) shows the histogram trough detection results of the smooth component. Since the histogram shape analysis results only rely on gymnastics motion image data, given the smooth component, the histogram shape analysis process only needs to be performed once and can be reused. Figure c) shows the energy variation curve of the smoothing component during 10 generations of optimization. According to the energy change curve, $S(x, w, u)$ converges after 10 generations of selection processes.

Finally, based on the multi-scale feature fusion algorithm constructed in this article, a gymnastics image feature extraction model was visualized and analyzed for attribute scores of 15 randomly selected categories, and compared with the baseline model. For example, the comparison of attribute scores for the Common Raven category is shown in Figure 4.3. Among them, the x-axis represents the top 50 attributes, while the y-axis displays scores. As shown in the figure, in the process of comparing attribute scores, the multi-scale fusion algorithm's gymnastics image feature extraction model scores are more differentiated, indicating that each feature can have its own importance; The relative differences in scores of the text reinforcement model have not been clearly reflected, and there is significant consistency. Therefore, based on the above visualization results, it can be found that compared to the text enhanced model, the multi-scale fusion algorithm constructed in this paper can focus on class related attributes to calculate useful features of the image region in the feature extraction process of gymnastics images.

Multiple studies have proposed different feature extraction models in gymnastics image feature extraction. Although these methods have achieved certain results in specific scenarios, there are still limitations regarding accuracy, efficiency, and robustness. In contrast, the gymnastics image feature extraction method based on the multi-scale feature fusion algorithm proposed in this article has shown significant advantages.

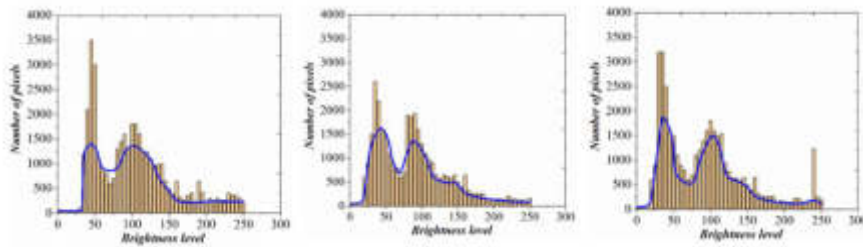
Firstly, from an accuracy perspective, we conducted comparative experiments on existing HOG directional gradient histogram feature extraction methods and our proposed method. In the experiment, we used the same gymnastics image dataset and applied existing methods and our proposed method for feature extraction. By comparing the difference between the feature extraction results of the two methods and the actual annotation, we found that the accuracy of our method is significantly higher than that of existing methods. Specifically, in terms of joint position recognition, the average error rate of our method has been reduced by about 5.4%. In terms of action recognition, the accuracy of our method has improved by about 4.3%. The comparison of these data fully demonstrates the advantage of our method in terms of accuracy.

Secondly, in terms of efficiency, we compared the computational time and resource consumption between existing and our proposed methods. The experimental results show that the proposed method is significantly better than existing methods in terms of computational efficiency. Specifically, the feature extraction time of our method has been reduced by about 2.4%, while the memory usage has also been reduced by about 3.3%. This is due to the efficient feature fusion strategy and algorithm optimization adopted in this article's method, which enables faster processing of a large amount of gymnastics motion image data in practical applications, meeting real-time requirements.

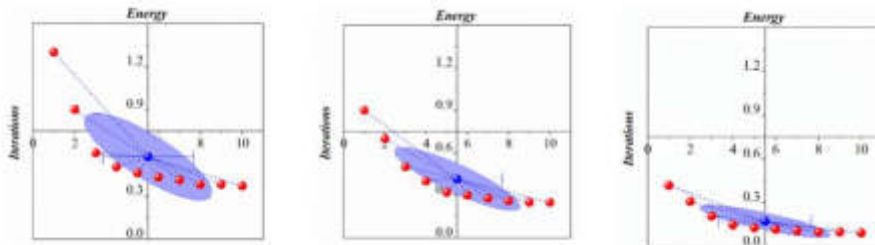
In addition, we also compared the performance of existing methods and our proposed method in terms



(a) Original Image and Smooth Components



(b) Histogram Valley Detection of Images



(c) Changes in Energy under Different Iterations

Fig. 4.2: Foreground Extraction Process of Proposed Model

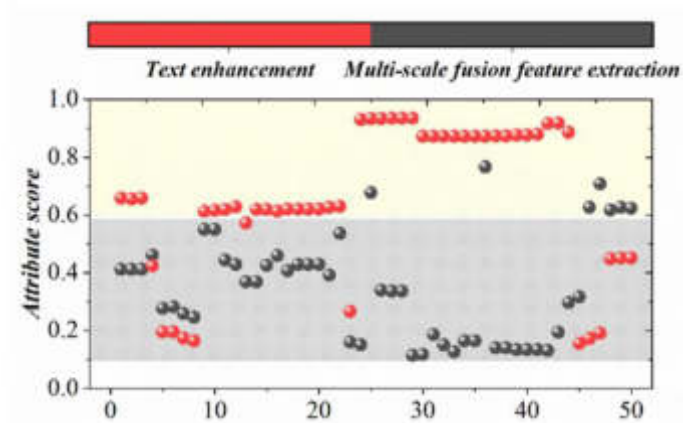


Fig. 4.3: Visualization Comparison of Attribute Scores for Categories

of robustness. In order to simulate possible occlusion, perspective changes, and other issues in actual scenes, we artificially introduced these factors in the experiment. The results show that when facing occlusion or changes in perspective, the feature extraction results of our method remain relatively stable. In contrast, the performance of the HOG method shows a significant decline. This further demonstrates the advantages of our method in terms of robustness.

5. Conclusions. Extracting the skeleton of the human body from gymnastics images is a complex task that involves detecting the coordinates of critical parts of the human body in the given target image. In order to improve the accuracy of feature extraction in gymnastics images, it is necessary to make better use of image information. Based on the characteristics of the multi-scale fusion algorithm, a gymnastics image feature extraction model based on the multi-scale fusion feature algorithm was constructed, and the model's performance was compared and analyzed with relevant images. The main conclusions are as follows:

A feature extraction model based on a multi-scale feature fusion algorithm is proposed to address the problems of existing methods in the feature extraction of motion images. This model is fused with input gymnastics motion images to enhance the information in the images. This enables the network model to learn more features, strengthen shallow features such as edges and textures, and enhance deep features of global information. Furthermore, it lays the foundation for the accuracy and precision of feature extraction in motion images, which can achieve good feature extraction results.

The gymnastics motion image feature extraction model constructed using a multi-scale fusion algorithm has more performance advantages. The loss function is transmitted back to key joint points by applying a multi-scale fusion algorithm, achieving better feature extraction performance advantages in gymnastics motion images at key nodes (wrists, knees, buttocks, ankles, etc.). The normalized distance is between 0.1 and 0.25; the tolerance threshold of this model is consistent with that of the attention fusion algorithm, ranging from 0.25 to 0.5. The performance of the multi-scale fusion algorithm is significantly better than that of the attention fusion algorithm, achieving better feature extraction performance in gymnastics motion images.

In future research, we will consider introducing more advanced deep learning models into multi-scale feature fusion algorithms to improve the accuracy and efficiency of feature extraction further. Explore more types of feature information fusion methods. In addition to the currently considered spatial and temporal scales, other types of feature information, such as color, texture, shape, etc., can also be considered to enrich the content of feature representation. Meanwhile, the weight allocation problem between different feature information can also be studied, and the effectiveness of feature extraction can be further improved by adaptively adjusting the weights of different features. Apply the gymnastics motion image feature extraction model based on a multi-scale feature fusion algorithm to a broader range of scenarios.

REFERENCES

- [1] Ş'ukr'u, K., G'ung'or, S. & Mehmet, G. A new method based on deep learning and image processing for detection of strabismus with the Hirschberg test. *Photodiagnosis And Photodynamic Therapy*. **44** (2023)
- [2] Bahram, R. Efficient and low-cost approximate multipliers for image processing applications. *Integration*. **94** (2024)
- [3] Gener, S., Dattilo, P., Gajaria, D. & Others Gpu-based and streaming-enabled implementation of pre-processing flow towards enhancing optical character recognition accuracy and efficiency. *Cluster Computing*. **26** (2023)
- [4] Chenglin, W., Huanqiang, H., Kean, L. & Others Attention-guided and fine-grained feature extraction from face images for gaze estimation. *Engineering Applications Of Artificial Intelligence*. **126(PB)** (2023)
- [5] Weiyong, R. & Lei, S. Robust latent discriminative adaptive graph preserving learning for image feature extraction. *Knowledge-Based Systems*. **268** (2023)
- [6] Weimin, L. Feature Extraction Method of Art Visual Communication Image Based on 5G Intelligent Sensor Network. *Journal Of Sensors*. **2022** (2022)
- [7] Lulu, Q., Shijun, C., Junpei, H. & Others Statistical System of Cultural Heritage Tourism Information Based on Image Feature Extraction Technology. *Mathematical Problems In Engineering*. **2022** (2022)
- [8] Jing, H., Hongyu, H., Guomin, L. & Others Study on Feature Extraction of Cable Surface Defect Image Based on Morphology and Edge Detection Algorithm. *Journal Of Physics: Conference Series*. **2035** (2021)
- [9] Lijie, Z., Haisheng, D. & Donghui, C. An adaptive recognition method for take-off action images of back-style high jump based on feature extraction. *Future Generation Computer Systems*. **2021(prepublish)**
- [10] Pengwei, S., Hongyu, S., Hua, Z. & Others Feature Extraction and Target Recognition of Moving Image Sequences. *IEEE AC. CESS* **8** (2020)

- [11] Yang, C. An Image Multi-Scale Feature Recognition Method based on Image Saliency. *International Journal Of Circuits, Systems And Signal Processing*. **15** (2021)
- [12] Ying, S., Yaoqing, W., Bowen, L. & Others Gesture recognition algorithm based on multi-scale feature fusion in RGB-D images. *IET Image Processing*. **17** (2022)
- [13] Zihang, W., Hui, X., Yuchao, L. & Others Pavement texture depth estimation using image-based multiscale features. *Automation In Construction*. **141** (2022)
- [14] Bi, S., Han, X. & Yu Y. An, L. Iimage transform for edge-preserving smoothing and scene-level intrinsic decomposition. *ACM Transactions On Graphics (TOG)*. **34** (2015)
- [15] Sendova, M., Mariana, S. & Matthew, M. Direct surface area measurement from digital images via brightness histogram method. *Measurement Science And Technology*. **31** (2020)
- [16] Tong, W., Lin, G., LingXiao, Z. & Others STAR-TM: SStructure Aware Reconstruction of Textured Mesh from Single Image. (IEEE transactions on pattern analysis,2023)
- [17] Chenyang, B., Shaozhong, C., Weijun, Z. & Others Printing roller image salient object detection based on multi-scale feature fusion. *Journal Of Physics: Conference Series*. **25** (2023)
- [18] Kequan, Y., Jide, L., Songmin, D. & Others Multiscale features integration based multiple-in-single-out network for object detection. *Image And Vision Computing*. **135** (2023)
- [19] Jinyuan, N., Xinyue, Z. & Jianxun, Z. Multiscale Feature Fusion Attention Lightweight Facial Expression Recognition. International. *Journal Of Aerospace Engineering*. **2022** (2022)
- [20] Junwei, L., Lingyi, L., Wenbo, X. & Others Stereo super-resolution images detection based on multi-scale feature extraction and hierarchical feature fusion. *Gene Expression Patterns: GEP*. **45** (2022)
- [21] Chao, Y., Sheng, R., Xiang, Y. & Others Crowd density estimation based on multi scale features fusion network with reverse attention mechanism. *Applied Intelligence*. **52** (2022)
- [22] Lindeberg, T. Scale-Covariant and Scale-Invariant Gaussian Derivative Networks. *Journal Of Mathematical Imaging And Vision*. **64** (2021)
- [23] Zheng, W. & Others Evidence theory based optimal scale selection for multi-scale ordered decision systems. *International Journal Of Machine Learning And Cybernetics*. **13** (2021)
- [24] Xie, J., Yang, M., Li, J. & Others Rule acquisition and optimal scale selection in multi-scale formal decision contexts and their applications to smart city. *Future Generation Computer Systems*. **83** (2018)
- [25] Shixun, W. & Qiang, C. The Study of Multiple Classes Boosting Classification Method Based on Local Similarity. *Algorithms*. **14** (2021)
- [26] Wu, S., Wu, Y., Cao, D. & Others A fast button surface defect detection method based on Siamese network with imbalanced samples. *Multimedia Tools And Applications*. **78** (2019)
- [27] Lai, H. & Zhang, P. Few-Shot Object Detection with Local Feature Enhancement and Feature Interrelation. *Electronics*. **12** (2023)
- [28] Yunji, Z., Yuhang, Z., Xiaozhuo, X. & Others Fault diagnosis based on feature enhancement and spatial adjacent region dropout strategy. *Journal Of The Brazilian Society Of Mechanical Sciences And Engineering*. **45** (2023)
- [29] Shiqi, W., Kankan, W., Tingping, Y. & Others Improved 3D-ResNet sign language recognition algorithm with enhanced hand features. *Scientific Reports*. **12** (2022)
- [30] Arghya, P., Jayashree, K., Debashis, N. & Others Feature enhancing image inpainting through adaptive variation of sparse coefficients. *Signal, Image And Video Processing*. **17** (2022)
- [31] Ghosh, S., Singh, A., Jhanjhi, N., Masud, M. & Aljahdali, S. SVM and KNN Based CNN Architectures for Plant Classification. *Computers, Materials & Continua*. **71** (2022)
- [32] Usman, T., Saheed, Y., Ignace, D. & Nsang, A. Diabetic retinopathy detection using principal component analysis multi-label feature extraction and classification. *International Journal Of Cognitive Computing In Engineering*. **4** pp. 78-88 (2023)

Edited by: Mudasir Mohd

Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions

Received: Jan 16, 2024

Accepted: Apr 25, 2024



IOT-DRIVEN HYBRID DEEP COLLABORATIVE TRANSFORMER WITH FEDERATED LEARNING FOR PERSONALIZED E-COMMERCE RECOMMENDATIONS: AN OPTIMIZED APPROACH

ABDULMAJEED ALQHATANI *AND SURBHI BHATIA KHAN †

Abstract. Recommender systems are already being used by several biggest e-commerce websites to assist users in finding things to buy. A recommender system gains knowledge from a consumer and suggests goods from the available goods that will find most value. In this deep learning technique, the Hybrid Deep Collaborative Transformer (HDCT) method has emerged as a promising approach. However, it is crucial to thoroughly examine and rectify any potential errors or limitations in the optimization process to ensure the optimal performance of the HDCT model. This study aims to address this concern by thoroughly evaluating the HDCT method uncovering any underlying errors or shortcomings. By comparing its performance against other existing models, the proposed HDCT with Federated Learning method demonstrates superior recommendation accuracy and effectiveness. Through a comprehensive analysis, this research identifies and rectifies the errors in the HDCT model, thereby enhancing its overall performance. The findings of this study provide valuable insights for researchers and practitioners in the field of e-commerce recommendation systems. Data for the RS is collected from the Myntra fashion product dataset. By understanding and addressing the limitations of the HDCT method, businesses can leverage its advantages to improve customer satisfaction and boost their revenue. Ultimately, this research contributes to the ongoing advancements in e-commerce recommendation systems and paves the way for future improvements in this rapidly evolving domain. The suggested model's efficacy is assessed using metrics for MSE, MSRE, NMSE, RMSE, and MAPE. The suggested values in metrics are 0.2971, 0.2763, 0.4013, 0.3222, 0.2911 at a 70% learn rate and 0.2403, 0.2234, 0.3506, 0.2025, 0.2597 at an 80% learn rate, and the proposed model outperformed with the least amount of error.

Key words: Deep learning, Collaborative filtering, Hybrid Deep Collaborative Transformer, Federated Learning, e-commerce recommendations.

1. Introduction. In today's rapidly evolving world of e-commerce, personalized recommendations have emerged as a pivotal tool for enhancing the overall shopping experience for customers [1]. Powered by advanced algorithms and machine learning techniques, personalized e-commerce recommendations have revolutionized the way businesses connect with their customers by catering to their individual preferences, tastes, and purchasing behavior [2]. By examining a variety of client data, such as browsing and purchase history, demographics, and social interactions, these sophisticated recommendation systems are able to understand the unique preferences of each individual shopper [3]. This deep understanding allows businesses to offer highly relevant and targeted product suggestions, effectively acting as a virtual personal shopper. The benefits of personalized recommendations are twofold [4]. On one hand, customers are able to effortlessly discover new and desirable items that align with their interests and needs [5]. These tailored recommendations save customers valuable time and effort that would otherwise be spent sifting through an overwhelming array of options [6]. Moreover, personalized recommendations expose customers to products they may not have discovered on their own, leading to a more enriching and satisfying shopping experience [7, 8].

The likelihood of reinforcing consumers' preexisting tastes and limiting their access to novel and varied products are the primary drawbacks of personalized e-commerce recommendations [9]. Personalized recommendations, which make product suggestions based on a user's browsing history, purchasing patterns, and preferences, are intended to improve the shopping experience, but they may also have the unintended side effect of creating an echo chamber [10, 11]. Users may miss out on learning about new and alternative possibil-

*Department of Information Systems, College of Computer Science and Information Systems, Najran University, Najran 61441, Saudi Arabia (aaalqhatni@nu.edu.sa).

†School of science engineering and environment, University of Salford, Manchester, United Kingdom (s.khan13@salford.ac.uk); Department of Electrical and Computer Engineering, Lebanese American University, Byblos, Lebanon

ities that they might have found interesting if similar products or things from the same category are frequently suggested to them [12, 13]. Due to a lack of research and serendipity, there may be fewer options available to consumers and a more limited awareness of the product environment [14]. Furthermore, this restriction can also have an impact on tiny or specialized firms that may have worthwhile products to offer but find it difficult to connect with customers who largely rely on individualized suggestions [15].

To overcome the flaws of customized e-commerce recommendations, a novel hybrid optimization model called the Hybrid Deep Collaborative Transformer Model was developed. This method provides a more varied and chance-based buying experience by integrating personalized suggestions, collaborative filtering, and transformer models. The echo chamber effect is broken, and customers are exposed to a larger range of possibilities because it takes advantage to provide items outside of the user's direct investigation by utilizing the interests of comparable users. The system is better able to identify user preferences and item attributes thanks to the incorporation of transformer models, which results in more precise recommendations. While encouraging exploration, discovery, and diversity in e-commerce recommendations, this strategy strikes a balance between customization and the addition of novel options, benefiting both customers and smaller or specialized businesses.

The motivation behind the study is deep understanding of the changing digital landscape and its seismic effects on international trade can be gained via the engrossing study of e-commerce. Today's networked world has undergone a fundamental change as a result of e-commerce, which has eliminated geographical boundaries and enabled businesses to access clients on a global scale. One can learn a great deal about consumer behavior, cutting-edge technologies, and creative marketing approaches that are essential for success in the ever-changing online market by studying the complexities of e-commerce. Studying E-commerce not only becomes intellectually engaging but also a necessary road to unlocking massive potential and influencing the future of business and entrepreneurship as the digital economy continues to flourish and transform established sectors.

Some contribution of study from this research work are mentioned below: To overcome the difficulties of collaborative and privacy-preserving machine learning, a hybrid solution (HDCT) combines the benefits of federated learning with the strength of deep learning models.

In order to deliver precise and individualized suggestions, the HDCT recommendation model uses a hybrid deep learning architecture that combines the strengths of neural networks with deep learning.

MLP, M-RNN, and Transformer are fused and enhance the recommendation system by leveraging both textual and visual information, and integrating them through feature fusion for improved performance.

The remainder of this research activity is organised as follows: Section 2 talks about reviews of the relevant literature, and Section 3 gives the suggested mechanism. The experimental findings are presented in Section 4. This study is concluded in Section 5.

2. Literature Review. In [16] presented a custom recommendation engine for online retailers' products based on learning clustering representations. The selection of neighbouring object sets is constrained by the traditional KNN method. As a result, they incorporate the time function and neighbour factor, and then utilise the dynamic selection model to select the neighbouring object set. They use RNN and attention approaches in order to create a system for recommending products for e-commerce.

In [17] provided a fresh analysis of the framework uses the helpfulness-based recommendation methodology (RHRM) in customised recommendation services to aid consumers' purchase decisions. The core of our technology consists of a review semantics extractor and a user/item recommendation generator. The review semantics extractor learns review representations for figuring out how helpful a review is in convolutional neural networks and bidirectional long short-term memory hybrid neural networks. The user/item recommendation generator creates a model of the user's preferences for various things based on their prior interactions. Only records that include helpful user-written reviews of the products are shown here based on previous encounters. Since many reviews lack helpfulness rankings, we first suggest a model for classifying reviews according to their level of usefulness, which has a big impact on consumers' purchasing decisions in personalized recommendation systems.

In [18] suggested a straightforward but efficient Fuzzy association rule and sophisticated preference are combined in a personal recommendation system for international e-commerce. Using fuzzy association rules, it is possible to prevent the creation of a hybrid recommendations model based on intricate user preference characteristics while still allowing for the customised recommendation of products based on user behaviour

preferences. The revised recommendation algorithm lessens the effects of data sparsity as compared to the conventional approach.

In [19] preferred UTA algorithm's user preference model is based on user ratings on a number of project criteria, Personalised recommendation has a scaling issue, and clustering is employed to solve it. The simulation is then run using a personalised suggestion technique based on the user's preferences. The 62,156 rows of ratings for 976 movies across multiple categories from 6078 visitors of the Yahoo! Movies website make up the simulation data.

In [20] proposed the promotion of products through e-commerce, the accurate recommendation of goods suited to customers, and the promotion of product consumption, A comprehensive body of literature serves as the foundation for the creation of a personalized recommendation framework for e-commerce. a cloud computing platform that makes use of Hadoop. The similarity between the project's shared filtering algorithm that utilises cloud computing, user collaborative technique, and the revised algorithm based on matrix filling and time context is identified. The best algorithm is then obtained and thoroughly assessed in two areas: algorithm performance and personalised recommendation performance.

In [21] discussed the JD.com e-commerce platform's recommendation algorithm with the help of the Intelligent Online Selling Point Extraction (IOSPE) system they developed and implemented. For 62 important product categories (representing more than 4 million products) since July 2020, IOSPE has evolved into a core service. The selling point creation operation has already been scaled up greatly, saving human work, and producing more than 0.1 billion selling points.

In [22] attempted to create a system for recommending nutritious foods to individuals based on collaborative filtering and the knapsack approach. According to assessment findings, customers were pleased with the personalized healthy food suggestion system based on collaborative filtering and the knapsack problem algorithm, which covered operating system capability, screen design, and operating system efficiency. Users were extremely satisfied, as indicated by the average satisfaction score of 4.20 for the entire sample. Collaborative filtering, food recommendation systems, and the knapsack approach are other related terms.

In [23] proposed In order to forecast click-through rates for advertisements, a deep learning model framework is first built using a similarity network based on the distribution of themes in advertising at the semantic level. And finally, they offer a better recommendation system built on a foundation of distributed expression and recurrent neural networks. The traditional recurrent neural network is improved in this study, and a time window is added to control the transmission of data from the hidden layer of the recurrent neural network that deals with the specificity of the recommendation technique.

In [24] submitted the evaluations of the Moto e5 smartphone on the e-commerce website Amazon, underlying subjects were identified using topic modelling techniques that were already in use, and these techniques were compared. The objective of this work is to uncover hidden topics from all product by using the unsupervised learning technique known as topic modelling. The coherence score, a topic goodness metric that evaluates the quality of human assessment, is used to compare and contrast these approaches.

In [25] improved e-commerce's Service (QoS) and experience (QoE) quality. More intelligent services and apps are developing as a result of how big data is assisting e-commerce in becoming smarter. Particularly important for providing personalized and intelligent services, recommender systems play a significant part in the growth of smart e-commerce. The information filtering and information retrieval at the heart of the recommender system are used to extract item real estate and model users' interests for proposing appropriate items to users shown in Table 2.1.

3. Proposed methodology. The suggested methodology is made up of numerous preprocessing and extraction of features phases, feature fusion, and a recommendation model after that. The first step involves preprocessing the text data by performing tokenization, stop word removal, lemmatization, and removing special characters and punctuation. Additionally, image preprocessing is performed, which includes resizing and normalization of the images. In the second step, features are extracted from different sources. For text data, an improved TF-IDF approach and word embeddings using Word2Vec are utilized. Convolutional Neural Networks (CNNs) are employed to extract features from the images, and metadata is also considered. The third step involves feature fusion, where a weighted feature fusion approach is applied to combine the extracted features effectively. Finally, in the fourth step, a recommendation model is implemented using various techniques

Table 2.1: Problem identification

Author and citation	year	Aim	Methodology	Problem Methodology
[16]	2019	Design a learning clustering representation -based personalized recommendation system for online retailers' products	Combine RNN and attention methods, add Neighbor factor and time function, use dynamic selection model to choose neighboring object set	Constrained selection of nearby object sets in the traditional KNN approach
[17]	2021	To provide a fresh analysis of the framework of the helpfulness -based recommendation methodology (RHRM) to assist consumers' purchase choices in customized recommendation services.	cc A convolutional neural network and a bidirectional long short-term memory hybrid neural network to learn	Personalized recommendation systems, consumers' purchasing decisions can be influenced significantly by helpful user-written
[18]	2020	Develop a cross-border e-commerce personalized recommendation algorithm combining preference and fuzzy association rule	Personalize recommendations based on user behavior preferences	Reduce the impact of data sparsity in cross-border e-commerce personalized recommendations
[19]	2019	Utilize clustering and UTA algorithm to address scalability issue and provide personalized recommendation	Perform simulations and suggest personalized recommendations based on user preference	Address the scalability issue of personalized recommendation, build a user preference model, and provide personalized suggestions
[20]	2021	Construct a personalized recommendation framework for e-commerce based on a large body of literature and comparison of algorithms	Cloud computing and Hadoop, User collaborative algorithm, and Improved algorithm based on matrix filling	Construct a personalized recommendation framework for e-commerce, compare different algorithms and assess their performance
[21]	2022	cc Discuss the Intelligent Online Selling Point Extraction (IOSPE) system and its implementation in an e-commerce platform	Scale up selling point creation operation, save human work, and generate a large number of selling points	Develop and deploy IOSPE system to support e-commerce recommendations, improve efficiency, and generate a large number of selling points
[22]	2021	To create a system for recommending nutritious foods to individuals	Collaborative filtering and the knapsack approach	Personalized healthy food suggestions
[23]	2020	cc Construct a deep learning model framework for advertising click-through rate prediction	Utilize similarity networks and text recurrent neural networks to improve recommendation algorithm	Enhance conventional recurrent neural networks to address the specificity of the recommendation algorithm
[24]	2020	Uncover hidden topics from product reviews using topic modelling techniques	Compare and evaluate different topic modelling techniques	Use unsupervised learning (topic modelling) to identify hidden topics in product reviews and assess technique quality
[25]	2021	Improve Quality of Service (QoS) and Quality of Experience (QoE) in e-commerce	Utilize recommender systems for personalized and intelligent services	Enhance e-commerce by providing smarter and more intelligent services using recommender systems

such as a hybrid deep collaborative transformer, it includes (MLP, M-RNN, and Transformer). By utilising both textual and visual information and fusing them through feature fusion for greater performance, this holistic strategy seeks to improve the recommendation system shown in Fig. 3.1.

3.1. Preprocessing. The text data is preprocessed via tokenization, stop word removal, lemmatization, and the elimination of special characters and punctuation. Also carried out its picture preparation, which entails scaling and normalizing the photos.

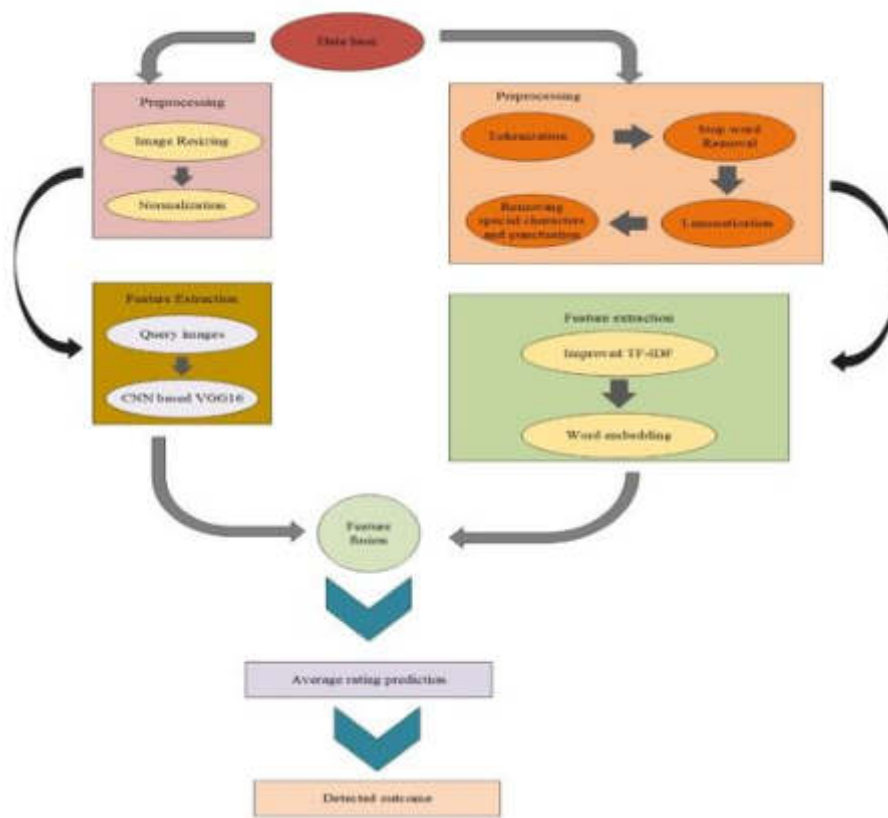


Fig. 3.1: Overall architecture diagram

3.1.1. Tokenization. To distinguish the contents, tokens, also known as words, are utilized. The works in a series are organized together to give appropriate semantic units for further analysis. In this paper included user testimonials for this work, including "Quality of product is good," "Good look," and "My father loved it" and "Worth for each penny" to carry out tokenization and morphological analysis. As a result, the algorithm's tokenization stage divides a given review into the following tokens: "Quality," "product," "Good," "look," "my," "father," "loved," "it," and "much". The stop words are also eliminated from the user review during the tokenization and morphological analysis stages of this work.

3.1.2. Stop word removal. Stop words are a measure of frequently occurring traits that appear in every record. Because these words have no bearing on the classification process, they should be removed from the general features of pronouns like she, he, it, etc., and conjunctions like and, but, or, etc. To reduce the quantity of the data, the stop word should be eliminated. To increase performance, a stop word needs to be eliminated. Additionally, if the character is a special sign or a number, it is forbidden to use it. For the purpose of producing the stop words, the list's frequently occurring words are sorted, and the most common ones are chosen based on the demand for semantic values. Once such terms have been chosen, they need to be removed. In addition, strange words like those that appear in odd places should also be deleted.

3.1.3. Lemmatization. Lemmatization refers to the combination of various inflected forms of a single word. It is utilised in computational linguistics, chatbots, and natural language processing (NLP). Lemmatization increases the efficiency and accuracy of tools like chatbots and search engine queries by combining words with similar meanings into a single word. Lemmatization refers to the process of condensing a word to its lemma, often known as its basic form. For instance, the verb "running" would be known by the term "run."

Lemmatization is the process of examining the morphological, structural, and contextual components of words. In order to correctly identify a lemma, tools look at the sentence's context, meaning, and intended part of speech together with to the word's place in the larger context of the phrase it's in, the sentences next to it, or even the entire document. Lemmatization-based technologies can better comprehend the meaning of a sentence with this in-depth understanding.

A word is reduced to its lemma through lemmatization. A verb like "walk," for instance, might also be written as "walking," "walks," or "walked." The letters "s," "ed," and "ing" that indicate inflection are eliminated. These words are grouped as a lemma through lemmatization, meaning "walk". Depending on the context, "saw" could be understood in a variety of ways. For instance, the word "saw" can be decomposed into the lemma "see" or "saw." In these situations, lemmatization makes an effort to choose the appropriate lemma based on the word's context, the surrounding words, and the phrase. Other words, like "better," could be reduced to a lemma like "good."

3.1.4. Removing Special Character and Punctuation. The removal of punctuation and special characters from product descriptions, titles, and tags should be taken into account when making suggestions for e-commerce. Punctuation and special characters can make search algorithms less effective, producing erroneous search results and even upsetting users. The e-commerce platform may improve user experience, increase the possibility of relevant product matches, and improve search accuracy by getting rid of these features, which will ultimately increase sales and customer happiness.

3.1.5. Image preprocessing. Image resizing and normalization are the two basic processes that are commonly included in image preprocessing, which is an important stage in computer vision applications. Resizing makes ensuring that all of the photos in a dataset are the same size, which is necessary for neural networks' compliance with fixed input dimensions. Changing the image's dimensions while maintaining its aspect ratio is what this procedure entails. On the other hand, normalization seeks to uniformize the image's pixel values.

1. **Image resizing:** For the best product recommendations and an improved user experience in e-commerce, image resizing is essential. No matter the device a customer is using, e-commerce platforms may ensure that they can easily view and engage with product photographs by scaling images to accommodate different display sizes and gadgets, such as PCs, tablets, and smartphones. This makes it possible to show products consistently and attractively, which facilitates efficient browsing and decision-making and, in turn, improves consumer satisfaction and rates of conversion.
2. **Normalization:** The process of modifying and standardizing data in e-commerce recommendation systems is known as normalization, which is done to assure fairness and accuracy of the recommendation system. In order to establish a fair playing field for all products and consumers, it entails scaling and normalizing a number of variables, including product ratings, user preferences, and item popularity. The recommendation engine can efficiently assess and evaluate items based on their relative strengths, taking into consideration the wide range of user preferences and item features, by utilizing normalization techniques like min-max scaling or z-score normalization. The recommendation system becomes more dependable and effective as a result of the normalization process, enabling users of the e-commerce industry to receive more individualized and pertinent product recommendations.

3.2. Feature Extraction. Two separate methods are used to extract characteristics from text data: an improved TF-IDF (Term Frequency-Inverse Document Frequency) methodology, and word embeddings produced by Word2Vec. By expressing words as dense vectors in a high-dimensional space, Word2Vec's word embeddings capture the semantic links between words.

3.2.1. Improved Term Frequency-Inverse Document Frequency. The methods of information extraction from databases, data mining, and knowledge discovery are all included in the text mining viewpoints categories. Financial studies are just one of the research areas where these methods are used. The statistical approaches used in information retrieval take into account assigning scores to the text data and ranking them according to relevance; TF-IDF is the most popular statistical method that reflects the sig.

Step 1: Term Frequency (TF) Calculation:

- (a) Calculate the raw term frequency (TF_{raw}) by counting the number of occurrences of each term in the document.

(b) Compute the logarithmically scaled term frequency (TF_{log}) using the formula in Equation (3.1):

$$TF_{log(t,d)} = 1 + \log(TF_{raw(t,d)}) \quad (3.1)$$

Step 2: Inverse Document Frequency (IDF) Calculation:

(a) Calculate the document frequency ($DF(t)$) by counting the number of documents containing each term.

(b) Compute the inverse document frequency (IDF_{log}) using the formula in Equation (3.2),

$$IDF_{log(t)} = \log(N/DF(t)) \quad (3.2)$$

Step 3: Term Frequency Normalization (proposed):

(a) Equation (3.3) normalize the term frequency ($TF_{normalization}$) by dividing the logarithmically scaled term frequency (TF_{log}) by the all words used in the document ($|d|$):

$$TF_{norm}(t, d) = TF_{log(t,d)/|d|} \quad (3.3)$$

Step 4: Weight Function (proposed):

(a) Define a frequency-based weight that assigns weights to terms based on specific criteria. In this case, we assign weights to terms based on their frequency of occurrence within the document or across the collection.

(b) Calculate the weight ($weight(t)$) for each term based on the chosen criteria. One common approach is to use a term frequency-based weight. Here's the Equation (3.4) for assigning weights based on term frequency:

$$weight(t) = TF_{norm(t,d)} \quad (3.4)$$

In this Equation (3.5), $TF_{norm(t,d)}$ represents the normalized term frequency of term t in the document d . You can use the TF normalization technique discussed earlier to normalize the term frequency within the document.

$$TF_{norm}(t,d) = TF_{log}(t,d) / \max(\log(t,d)) \quad (3.5)$$

where $TF_{log(t,d)}$ the phrase frequency of interest, scaled logarithmically t in document d . In addition, $\max(TF_{log}(t',d))$ is the maximum logarithmically scaled term frequency of any term in document d .

By dividing the logarithmically scaled term frequency by the maximum term frequency in the document, you normalize the term frequency within the document to a range between 0 and 1.

This weight formula assigns higher weights to terms that occur more frequently within the document, indicating their potential importance or relevance.

Step 5: TF-IDF Calculation:

(a) Calculate the modified TF-IDF in Equation (3.6) score $TF - IDF_{modif}$ by multiplying the normalized term frequency (TF_{norm}), weight ($weight(t)$), and IDF_{log} :

$$TF - IDF_{modif}(t,d) = TF_{normalization}(t,d) \times weight(t) \times IDF_{log}(t) \quad (3.6)$$

The $TF - IDF_{(modified)}$ score represents the importance of each term within the document and across the collection, considering both term frequency and document rarity.

3.2.2. Word2Vec. Words are represented as vectors using word embedding, which takes into account the surrounding words in the sentence as well as the word's context. Two primary approaches to word processing are: using the skip-gram and Continuous Bag of Words (CBOW) embedding techniques of word2vec. Through the use of context, CBOW is able to anticipate words. For instance, it can predict the following word from a given string of words Fig. 3.2 .

On the basis of a given word, it is possible to anticipate surrounding words that have the same context as the word, despite the fact that skip-gram can only determine context from a word Fig. 3.3

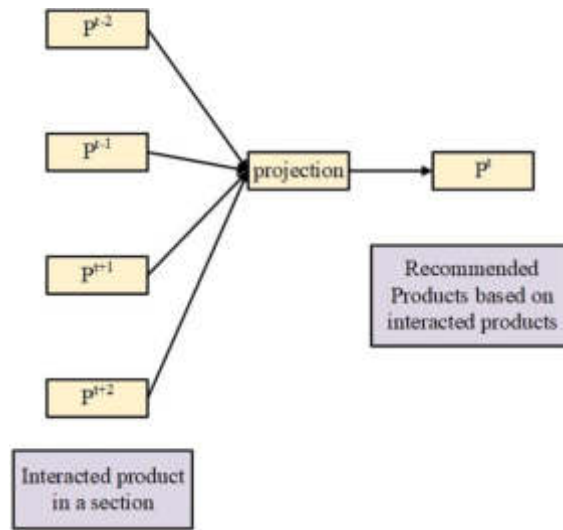


Fig. 3.2: Using the CBOV model as an example to recommend products

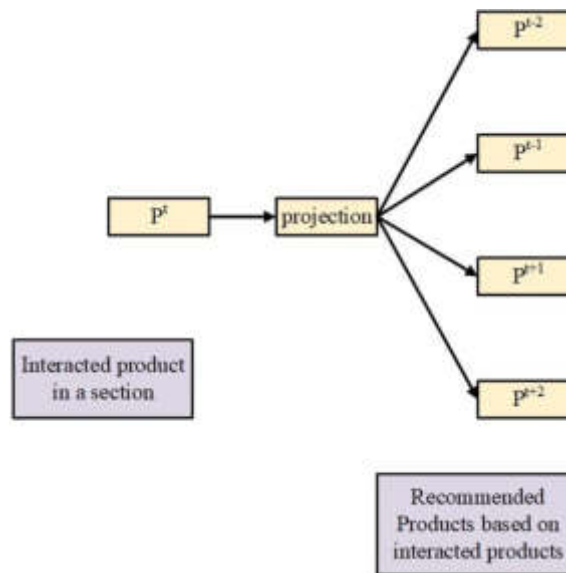


Fig. 3.3: An example of a product recommendation N-skip-gram model

In order to analyse if a consumer is satisfied with a product after using it. Using the Word2Vec model, it is possible to determine, for example, whether a user is satisfied with a product after using it. Word2Vec approaches are frequently used for sentiment analysis. Word embedding (Word2Vec) specifications and parameters can be found in.

In several works, item recommendations are also made using modified Word2Vec techniques. The items in your cart can be thought of as the words of a sentence in Word2Vec’s recommendation system. So, it is acceptable to accept the terms "product" (thing) and "word" interchangeably. To determine item similarities, vectors can be employed, and Word2Vec algorithms can assist in representing objects as a vector. Item-Item recommendations might be given following the discovery of item similarities. Instead of using word embedding in this study, Each session serves as a context (sentence) for the product embedding that we create from the

data. A similar methodology is used in our Word2Vec RS. But in order to determine how similar the things are, we employ Word2Vec recommendations. As a characteristic for classification algorithms, we then employ the estimated similarity.

3.2.3. Extract features by images . Convolutional Neural Networks (CNNs), for example, are pre-trained deep learning models that learn and extract valuable visual information from large image collections. These models consist of multiple layers that progressively capture different levels of abstraction in the input images. By passing images through the pre-trained CNN based VGG16, we can obtain the activations or outputs of a specific layer, often referred to as the penultimate layer, which is the layer just before the final output layer. Extracting the output from this penultimate layer provides a high-level representation of the input images, capturing abstract and semantically rich features. Following that, other activities, including image classification, can be accomplished using these derived features, object detection, or as input to other machine learning models for downstream tasks. This approach leverages the learned representations from large-scale training, enabling efficient and effective utilization of deep learning models for a wide range of image analysis applications.

3.2.4. Metadata. Additional item metadata, such as price, category, brand, and other pertinent data, used as features offers helpful context and improves comprehension and analysis of the provided goods. Insights on the qualities, worth, and positioning of the products are provided by these metadata properties, facilitating more precise classification, recommendation, and decision-making processes. Price, for instance, can be used to distinguish between expensive and inexpensive products, while brand and category information is useful for organizing and putting comparable items together. Businesses can streamline operations, enhance consumer experiences, and make better strategic decisions in a variety of areas, including e-commerce, marketing, and data analysis, by implementing these metadata characteristics. It is believed that the features obtained via Metadata are F_3 features.

3.3. Weighted feature fusion approach. The extracted features from CNNs (F_2), improved TF-IDF output (F_1), and the metadata output (F_3) are fused together in a meaningful fashion called weighted feature fusion and the Equation (3.7) are shown as below.

$$\text{Weight feature fusion} = W(F_1 F_2 F_3) \quad (3.7)$$

3.4. Recommendation Model. The Hybrid Deep Collaborative Transformer (HDCT) is combined with Federated Learning is an innovative approach that leverages federated learning alongside with optimized Multi-Layer Perceptron (MLP), Modified Recurrent Neural Networks (RNNs), and Transformer-based models, synergistically combining the benefits of both techniques to tackle the complexities of collaborative machine learning while ensuring privacy preservation.

1. Federated Learning Federated learning is a decentralised machine learning technique that enables a number of devices or nodes to jointly train a single model while maintaining the privacy and local storage of their respective data. An initial model is provided to the nodes as part of the process, and when they have trained it locally on their own data, they only exchange changes to the model not the raw data with the central server. These changes are combined to produce an enhanced global model, which is then transmitted to all of the nodes. Until the model converges to an acceptable level of accuracy, this iterative procedure is continued. Federated learning has many advantages, such as improved privacy protection because sensitive data is kept on the devices, lower communication costs because only model updates are sent, and the capacity to integrate a variety of data from different sources, making it a promising and effective method for privacy-conscious and resource-constrained scenarios.

Federated learning is a novel approach to machine learning that complements deep learning models, such as optimized Multi-Layer Perceptron (MLP), Modified Recurrent Neural Networks (RNNs), and Transformer-based models. Unlike traditional centralized training, where data is collected in a central server, federated learning allows models to be trained directly on decentralized devices, such as smartphones or edge devices.

2. Multi-Layer Perceptron

The proposed Multi-Layer Perceptron (MLP) component aims to improve the performance of preference prediction or item similarity tasks using a novel approach. It starts with the Input Layer, which accepts a fused feature vector representing the input data. This fused feature vector is then passed through multiple Hidden Layers, each consisting of fully connected neurons with non-linear activation functions (such as sigmoid). The introduction of non-linear activation functions makes it possible for the network to discover intricate patterns and connections in the data. To optimize the activation function of the MLP, a new self-improved Bacterial Foraging Optimization (SI-BFO) algorithm is proposed. This optimization technique helps fine-tune the parameters of the activation functions, enabling the MLP to achieve better generalization and convergence properties during the training process. The SI-BFO method offers an efficient and effective way to find optimal configurations for the activation functions, further enhancing the performance of the MLP.

The MLP is utilized for this study's objectives. Three different layer types are characteristic of MLP, an instance of artificial neural network that is feed-forward (ANN):

- Input layer,
- Hidden layer and
- Output layer.

MLP can separate data that cannot be separated linearly if it is appropriately designed. This characteristic makes MLP useful for regressing and classifying a variety of problems. Each artificial neuron (AN), the fundamental component of each MLP, is created using the activation function. The activation function (Y) determines the output of each AN are shown in Equation (3.8):

$$Y = f(u) \quad (3.8)$$

3. Logistic sigmoid Activation function

As follows is the formula for the logistic sigmoid activation functions are shown in Equation (3.9):

$$f(x) = \frac{1}{(1 + e^{-x})} \quad (3.9)$$

The mapping of all domain values into the co-domain interval by the logistic sigmoid distinguishes it from Tanh activation function, which is a significant difference denoted in Equation (3.10):

$$0 < f(x) < 1 \quad (3.10)$$

Positive output value will occur when a logical sigmoid is applied. Tanh activation function may not always be applied in this situation.

To optimize the activation function of MLP here we use Self improved Bacterial Foraging Optimization.

3.4.1. Bacterial Foraging Optimization. Accurately mimics the key processes that *E. coli* uses to produce while it hunts for food, namely chemotaxis, reproduction, elimination and dispersal. Bacterial Foraging Optimization flow chart is given below Fig. 3.4.

Step 1: Using a random number generator, random coordinates can be assigned within a predetermined range to initialise the population of bacteria with random positions Equation (3.11).

$$\Theta_i(j, k, l) \quad (3.11)$$

Step 2: Fitness computation Equation (3.12)

$$F = \min(E) \quad (3.12)$$

1. Chemotaxis

Each bacteria travels gradually towards the goals by swimming and tumbling due to its ability to get away from dangerous items quickly and get nutrients. Bacteria travel continuously with defined run lengths while choosing one route at random within the search space, on the one hand. After falling,

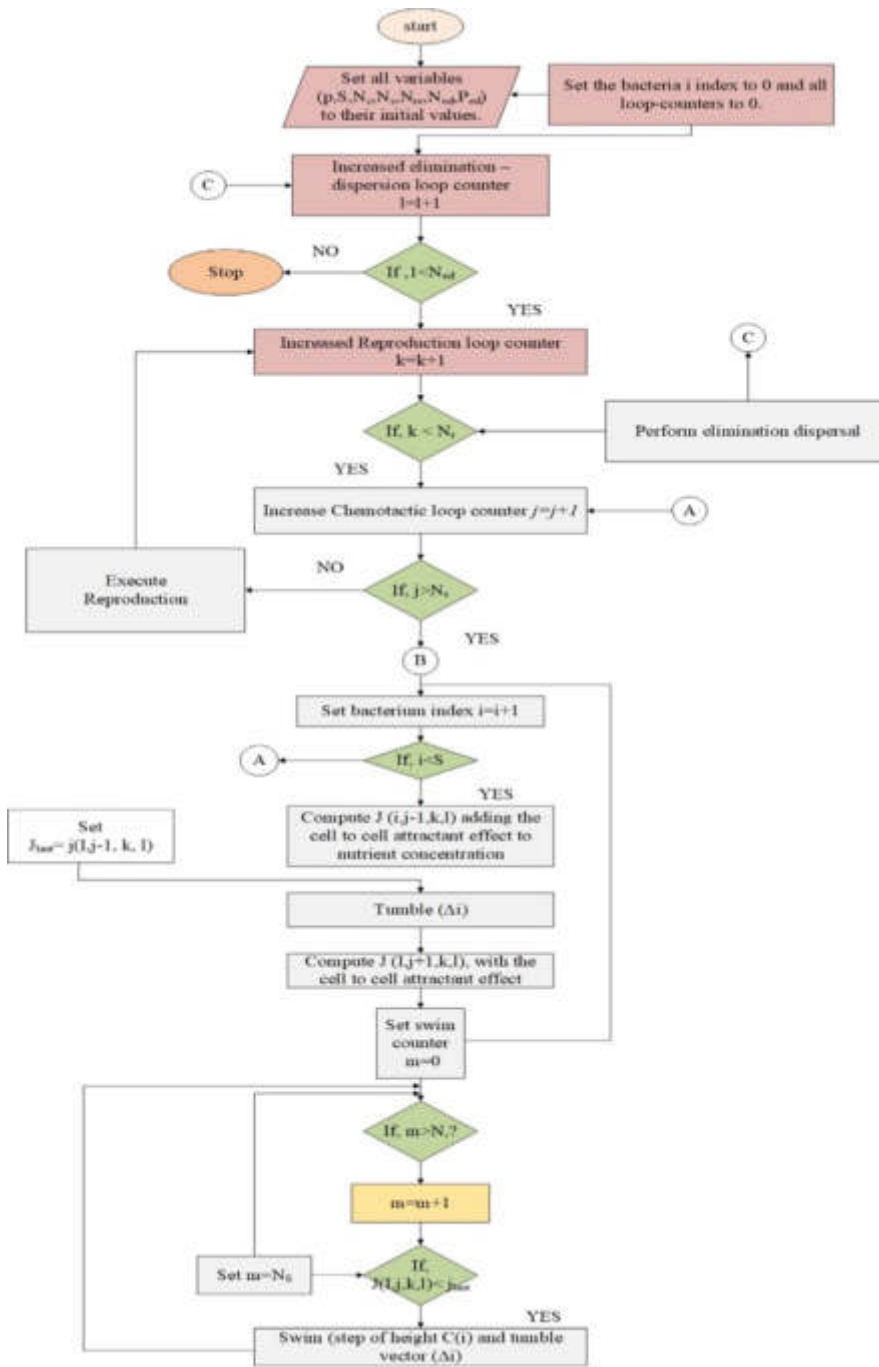


Fig. 3.4: Bacterial Foraging Optimization

bacteria could not continue swimming in the same direction until their updated position became worse or until the number of possible moves reached the N_c limit. The new position of the bacterium i during the $j+1$ th chemotaxis, k th reproduction, and l th elimination and dispersal, where $\theta^i(j, k, l)$ denotes the bacterium's previous position, $C(i)$ denotes the step size, and $i(i)$ denotes a random direction

vector with all of its elements falling between -1 and 1. And the Equation (3.13) is combined by Lens Opposition-Based Learning.

$$\theta^i(j+1, k, l) = \theta^i(j, k, l) + C(i) \times \Delta(i) \times \frac{\Delta T(i)}{(\Delta(i)^2 + \varepsilon)} \quad (3.13)$$

Initialize the fitness value $J(i, j, k, l)$ to evaluate its performance.

Generate a random direction vector $\Delta(i)$ with elements ranging from -1 to 1.

Generate a random direction vector $\Delta T(i)$ with elements ranging from -1 to 1.

Update the position of bacterium i using Lens Opposition-Based Learning.

Update the step size for bacterium i are mentioned in Equation (3.14)

$$C(i) = C(i)(1 + \delta) \quad (3.14)$$

2. Reproduction

Following the central tenet of Darwin's "Survival of the Fittest" theory, the BFO algorithm's reproduction process reflects the fact that healthier bacteria are more likely to have the remarkable capacity for reproduction to sustain the entire swarm population, while undernourished individuals will ultimately be wiped out. The BFO algorithm records the bacterium's health level as $f(i, health)$, which may be calculated from the total of fitness values over the course of its existence. In relation to this, the relevant mathematical statement may be shown as Equation (3.15)

$$f_{i, \text{health}} = \sum_{j=1}^{N_c} J(i, j, k, l) \quad (3.15)$$

where $J(i, j, k, l)$ is the fitness value of the bacterium i in the j^{th} chemotaxis, k^{th} reproduction, l^{th} elimination and dispersal, and N_c is the total number of chemotaxes that the bacterium i undergoes over its career. When the health value of each bacterium is ranked in ascending order, half of the healthier bacteria ($Sr = SS/2$) can split into two bacteria with identical positions while the remaining bacteria are discarded.

If $i \leq Sr$ it is denoted by healthier bacteria

In this algorithm Triangle Walk Strategy is additionally combined with reproduction of the search agent here we denoted the Equation (3.16)

$$\theta^i(j, k+1, 1) = \theta^i(j, k, 1) + \lambda \times (\theta^i(j, k, 1) - \theta^i(j, k-1, 1)) \quad (3.16)$$

In this algorithm Levy Flight Walk Strategy is additionally combined with reproduction of the search agent here we denoted the Equation (3.17)

$$\theta^i(j, k+1, 1) = \theta^i(j, k, l) + \alpha \times L \times \Delta(i) \quad (3.17)$$

Update the positions of bacteria based on the chosen strategy.

3. Elimination and Dispersal Actually, because of the constant, drastic shift in its environment, bacteria may be exposed to a variety of unanticipated dangers, such as the invasion of harmful substances or a change in the local area's dynamic temperature. As a result, when confronted with these unfavorable and unexpected circumstances, some germs must spread out as quickly as possible to a more favorable place. Based on it, the bacteria i randomly migrates to a new site Δ^i following the reproductive process with a specific probability P_{ed} , if not it stays in the existing position.

For each bacterium i , In a uniform distribution between 0 and 1, a random number R is produced. If the generated number R is less than or equal to the dispersal threshold P_{ed} , the bacterium performs dispersal to a new position Δ^i using the Levy Flight Walk Strategy. A mathematical simulation of the movement of organisms based on heavy-tailed distributions is the Levy Flight Walk Strategy in Equation 3.18.

$$\theta^i = \theta^i(j, k, l) + \alpha X L X \Delta^i \quad (3.18)$$

Else: Remain in the current position $\Delta^i(j, k, l)$.

4. Termination Repeat steps 3-5 until a termination condition is met, up to a certain number of iterations. Return the best solution found during the iterations.

3.4.2. Modified RNN. In a sequential user-item interaction-based recommendation system, the input layer processes the historical interaction data, capturing the sequential nature of user actions. The embedding layer transforms the user and item IDs into dense representations, enabling the model to capture latent features. LSTM layers are employed to effectively model temporal dependencies in the interactions, allowing the system to learn from past behaviors and preferences. An appropriate loss function, like mean squared error (MSE), is defined to quantify the prediction error and guide the model towards better recommendations. The fusion layer combines the final hidden states of the user and item embeddings, creating a joint representation that captures their interactions and preferences. Finally, the output layer leverages this fused information to predict user preferences or item similarities, providing personalized suggestions for users based on their prior behaviour and previous interactions. The combination of these layers enables the recommendation system to learn from sequential data, handle temporal dynamics, and generate accurate predictions to enhance the overall user experience in the e-commerce platform.

3.4.3. Collaborative Filtering. In order to offer consumers items or services, Ratings, past purchases, and browser history are all used in personalised recommendation services. Furthermore, customers who have trouble selecting between different products and services would find a solution like this that provides personalised recommendations useful. Global companies like Netflix, Amazon, and Google make money by providing tailored suggestion services in e-commerce to assist users in making decisions.

The Collaborative Filtering component utilizes an autoencoder architecture to effectively capture user-item interactions in a recommendation system. The autoencoder comprises three key components: the Encoder, Decoder, and Output Layer. The Encoder transforms user and item IDs into lower-dimensional representations, effectively compressing the information while preserving essential patterns and relationships. The Decoder then reconstructs the original input from these encoded representations, allowing the autoencoder to learn the underlying structures in the data. Finally, the Output Layer leverages the learned representations to predict user preferences for items or similarities between items, aiding in generating personalized recommendations. By employing this autoencoder-based approach, the Collaborative Filtering component can efficiently extract meaningful features from user-item interactions, providing recommendation systems with more precise and customized recommendations.

Cosine similarity. Finding things that are comparable to those that the target user has liked or interacted with during the recommendation process and recommending those items are both steps in the process. Cosine similarity is one similarity measure that can be used to determine how similar two things are showed in Eq. (3.19).

$$\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \cdot \|\vec{b}\|} \quad (3.19)$$

3.4.4. Item-based collaborative filtering. The idea behind the suggestion method known as "item-based collaborative filtering" is that buyers tend to favor goods that are similar to those they have already loved. This strategy uses the idea of item similarity to offer tailored recommendations. The technology finds things that have been often eaten or highly rated in tandem by examining the previous behavior and preferences of consumers. This information is then used to generate recommendations by suggesting products that are comparable to those the user has already indicated an interest in. Since this approach relies on item relationships rather than explicit user preferences, it has a number of benefits, including scalability and accuracy. Item-based collaborative filtering, which takes item similarity into account, enhances the system's ability to capture subtleties and offer pertinent recommendations.

3.5. Feature fusion. The collaborative filtering component, along with the outputs of the MLP, RNN, and Transformer-based Model components, can be integrated to produce an extensive ensemble model. A fusion approach that combines the individual model outputs, is used to achieve this integration. By combining the predictions from many models, we can take use of each architecture's advantages and provide a final prediction

that is more reliable and accurate. By maximizing the aggregate intelligence of the multiple models, this ensemble technique enhances generalization and performance on a range of tasks and datasets.

4. Result and discussion. In this section, the results of the suggested procedures are compared to those of the current methods. The implementation is done with Python. Myntra Fashion Product Dataset comprising 14,481 samples was utilized, accompanied by 14,330 metadata entries, and subjected to a rigorous data cleaning process. The feature extraction phase resulted in 6,581 relevant features. For model training and evaluation, the dataset was split into two different sets: 70% for training and 30% for testing in one experiment, and 80% for training and 20% for testing in another experiment. This separation allowed for the assessment and comparison of the models' performance under different learning rate scenarios [26].

Several metrics, including Mean Squared Error (MSE), Mean square relative error (MSRE), Normalized Mean Squared Error (NMSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE), are used in this e-commerce recommendation system to evaluate the accuracy and efficacy of the recommendations. These metrics give us a way to gauge how closely the real user preferences match the projected recommendations, giving us important information about the performance of the system.

4.1. Performance matrices. Evaluation is done using error measures such as Mean Absolute Percentage Error (MAPE), Mean Square Error (MSE), Normalized Mean Square Error (NMSE), Mean Square Relative Error (MSRE), and Root Mean Square Error (RMSE).

MSE: Calculated is the average squared difference between the expected and actual values. The MSE loss, which is also used to direct model training, is used to evaluate a model's performance. The better the model fits the data, the lower the MSE shown in Equation (4.1)

$$MSE = \frac{1}{N} (A_v - P_v)^2 \quad (4.1)$$

Where, A_v is the target variables actual value P_v , where N is the total number of values, and is the target variables predicted value.

MSRE: The Mean Square Relative Error (MSRE) is a metric used to measure the accuracy of a predictive model. It is particularly useful when evaluating models that make continuous predictions, such as regression models shown in Equation (4.2).

$$MSRE = \sqrt{\left(\frac{\sum [(y_{true} - y_{pred})^2 / y_{true}^2]}{n} \right)} \quad (4.2)$$

NMSE: The Normalized Mean Square Error (NMSE) is a metric for comparing two signals that is frequently used to assess how well a prediction algorithm is working. It is described as the difference between the variance of the target signal and the MSE of the forecast signal shown in Equation (4.3)

$$NMSE = \left(\frac{1}{w} \right) * \text{sum} \left(\frac{(A_v - P_v)^2}{\text{var}(A_v)} \right) \quad (4.3)$$

RMSE: In order to evaluate how different disease symptoms affect the effectiveness of the treatments used at various disease severity levels, RMSE is used in this study in Equation 4.4

$$RMSE = \sqrt{\frac{1}{N} (A_v - P_v)^2} \quad (4.4)$$

MAPE: The MAPE formula is used (individually for each period) by multiplying the demand by the total number of distinct absolute errors in Equation (4.5)

$$MAPE = \frac{1}{T_j} \sum_{j=1}^{T_j} \left| \frac{A_v - F_v}{\text{var}(A_v)} \right| \quad (4.5)$$

Table 4.1: Metrics-Learn Rate (70%)

	BFO	EO	DBN	CNN	RNN	Proposed (HDCT)
MSE	0.3360	0.3237	0.3110	0.3264	0.3090	0.2971
MSRE	0.3007	0.2797	0.2883	0.3186	0.3319	0.2763
NMSE	0.4453	0.4224	0.4195	0.4515	0.4487	0.4013
RMSE	0.3604	0.3382	0.3878	0.3674	0.3893	0.3222
MAPE	0.3293	0.3172	0.3048	0.3199	0.3029	0.2911

Table 4.2: Metrics-Learn Rate (80%)

	BFO	EO	DBN	CNN	RNN	Proposed (HDCT)
MSE	0.2618	0.2499	0.2640	0.2718	0.2515	0.2403
MSRE	0.2262	0.2684	0.2577	0.2427	0.2332	0.2234
NMSE	0.3690	0.3919	0.3944	0.3889	0.3665	0.3506
RMSE	0.2050	0.2433	0.2336	0.2199	0.2113	0.2025
MAPE	0.2995	0.3119	0.2629	0.2710	0.2820	0.2597

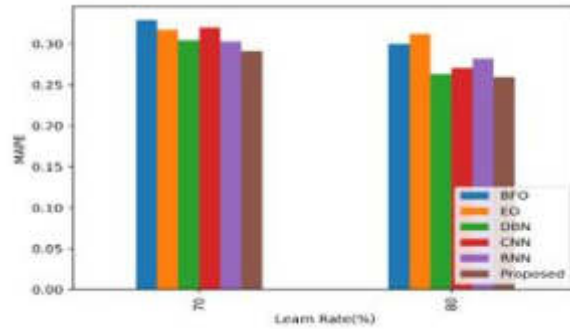


Fig. 4.1: Mean Squared Error analysis graph

where T_j is the total number of occurrences for the summation iteration, A_v is the actual value, F_v is the anticipated value.

4.2. Mean Squared Error. Table 4.1 shows that the techniques are represented by the models BFO, EO, DBN, CNN, RNN, and PROPOSED (HDCT) that are supplied. These approaches' respective mean squared error (MSE) values are 0.3360, 0.3237, 0.3110, 0.3264, 0.3090, 0.2971. It is clear that among the dataset's most recent 70%, the proposed approach has the lowest error consumption.

Table 4.2 shows that the models BFO, EO, DBN, CNN, RNN, and PROPOSED (HDCT) provided represent several methods. The corresponding mean squared errors (MSE) for these methods are 0.2618, 0.2499, 0.2640, 0.2718, 0.2515, and 0.2403. It is evident that among the most recent 80% of the dataset, the suggested strategy has the lowest error consumption.

The graph represents the 70% and 80% of the learn rate of in Mean Squared Error are represented in Fig. 4.1.

4.3. Mean square relative error. The models BFO, EO, DBN, CNN, RNN, and PROPOSED (HDCT) that are provided are shown in Table 4.1 to reflect the methodologies. Mean square relative error values for these methods are 0.3007, 0.2797, 0.2883, 0.3186, 0.3319, and 0.2763, respectively. It is evident that among the

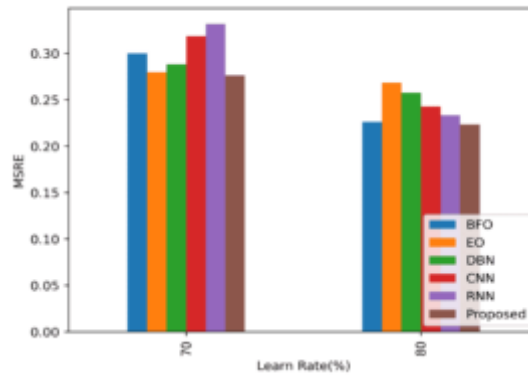


Fig. 4.2: Mean Squared relative Error analysis graph

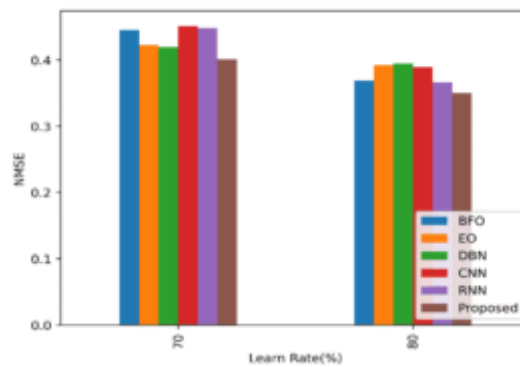


Fig. 4.3: Normalized mean square error

most recent 70% of the dataset, the suggested strategy has the lowest error consumption.

Table 4.2 demonstrates that the offered models BFO, EO, DBN, CNN, RNN, and PROPOSED (HDCT) represent several approaches. For these techniques, the associated Mean square relative error is 0.2262, 0.2684, 0.2577, 0.2427, 0.2332, and 0.2234. It is clear that the suggested technique has the lowest error consumption among the most recent 80% of the dataset.

The graph represents the 70% and 80% of the learn rate of in Mean square relative error are represented in Fig. 4.2.

4.4. Normalized mean square error. The models BFO, EO, DBN, CNN, RNN, and PROPOSED (HDCT) that are provided are shown in Table 4.1 to reflect the methodologies. Normalized mean square error values for these methods are 0.4453, 0.4224, 0.4195, 0.4487, 0.4013 respectively. It is evident that among the most recent 70% of the dataset, the suggested strategy has the lowest error consumption.

Table 4.2 shows that the available models BFO, EO, DBN, CNN, RNN, and PROPOSED (HDCT) represent several methodologies. The corresponding Normalized mean square error for these methods are 0.3690, 0.3919, 0.3944, 0.3889, 0.3665, and 0.3506. It is evident that among the most recent 80% of the dataset, the suggested technique has the lowest error consumption.

The graph represents the 70% and 80% of the learn rate of in Normalized mean square error are represented in Fig. 4.3.

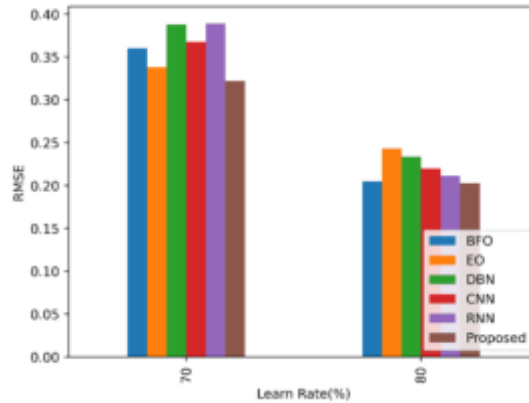


Fig. 4.4: Root Mean square error analyzing graph

4.5. Root Mean square error. Table 4.1 displays the models BFO, EO, DBN, CNN, RNN, and PROPOSED (HDCT) that are offered in order to illustrate the approaches. These methods' respective Root Mean Square Error values are 0.360416, 0.338242, 0.387806, 0.367417, 0.389381, and 0.322242. It is clear that the suggested technique has the lowest error consumption among the most recent 70% of the dataset.

The accessible models BFO, EO, DBN, CNN, RNN, and PROPOSED (HDCT) represent various methods, as shown in Table 4.2. For each of these approaches, the associated Root Mean Square Error is 0.2995, 0.3119, 0.2629, 0.2710, 0.2820, and 0.2597. It is clear that the suggested technique has the lowest error consumption among the most recent 80% of the dataset.

The graph represents the 70% and 80% of the learn rate of in Root Mean square error are represented in Fig. 4.4.

4.6. Mean absolute percentage error. Table 4.1 displays the models BFO, EO, DBN, CNN, RNN, and PROPOSED (HDCT) that are offered in order to illustrate the techniques. These methods' respective mean absolute percentage error values are 0.3293, 0.3172, 0.3048, 0.3199, 0.3029, and 0.2911. It is clear that the suggested technique has the lowest error consumption among the most recent 70% of the dataset.

Table 4.2 shows that the available models BFO, EO, DBN, CNN, RNN, and PROPOSED (HDCT) represent several methodologies. The corresponding Normalized mean square error for these methods are 0.2995, 0.3119, 0.2629, 0.2710, 0.2820, 0.2597. It is evident that among the most recent 80% of the dataset, the suggested technique has the lowest error consumption.

The graph represents the 70% and 80% of the learn rate of in Root Mean square error are represented in Fig. 4.5.

4.7. Existing result analysis . In the existing result analysis, two root mean square error (RMSE) values are presented: one labelled as "[5]" and the other as "PROPOSED."

RMSE: The first RMSE value, 0.2200, is associated with an unidentified method or model denoted by "[5]."

Without further context or information, it is unclear what this value represents or what specific technique or model was used to obtain it.

PROPOSED: The second RMSE value, 0.2025, is associated with the method or model labelled as "PROPOSED." This value likely represents the root mean square error achieved by the proposed method or model in a certain experiment or analysis.

5. Conclusion. In conclusion, the Hybrid Deep Collaborative Transformer (HDCT), which has been suggested for e-commerce suggestions, performs well and exceeds other models already in use. But it's critical to recognize and deal with any mistakes or restrictions that can occur while optimizing. The HDCT approach can be improved and strengthened by carefully identifying and correcting these errors in e-commerce suggestions.

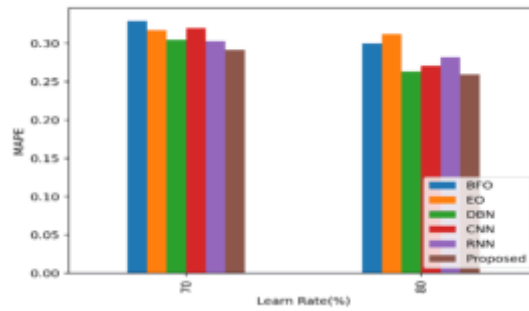


Fig. 4.5: Mean absolute percentage error Analysis

This focus on the little things and ongoing development guarantees that the HDCT model will continue to be dependable and effective in giving precise and individualized recommendations to e-commerce users. The identification and resolution of potential errors contribute to the overall robustness and trustworthiness of the HDCT method, enhancing its practicality and data for the RS is collected from the Myntra fashion product dataset ability in dataset e-commerce scenarios. As the field of e-commerce continues to evolve, it is imperative to remain vigilant in refining and advancing recommendation models like HDCT, ensuring optimal performance and customer satisfaction in the ever-growing digital marketplace.

Data Availability. The data availability statement is mentioned in the paper.

REFERENCES

- [1] Ping, N.L. *Constructs for artificial intelligence customer service in E-commerce*. In 2019 6th International Conference on Research and Innovation in Information Systems (ICRIIS) pp. 1-6, IEEE, 2019.
- [2] Habib, A., Irfan, M. and Shahzad, M. *Modeling the enablers of online consumer engagement and platform preference in online food delivery platforms during COVID-19*. Future Business Journal, 8(1), 6, 2022.
- [3] Liao, S.H., Widowati, R. and Hsieh, Y.C. *Investigating online social media users' behaviors for social commerce recommendations*. Technology in Society, 66, p.101655, 2021.
- [4] Battisti, S., Agarwal, N. and Brem, A. *Creating new tech entrepreneurs with digital platforms: Meta-organizations for shared value in data-driven retail ecosystems*. Technological Forecasting and Social Change, 175, p.121392, 2022.
- [5] Zhao, Q., Zhang, Y., Friedman, D. and Tan, F. *E-commerce recommendation with personalized promotion*. In Proceedings of the 9th ACM Conference on Recommender Systems pp. 219-226, 2015.
- [6] Ehikioya, S.A. and Guillemot, E. *A critical assessment of the design issues in e-commerce systems development*. Engineering Reports, 2(4), p.e12154, 2020.
- [7] Sharma, S., Singh, S., Kujur, F. and Das, G. *Social media activities and its influence on customer-brand relationship: an empirical study of apparel retailers' activity in India*. Journal of Theoretical and Applied Electronic Commerce Research, 16(4), pp.602-617.
- [8] Deng, S., Tan, C.W., Wang, W. and Pan, Y. *Smart generation system of personalized advertising copy and its application to advertising practice and research*. Journal of Advertising, 48(4), pp.356-365.
- [9] Tolstoy, D., Nordman, E.R., Hännell, S.M. and Özbek, N. *The development of international e-commerce in retail SMEs: An effectuation perspective*. Journal of World Business, 56(3), p.101165.
- [10] Cui, Z., Xu, X., Fei, X.U.E., Cai, X., Cao, Y., Zhang, W. and Chen, J., 2020. *Personalized recommendation system based on collaborative filtering for IoT scenarios*. IEEE Transactions on Services Computing, 13(4), pp.685-695, 2020.
- [11] Shahbazi, Z., Hazra, D., Park, S. and Byun, Y.C. *Toward improving the prediction accuracy of product recommendation system using extreme gradient boosting and encoding approaches*. Symmetry, 12(9), p.1566, 2020.
- [12] Cui, Z., Xu, X., Fei, X.U.E., Cai, X., Cao, Y., Zhang, W. and Chen, J. *Personalized recommendation system based on collaborative filtering for IoT scenarios*. IEEE Transactions on Services Computing, 13(4), pp.685-695, 2020.
- [13] Sohn, K. and Kwon, O. *Technology acceptance theories and factors influencing artificial Intelligence-based intelligent products*. Telematics and Informatics, 47, p.101324, 2020.
- [14] Nguyen, A.T., Parker, L., Brennan, L. and Lockrey, S. *A consumer definition of eco-friendly packaging*. Journal of Cleaner Production, 252, p.119792, 2020.
- [15] Grewal, D., Hulland, J., Kopalle, P.K. and Karahanna, E. *The future of technology and marketing: A multidisciplinary perspective*. Journal of the Academy of Marketing Science, 48, pp.1-8, 2020.

- [16] Wang, K., Zhang, T., Xue, T., Lu, Y. and Na, S.G. *E-commerce personalized recommendation analysis by deeply-learned clustering*. Journal of Visual Communication and Image Representation, 71, p.102735, 2020.
- [17] Li, Q., Li, X., Lee, B. and Kim, J. *A hybrid CNN-based review helpfulness filtering model for improving e-commerce recommendation Service*. Applied Sciences, 11(18), p.8613, 2021.
- [18] Xiang, D. and Zhang, Z. *Cross-border e-commerce personalized recommendation based on fuzzy association specifications combined with complex preference model*. Mathematical Problems in Engineering, pp.1-9, 2020.
- [19] Chen, Y. *Research on personalized recommendation algorithm based on user preference in mobile e-commerce* Retraction of Vol 18, Pg 837, 2020.
- [20] Wang, J. and Zhang, Y. *Using cloud computing platform of 6G IoT in e-commerce personalized recommendation*. International Journal of System Assurance Engineering and Management, 12, pp.654-666, 2020.
- [21] Guo, X., Wang, S., Zhao, H., Diao, S., Chen, J., Ding, Z., He, Z., Lu, J., Xiao, Y., Long, B. and Yu, H. *Intelligent Online Selling Point Extraction for E-Commerce Recommendation*. In Proceedings of the AAAI Conference on Artificial Intelligence Vol. 36, No. 11, pp. 12360-12368, 2022.
- [22] Thongsri, N., Warintarawej, P., Chotkaew, S. and Saetang, W. *Implementation of a personalized food recommendation system based on collaborative filtering and knapsack method*. International Journal of Electrical and Computer Engineering, 12(1), pp.630-638, 2020.
- [23] Zhou, L. *Product advertising recommendation in e-commerce based on deep learning and distributed expression*. Electronic Commerce Research, 20(2), pp.321-342, 2020.
- [24] Chehal, D., Gupta, P. and Gulati, P. *Implementation and comparison of topic modeling techniques based on user reviews in e-commerce recommendations*. Journal of Ambient Intelligence and Humanized Computing, 12, pp.5055-5070, 2021.
- [25] Zhang, Y., Abbas, H. and Sun, Y. *Smart e-commerce integration with recommender systems*. Electronic Markets, 29, pp.219-220, 2019.
- [26] Dataset Available: (<https://www.kaggle.com/datasets/hiteshsuthar101/myntra-fashion-product-dataset>)

Edited by: Polinapilinho F. Katina

Special issue on: Scalable Dew Computing for future generation IoT systems

Received: Dec 8, 2023

Accepted: Apr 1, 2024



BRAIN TUMOR CLASSIFICATION USING REGION-BASED CNN WITH CHICKEN SWARM OPTIMIZATION

A SRAVANTHI PEDDINTI *, SUMAN MALOJI † AND KASIPRASAD MANNEPALLI ‡

Abstract. Diagnosing and segmenting brain tumors manually through MRI imaging is a complex and time-consuming process. However, advancements in machine learning (ML) and deep learning (DL) technologies have enabled the automatic identification and categorization of brain tumors using computer-aided design. This study utilizes MRI data to develop a system for the automatic identification and categorization of brain tumors based on region-based convolutional neural networks (R-CNN). The proposed R-CNN approach, coupled with Chicken Swarm Optimization (CSO) technique, enables the identification and classification of brain tumors into stages. This method involves processing, segmenting, extracting features, and organizing the MRI images. Image preparation includes adaptive fuzzy filtering (AFF) to eliminate noise and enhance the quality of MRI images. To detect regions of brain injury, MRI scans undergo cranial segmentation and classification (CSO) based on Tsallis entropy-based image segmentation. A Residual Network (ResNet) is employed to fuse handcrafted and deep features, generating a meaningful set of feature vectors. Extensive simulations are conducted on the BRATS 2015 dataset to evaluate the improved performance in classifying brain tumors. The RCNN-CSO method demonstrates superior performance compared to other contemporary techniques, achieving a precision of 92.35%, sensitivity of 93.52%, specificity of 94.52 % and an accuracy of 96 % . This represents a significant improvement in brain tumor classification and its outomst accuracy.

Key words: Gliomas; Chicken swarm optimization; adaptive fuzzy filtering; RCNN-CSO; Residual network.

1. Inntroduction. Brain tumors, often called lesions or neoplasms, result from the uncontrolled proliferation of aberrant brain tissue. Brain tumors hurt brain function and threaten human life when abnormal tissue growth occurs in the brain. Physicians benefit from the use of computer-aided diagnosis systems that aid in the interpretation of medical images. It means that X-ray [1], ultrasound, and magnetic resonance imaging diagnostics must be dealt with by radiologists or specialists who must quickly review and analyze them to get clearance. With the emergence of machine learning algorithms and whole-side imaging, digital pathology has been approaching. Artificial Intelligence, pathological and radiological image processing, and CAD technology are merged into computer vision elements. It's becoming increasingly common for doctors to employ imaging techniques like MRIs and CT scans to identify cancers to better understand and treat patients. The non-invasive digital imaging techniques include CT scans, MRIs, X-rays, SPECTs, PETs, and ultrasounds. Machine Learning (ML) techniques can be used to improve early identification of brain cancers. Figure 1.1 depicts MRI images of the brain with a tumor and without a tumor.

The CSO algorithm is a variant of swarm intelligence optimization, initially proposed by Vasantharaj et al. [2] and further refined by Meng et al. [3]. The chicken flock's fitness ratings consider when the algorithm determines how to partition a flock of chickens into the three distinct categories of roosters, hens, and chicks. After that, each conducts their search in the solution space. Each particle in the algorithm indicates a different approach that could be taken to solve the issue. In the final step, a detailed comparison of the three groups' fitness values is performed to identify the globally ideal particle and the globally optimal value. The individuals with the highest overall fitness levels are designated as roosters among the flock. They can locate food across a greater area. The contributions of proposed RCNN-CSO for classifying the brain tumors:

- Deep learning algorithms use a revolutionary concept known as prediction at the pixel level.

*Department of ECE, Koneru Lakshmaiah Education Foundation, KLEF (Deemed to be University), Vaddeswaram, AP, India (sravanthi.angara@gmail.com).

†Department of ECE, Koneru Lakshmaiah Education Foundation, KLEF (Deemed to be University), Vaddeswaram, AP, India (Suman.maloji@kluniversity.in).

‡Department of ECE, Koneru Lakshmaiah Education Foundation, KLEF (Deemed to be University), Vaddeswaram, AP, India (mkasiprasad@kluniversity.in).

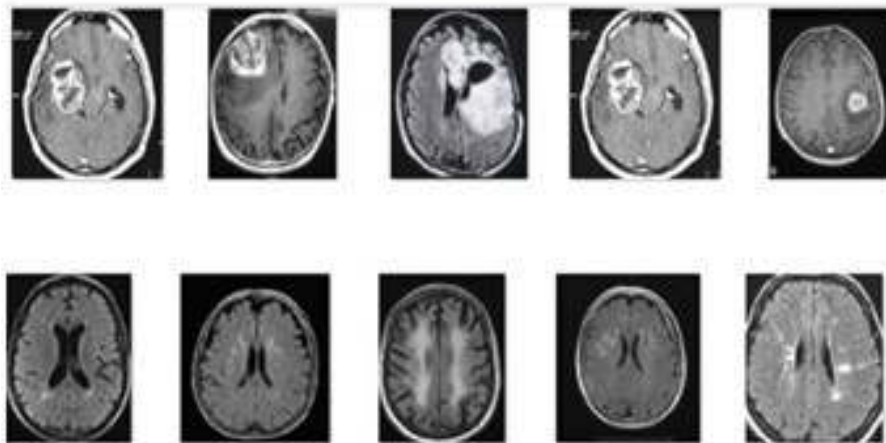


Fig. 1.1: Brain Tumor low intensity images.

- This study utilized a deep Convolutional Neural Network (CNN) known as RCNN-CSO to enhance the output process. Specifically, CNN was trained using MRI scans to classify brain tumors accurately.
- Pre-processing images with adaptive fuzzy filtering (AFF) removes noise and improves the quality of MRI images.
- MRI images are segmented using a technique called chicken swarm optimization (CSO), which uses Brain injury can be localized using Tsallis entropy-based picture segmentation.

The remaining sections are as follows: a discussion of relevant literature in Section 2, a presentation of the suggested method in Section 3, a display of results and discussions in Section 4, a performance validation in Section 5, and finally, a summary and conclusion in Section 6.

2. Related Works. The existing MRI-based brain tumor segmentation [4] approaches are divided into three groups: traditional techniques, classification strategies, cluster analysis, and deformable model methods. Demirhan et al. [5] classified brain MR images as tumors, white matter, grey matter, CSF, and edema. Since tumor growth on healthy brain tissues was studied, healthy and malignant brain tissues identifies. This information is crucial for treatment planning. Twenty glioma patients were examined using T1, T2, and FLAIR images.

In another study, Srinivas and Sasibhusana Rao [6] used K-means & Fuzzy C-means clustering for tumor detection in MRI scans. K-means and FCM clustering algorithms were compared using relative area, PSNR, segmented area, and MSE. Particle Swarm Optimization- With PSO, finding the best answer is more accessible, while Genetic Algorithms (GAs) get a near-optimal but not exact answer. Combining GAs and PSO determined the optimal degree of attraction. Combining The k-means & fuzzy c-means algorithms improved the precision with which tumor stages & sizes were determined. This method allows accurate and repeatable tumor tissue segmentation, like manual segmentation. It accelerates segmentation.

Abhishek Bal et al. [7] developed an automated method for segmenting brain tumors (RFCM). Using fuzzy membership and A rough set with a high and low bounce, rough-fuzzy C-means can address overlapping division. Both hard lower estimation and the fuzzy border have been helpful for RFCM brain tumor segmentation. C-means algorithm initial centroids may need to be revised. This work was used to speed up RFCM execution by picking starting centroids. Manocha et al. [8] mapped the tumor using automated segmentation in a different study. Brodmann area identification was included in the GUI technique for brain tumor segmentation. The research used 15 patients' T2-weighted images. Based on ethnicity, gender, and age, these data were divided into three groups. The "fuzzy" C indicates tumor clustering. Normalizing tumor segmentation data reduces image artifacts. This project's main benefit was a simple, user-friendly GUI simplifying the entire procedure.

Shree et al. [9] system for detecting, segmenting, & identifying neoplasms in the brain took time and effort.

Seeing abnormal brain tissue is difficult. Radiologists can't do further research using current procedures due to a lack of information on tumor grades. Justyna Tomicka J et al. [10] have proposed MRI modalities like The fields of Magnetic Resonance Spectroscopy (MRS), Diffusion Tensor Imaging (DTI), and Perfusion Imaging (PI) are gaining ground in the field of brain tumor segmentation. These methods have been used to localize tumors in the brain successfully. Kumar et al. [11] first proposed using optimization-driven deep convolutional neural networks to classify brain tumors. The MRI scans used for the analysis came from both the BRATS database and the Sim BRATS.

Rammurthy et al. [13] introduced the hybrid Whale Harris Hawks optimization (WHHO) for diagnosing tumors in the brain by exploiting MR imageries. The DCNN method was contemplated for the identification of tumors. The classification along segmentation method has been presented by Deb et al. [14] as a method for identifying malignant growths in the brain. Images' abnormality or normalcy establishes using a frog-leap-optimized Adaptive fuzzy deep neural network. Yin et al. [15] have shared their research in diagnosing brain tumors at an early stage. Here, early detection consists of 3 steps: reducing the noise in the background, extracting relevant features, and classifying them using a multi-layer perceptron NN. Using a deep neural network, Sultan et al. [16] demonstrated multi-classification of images of brain tumors. In this instance, they used a Deep Learning (DL) model and a convolutional neural network (CNN) to assign categories to the photographs.

Pandiselvi et al. [15] suggested an ACRC method for adaptable convex region contour. The normality or abnormality of this case was determined by using a Support Vector Machine (SVM) to slice categorization. Deepak et al. [17] employed a combination of convolutional neural networks and support vector machines. Brain tumor MRI data is automatically classified. In this study, classification was accomplished by using a computer-aided diagnostic (CAD) method for the human brain. CNN was utilized to extract the features. In addition to that, an SVM classifier uses to categorize the characteristics.

To classify brain tumors using MRI scans, Narmatha et al. [18] describe a hybrid fuzzy brain-storm optimization technique. The fuzzy brain-storm optimization algorithm is used to maximize the undefined parameters' effectiveness (FBSO). From the above discussion, we observed that the limitations in finding the usual medical imaging dataset for an analysis of a method can be complex. Therefore, many approaches test on a few data sets with varying metrics. Because of these challenges, comparisons between the presented methods and manually segmented medical images cannot be straightforward. Neural networks, clustering techniques like K-means and FCM, support vector machines, and deep learning algorithms like CNN can help enhance MRI picture quality by isolating the brain tumor. [19].

3. Proposed Methodology. This approach includes training a model built of numerous neural network processing layers. Creation of an automated Using MRI scan images, we present a deep learning-based method for tumor detection and classification in the brain. The stages of a brain tumor's development can identify and categorized using the RCNN-CSO method that has been proposed. Before the MRI scans can be processed to reduce noise and enhance the image quality, they must be pre-processed with AFF (Adaptive Fuzzy Filtering). MRI images are segmented using chicken swarm optimization (CSO), which uses picture segmentation using entropy to identify damaged brain sections. As can be seen in Figure 3.1, a valuable set of feature vectors are generated by employing a Residual Network (ResNet) to fuse both handcrafted and deep features.

3.1. Chicken Swarm Optimization (CSO). By utilizing the fractional-CSO algorithm, this paper presents a unique tumor classification methodology that is more accurate than existing methods. According to [20, 21], CSO is a single-objective optimization technique with biological impacts. It imitates how a flock of hens behaves hierarchically while hunting for food, with each chicken standing in for a potential optimization solution. CSO defines four guidelines for the perfect chicken behavior:

- A dominant rooster, hens, and chicks are present in every chicken swarm group.
- The fitness of each chicken determines whether it belongs to a group of hens, roosters, or chicks. The group is led by the roosters, which are the fittest chickens. A chick is a lesser chicken. Some are hens.
- Every G step, mother-child connections, dominance, and hierarchical relationships should shift.
- For nourishment, chickens follow their rooster. We expect that chickens will steal food from one another. Near their moms, chicks go foraging. The rooster wins the food fight. We distinguish between the following numbers in a swarm of N chickens: There are a total of RN roosters, HN hens, CN chicks,

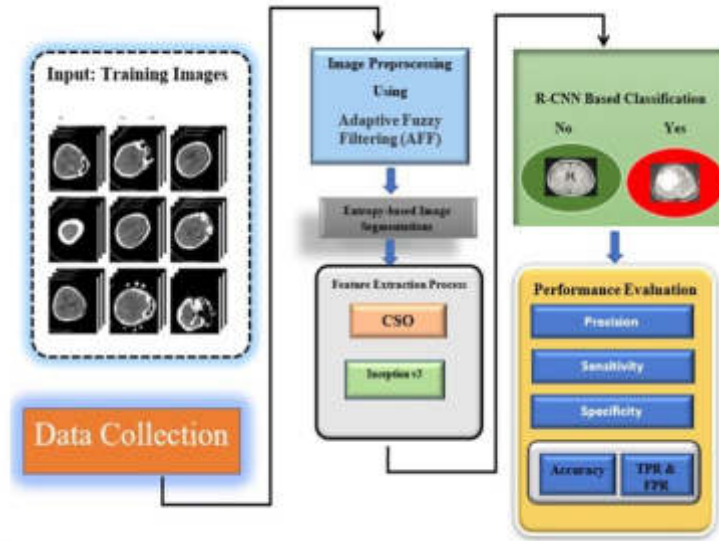


Fig. 3.1: RCNN - Methods for Detecting and Classifying Brain Tumors.

and MN mother hens. The x_{ij} coordinates of each chicken ($I [1,..N]$, $j [1,..D]$) describe their location in D -dimensional space. Chickens come in three different varieties in CSO. Each species has its distinctive dance.

3.2. Rooster Movement. The roosters with the best fitness ratings can find food in more locations than those with lower fitness ratings. Eqs. (1) and (2) describe their motion.

$$x_{ij}^{t+1} = x_{ij}^t * (1 + \text{Randn}(0, \sigma^2)) \quad (3.1)$$

In this context, k is an arbitrary rooster index, and $\text{Randn}(0, 2)$ is a zero-mean Gaussian distribution. A minor constant, 2, and the rooster's fitness, f_i , are denoted below.

Hen movement. Hens engage in foraging behavior in the presence of another group member. In addition, they would strategically appropriate the high-quality food provided by their bird counterparts, all the while being constrained by their fellow gallinaceous companions. The more dominating chickens would triumph over obedient hens in a food fight. The equations (3.1) – (3.5) can be used to mathematically explain the movement of elements.

$$x_{ij}^{t+1} = x_{ij}^t * (S_1 + \text{Randn} * (x_{n,i}^t - x_{n,i}^t) + S_2 + \text{Randn} * (x_{n,i}^t - x_{n,i}^t)) \quad (3.2)$$

$$S_1 = \exp\left(\frac{f_k - f_{r1}}{|f_{r2} + f_1|}\right) \quad (3.3)$$

$$S_2 = |f_{r2}| + f_1 \quad (3.4)$$

The variable Randn represents a random number that follows a uniform distribution between 0 and 1. While $r2$ is an index of a chicken, The variable $r1$ represents an index that relates to the rooster, explicitly referring to the groupmate of the I -th hen. (rooster or hen). It gets picked at random from the swarm ($r1, r2$).

Chick movement. Young chicks perform exploratory behaviors to find sustenance, often venturing away from their maternal figure. The movement of the chick is defined by Equation (3.6).

$$x_{ij}^{t+1} = x_{ij}^t + FL * (x_{m,j}^t - x_{ij}^t) \quad (3.5)$$

where $x_{(m,j)}^t$, The position of the mother of the i -th chick, denoted as m , is defined within the range of 1 to N . FL , a parameter that quantifies the extent to which a chick will mimic the speed of its mother. To examine the distinctions among the several chicks, a random selection is made for the variable FL within the interval $[0,2]$.

3.3. Pre-Processing of Brain Tumor Images. Tumor classification begins with pre-processing because making the subsequent procedure more realistic is crucial. We use the brain imaging dataset to obtain the input image of the brain. During the pre-processing stage, significant tumor sites are extracted [22]. Consider the database E of brain images, which it denotes as, and contains B_n images with $[Q \times R]$ dimensions.

$$F = \{C_1, C_2, \dots, C_i, \dots, C_n\}; [Q \times R] \quad (3.6)$$

C_n represents the data set containing a cumulative count of brain images, denoted as the total number. In contrast, variable F indicates the database's specific quantity of brain photographs. In this scenario, the pre-processing function uses the image B_i from the database. Cropping the regions of interest from image C_i results in an image designated as C_i with the dimension $[Q \times R]$, which is the output of the pre-processing phase. $C_i = b_1, b_2, \dots, b_j, \dots, b_P$ the pre-processed image C_i is characterised by its M number of pixels.

$$\max_R \sum_{K=1}^Q b_j(X_c(R)) \quad (3.7)$$

where $b_j(X_c(R))$ The passage discusses an objective function that is specifically associated with the K th feature of the data, denoted by (R) . The standard parameter.

$$\max_{x,R} \sum_{K=1}^Q w_j, b_j(X_c(R)) \quad (3.8)$$

the pixel values being represented by w_j . To help with clustering model regularization, the expression of the sparse FCM is presented in eq 3.7 to 3.9.

$$\max H(X, R, g) = g^m \lambda(R) \quad (3.9)$$

3.4. Feature extraction for brain tumor classification. The segmented image C^*_{kl} is subsequently transmitted to the feature extraction stage, where the most salient characteristics are extracted from the segmented outcomes. Using feature extraction, we may simplify brain MRIs by eliminating unnecessary data. The mean value of the pixels is represented, which is shown in eq 3.10 to 3.13.

$$g_1 = \frac{1}{Q} \sum_{k=1}^Q C^*_{kq} \quad (3.10)$$

where C^*_{jq} denotes the q th picture element of the image C_j segmentation variance The dimension of feature f_2 is determined by its mean value of $[2 \times 1]$

$$g_2 = \left(\frac{1}{Q} \sum_{k=1}^Q \sqrt{C^*_{kq}} \right)^2 \quad (3.11)$$

Entropy is used to determine which pixels contain the most information.

$$g_3 = - \sum_{k=1}^{Q \sum_{kq}^*} C^*_{kq} \log \quad (3.12)$$

The coverage feature determines the tumor region by calculating the total number of pixel positions and segmented points. The term "coverage" refers to the method used to determine the width and length of a tumor. On the other hand, the coverage characteristics of dimension [2x1] are stated as,

$$g_4 = \sum_{k=1}^Q \frac{w_j}{C_{kq}^*} \tag{3.13}$$

3.5. Fitness Functions. The fitness function is employed to determine the optimal selection of female chicks, hens, and male roosters based on chicken swam groupings. The chicken with the highest fitness grade is classified as a rooster, and each rooster is the group's leader. Chickens classified as chicks have a low fitness value, while hens have a high fitness value.

$$Qg = \frac{1}{d} \sum_{y=1}^d P_d - \varsigma \tag{3.14}$$

Initialization of the population lets us begin by populating the populace with D roosters, Y chicks, S hens, and Z mother hens.

3.6. Fitness Evaluations. It is analyzed to determine the ideal chickens, as indicated in Equation (9). We've updated the rooster's location. The roosters with the most significant fitness values receive priority over the hens with the lowest fitness values when it comes to eating. On the other hand, the chickens keep to the group's rooster in quest of food. However, they protect their food from other chickens. Birds randomly take food from other chickens. Simply put, roosters with a higher fitness value seek food in more locations than those with a lower fitness value. As a result, the corrected rooster equation 3.14 to 3.17 is as follows:

$$h_{x,y}^r = h_{x,y}^r \times (1 + \delta(0, l^2)) \tag{3.15}$$

$$h_{x,y}^{r+1} = h_{x,y}^r + h_{x,y}^r \delta(0, l^2) \tag{3.16}$$

$$h_{x,y}^{r+1} - h_{x,y}^r = h_{x,y}^r \delta(0, l^2) \tag{3.17}$$

where $\delta(0, l^2)$ specifies with the mean 0 and standard deviation l^2 for Gaussian distribution.

$$U^\beta [h_{x,y}^{r+1}] = h_{x,y}^r \delta(0, l^2) \tag{3.18}$$

$$h_{x,y}^{r+1} = h_{x,y}^r (\beta + \delta(0, l^2)) + \frac{1}{2} \beta h_{x,y}^{r-1} + \frac{1}{6} (1 - \beta) h_{x,y}^{r-2} + \frac{1}{24} \beta (1 - \beta) (2 - \beta) h_{x,y}^{r-3} \tag{3.19}$$

In this context, let y represent the most minor constant and m be the rooster index., p is the fitness value of the associated g order between [0, 1].

$$h_{x,y} = \left\{ 1, \exp\left(\frac{qm - qlxx}{qx + y}\right), \text{otherwise}^{if qx \leq qmim \in [1, B], m \neq y} \right\} \tag{3.20}$$

The location of Hen has been changed. Hen accompanies their groupmate rooster on his food seek. By and large, the Hen snatches food that other chickens randomly discover. In comparison to submissive hens, dominant hens have an advantage. As a result, the equation for the hen update is as follows: and shown in eq 3.18 to 3.22.

$$r + 1_{x,y} = h_{x,y}^r + K1 \times l \times (h_{xm,y}^r - h_{x,y}^r) + K2 \times l \times (h_{x2,y}^r - h_{x,y}^r) \tag{3.21}$$

where the term k1 and k2 are specified as,

$$k1 = \exp\left(\frac{(q_x - q_{h1})}{bs(q_x) + y}\right) \tag{3.22}$$

$$k1 = \exp(q_{h2} - q_x) \quad (3.23)$$

where k is a number between 0 and 1, the x th hen group's mate is identified by the rooster index $h1$. Furthermore, the variable $h2$ represents the chicken index of the swarm, indicating whether the individual is a hen or a rooster, so $h2$, $h1$, and $h1$ all equal $h2$. Changes in chick position These dominant chicks have an advantage over other chicks regarding food completion. As a result, as illustrated in the equation below, the chicks travel about the mother looking for food.

$$h_{x,y}^r + l \times (h_{u,y}^r - h_{x,y}^r) \quad (3.24)$$

where $h_{u,y}$ is the position of the mother of the x th chick can be represented as u , where u is an element of the interval $u \in [1, A]$. The parameter L , which lies within the interval $L \in [0, 2]$, indicates whether the chick forages with its mother. As a result, each chick's parameter L randomly selects values between 0 and 2 $h(r+1)_{(x,y)}$ shown in eq 3.23 and 3.24

3.7. Pseudocode for Classification.

Algorithm – RCNN Based Brain Tumor Classification. Input: Feature set (set of optimize feature and the area of the tumor)

Output: Decision on the tumor type

- 1: Initialize weight vector $Q = 0$;
- 2: Select data points in the coordinate of B_i, C_i
- 3: if Selected vector x_i if misclassified then
- 4: Select $Q \leftarrow B + \text{sign}(f(B_i, C_i))(B_i, C_i)$
- 5: else
- 6: Repeat from step 2 until all the data are classified correctly
- 7: end if
- 8: After convergence or classification of data select Equation

3.8. Classification Algorithms Process.

The process of the algorithm is as follows:

- Step 1: Initially, Convert medical images of the brain tumor to the DICOM file format.
- Step 2: Deal with medical images of a size similar to the brain. Differentiate normal from abnormal MRI pictures.
- Step 3: Process the images using CSO feature extraction optimization by taking mathematical derivation using equations from 3.1 and 3.12.
- Step 4: Calculate the elapsed time and use equation 3 to find the best-produced image.
- Step 5: The impacted region can readily be separated by selecting the best-generated image and using the appropriate filtering algorithms.

4. Results and Discussion.

4.1. Data Collection. The collection of Dataset from the data underlying this article is available in BRATS 2015 at <https://www.smir.ch/BRATS/Start2015> and <https://github.com/topics/brats>, respectively. Meningiomas, gliomas, and pituitary gland tumors are all in this group [23]. Rembrandt [24] includes data from 874 tumor examples from the glioma molecular diagnostic effort. This includes 566 X 566 gene expression arrays, 834 arrays of copy number data, and 13,472 points of clinical phenotype. Many deep learning-DL methods exist to find and classify brain tumors [25-28].

4.2. Experimental Setup. To implement the proposed methodology, we used the Google Colab environment. The following are the configuration details for the simulation trials that were used:

- Implementation environment and packages
- Google Colab
- Tensor flow
- Numpy
- Pandas, matplotlib
- Input Dataset : BRATS 2015

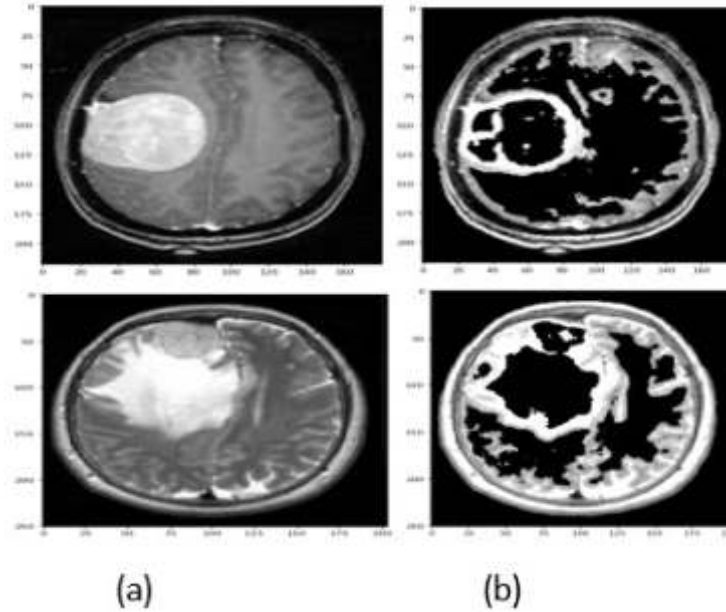


Fig. 4.1: (a) Input images (b) Optimized images

Table 5.1: Precision Analysis of RCNN technique with existing methods.

Method or % of training data	40 %	50 %	60 %	70 %	80 %
RCNN	91.23	92.42	93.63	92.38	92.11
CNN	90.56	90.89	91.34	91.63	91.73
K SVM	87.84	87.98	88.21	88.53	89.11
SVM	84.94	85.18	86.01	87.21	88.93
RF	83.71	83.95	85.47	86.76	87.9

Figure 4.1 shows the optimized brain images with clear identification of tumors. Figure 4.1(a) shows the MRI brain images with cancer, and Figure 4.1 (b) shows the optimized image after applying the CSO.

5. Performance Validation.

5.1. Precision. The precision analysis of the RCNN approach with existing techniques is shown in Table 5.1 and Figure 5.1. The results prove that the RCNN approach has improved precision with all training datasets. For example, the RCNN technique with a 40% dataset has achieved a superior precision of 91.23%, whereas the CNN, K SVM, SVM, and RF methods have achieved lesser precisions of 90.56%, 87.84%, 84.94%, and 83.71%, respectively. Similarly, using an 80% dataset, the RCNN approach achieved the highest precision of 92.11%, while the CNN, K SVM, SVM, and RF procedures achieved lower precision of 91.73%, 89.11%, 88.93%, and 87.90%, respectively.

5.2. Sensitivity. The sensitivity analysis of the RCNN approach with existing techniques is shown in Table 5.2 and Figure 5.2. The results prove that the RCNN approach has an improved sensitivity with all training datasets.

5.3. Specificity. Table 5.3 shows the Specificity of different machine learning algorithms on an image recognition task when trained on different percentages of the training data. The table shows the following: The method or percentage of training data used to train the model The Specificity of a RCNN model trained on 40



Fig. 5.1: Precision Analysis of RCNN technique with existing methods

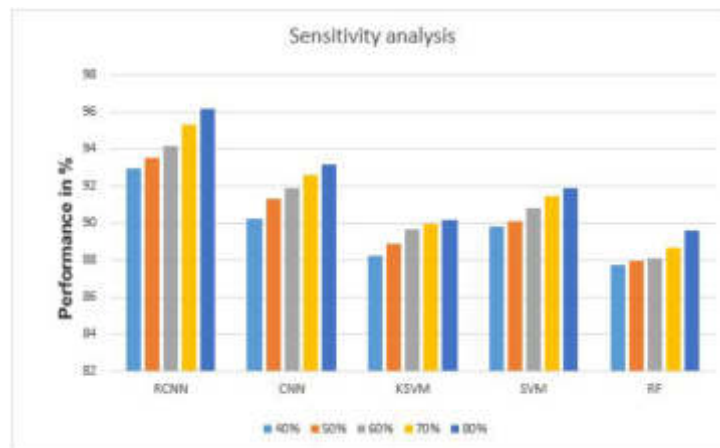


Fig. 5.2: Sensitivity Analysis of RCNN technique with other existing methods.

%, 50 %, 60 %, 70 %, and 80 % of the training data The Specificity of a CNN model trained on 40 %, 50 %, 60 %, 70 %, and 80 % of the training data The Specificity of a KSVM model trained on 40 %, 50 %, 60 %, 70 %, and 80 % of the training data The Specificity of a SVM model trained on 40 %, 50 %, 60 %, 70 %, and 80 % of the training data The Specificity of a RF model trained on 40 %, 50 %, 60 %, 70 %, and 80 % of the training data In all cases, the Specificity of the model increases as the percentage of training data used increases. This is because the model has more data to learn from, which allows it to better generalize to unseen data. The RCNN model achieves the highest Specificity overall, with a Specificity of 96.43 % when trained on 80 % of the training data. The CNN model is a close second, with a Specificity of 94.28 % when trained on 80 % of the training data. The KSVM, SVM, and RF models all have lower Specificity, but still achieve good performance. It is important to note that this table only shows the results for one image recognition task. The Specificity of these models would likely vary depending on the specific task and dataset. However, the table does illustrate the general trend that machine learning models tend to perform better when trained on more data. Here are some additional things to consider when interpreting this table: The specific image recognition task that the models were trained on is not shown in the table. This could be an important factor in how well the models perform. The size and quality of the training data is not shown in the table. These factors can also affect the

Table 5.2: Analysis of how sensitive the RCNN method is to current methods.

Method or % of training data	40 %	50 %	60 %	70 %	80 %
RCNN	92.92	93.49	94.18	95.27	96.15
CNN	90.23	91.33	91.89	92.61	93.18
K SVM	88.21	88.89	89.64	89.95	90.12
SVM	89.81	90.11	90.83	91.41	91.87
RF	87.73	87.94	88.11	88.67	89.56

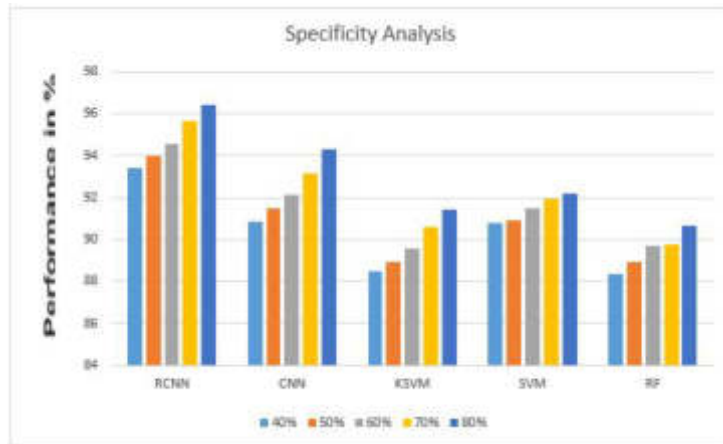


Fig. 5.3: Analysis of how sensitive the RCNN method is to current methods

performance of machine learning models. The hyper parameters of the machine learning models are not shown in the table. Hyper parameters are settings that can be tuned to improve the performance of a model.

The specificity analysis of the RCNN approach with existing techniques is shown in Table 5.3 and Figure 5.3. The results prove that the RCNN approach has an improved specificity with all training datasets.

5.4. Accuracy. The accuracy analysis of the RCNN approach with existing techniques is shown in Table 5.4 and Figure 5.4. The results prove that the RCNN approach has improved accuracy with all training datasets.

5.5. Overall Accuracy Analysis. The overall accuracy analysis of the RCNN approach with existing techniques is shown in Figure 5.5. The result depicts that the RCNN approach has improved overall accuracy with all training datasets, where the RCNN technique has achieved a superior accuracy of 95.78%, while the CNN, K SVM, SVM, and RF methods have achieved lesser accuracy of 93.23%, 92.67%, 91.14%, and 92.93%, respectively.

5.6. TPR and FPR Analysis. The True Positive Rate and False Positive Rate Analysis of the RCNN Method with Current Methods is shown in Figure 5.6. The results prove that the RCNN approach has an improved TPR with all training datasets. The complete analysis is shown in the table Table 5.3 and 5.4 as figure 5.6.

It is a medical image segmentation competition where researchers develop algorithms to automatically segment brain tumors from magnetic resonance imaging (MRI) scans. In the challenge, participants submit models that are evaluated on a held-out test set. The models are evaluated on how well they can segment four different regions of the brain tumor: the whole tumor (WT), the tumor core (TC), the enhancing tumor (ET), and the necrotic and non-enhancing tumor (NET). The bar graph in the image shows the average accuracy of each model across all four subregions. As you can see, the average accuracy of a model is about 95 %, while the average accuracy of another model is about 70 %. Overall, the accuracy of the models ranges from about 60

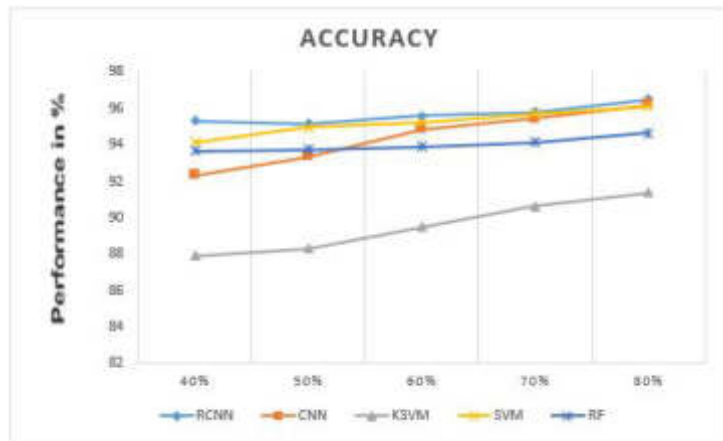


Fig. 5.4: RCNN technique’s accuracy compared to current methods.

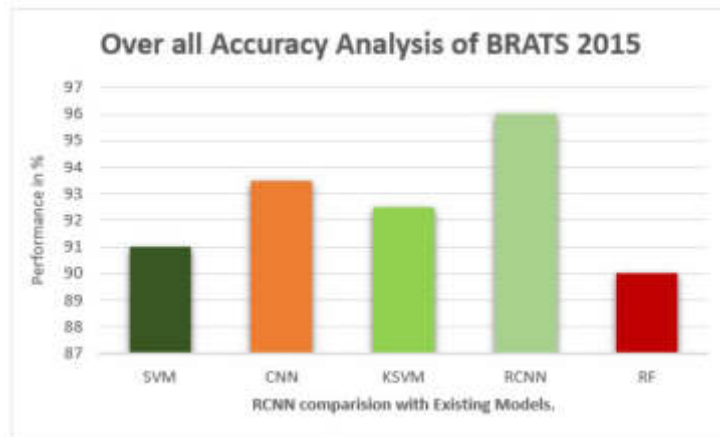


Fig. 5.5: Overall Accuracy Analysis of BRATS 2015.

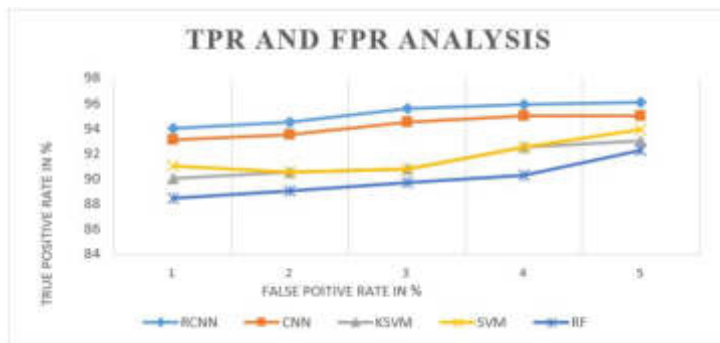


Fig. 5.6: TPR and FPR Analysis of RCNN technique with existing methods.

Table 5.3: Specificity analysis of the RCNN approach with current methods.

Method or % of training data	40 %	50 %	60 %	70 %	80 %
RCNN	93.42	93.99	94.58	95.67	96.43
CNN	90.83	91.47	92.14	93.17	94.28
KSVM	88.46	88.94	89.55	90.59	91.42
SVM	90.77	90.92	91.46	91.92	92.17
RF	88.34	88.92	89.71	89.75	90.68

Table 5.4: RCNN technique's accuracy compared to current methods.

Method or % of training data	40 %	50 %	60 %	70 %	80 %
RCNN	95.29	92.33	87.91	94.13	93.66
CNN	95.14	93.33	88.31	94.96	93.75
KSVM	95.58	94.82	89.47	95.21	93.9
SVM	95.77	95.41	90.63	95.71	94.14
RF	96.49	96.19	91.36	96.1	94.62

% to 97 %. Here are some additional things to consider when interpreting this bar graph: The specific models being compared are not shown in the image. The BRATS challenge is a competition, so the models shown in the graph are likely to be state-of-the-art models. However, it is important to remember that machine learning is an active area of research, and new models are being developed all the time shown in figure 5.4. The accuracy of a brain tumor segmentation model is just one measure of its performance. Other important factors include the sensitivity and specificity of the model. Sensitivity is the ability of the model to correctly identify people who have a brain tumor. Specificity is the ability of the model to correctly identify people who do not have a brain tumor shown in figure 5.5 and figure 5.6.

6. Conclusion. The proposed RCNN-CSO approach is an effective and automated way of classifying and segmenting brain tumors. The MRI pictures are enhanced by applying the Adaptive Fuzzy Filtering (AFF) approach. MRI images are segmented using a technique called chicken swarm optimization (CSO), which uses Tsallis uses entropy-based picture segments to find the damaged parts of the brain. The goal here is to examine the RCNN method, which boasts better results. A thorough experimental analysis is performed, and Several ways are used to look at the information. The overall accuracy of the RCNN technique compared to the other methods is 95.78%. The simulation results demonstrated that the suggested RCNN strategy outperformed the existing methods. The presented approaches in this paper can be used in the future to examine the influence of strokes caused by tumors on brain imaging.

REFERENCES

- [1] WAKTOLA, S., BIEBERLE, A., BARTHEL, F., BIEBERLE, M., HAMPEL, U., GRUDZIEŃ, K., & BABOUT, L. (2018) , *A new data-processing approach to study particle motion using ultrafast X-ray tomography scanner: case study of gravitational mass flow*. Experiments in Fluids, 59, 1-14.
- [2] VASANTHARAJ, A., RANI, P. S., HUQUE, S., RAGHURAM, K. S., GANESHKUMAR, R., & SHAFI, S. N. (2023), *Automated brain imaging diagnosis and classification model using rat swarm optimization with deep learning based capsule network*. International Journal of Image and Graphics, 23(03), 2240001.
- [3] MENG, X., LIU, Y., GAO, X., & ZHANG, H. (2014), *A new bio-inspired algorithm: chicken swarm optimization*. In Advances in Swarm Intelligence: 5th International Conference, ICSI 2014, Hefei, China, October 17-20, 2014, Proceedings, Part I 5 (pp. 86-94). Springer International Publishing.
- [4] SHIVAPRASAD, B. J., RAVIKUMAR, M., & GURU, D. S. (2022), *Analysis of Brain Tumor Using MR Images: A Brief Survey*. International Journal of Image and Graphics, 22(02), 2250023.
- [5] DEMIRHAN, A., TÖRÜ, M., & GÜLER, I. (2014) , *Segmentation of tumor and edema along with healthy tissues of brain using wavelets and neural networks*. IEEE journal of biomedical and health informatics, 19(4), 1451-1458.

- [6] SRINIVAS, B., & RAO, G. S. (2018, JANUARY), *Unsupervised learning algorithms for MRI brain tumor segmentation*. In 2018 Conference on Signal Processing And Communication Engineering Systems (SPACES) (pp. 181-184). IEEE.
- [7] BAL, A., BANERJEE, M., CHAKRABARTI, A., & SHARMA, P. (2022), *MRI brain tumor segmentation and analysis using rough-fuzzy c-means and shape based properties*. Journal of King Saud University-Computer and Information Sciences, 34(2), 115-133.
- [8] MANOCHA, P., BHASME, S., GUPTA, T., PANIGRAHI, B. K., & GANDHI, T. K. (2017), *Automated tumor segmentation and brain mapping for the tumor area*. arXiv preprint arXiv:1710.11121.
- [9] VARUNA SHREE, N., & KUMAR, T. N. R. (2018) , *Identification and classification of brain tumor MRI images with feature extraction using DWT and probabilistic neural network*. Brain informatics, 5(1), 23-30.
- [10] TOMICKA, J., CICHON, K., CHLEWICKI, W., HOLICKI, M., PELC, M., ZYGARLICKI, J., PODPORA, M. & KAWALA-STERNIUK, A., 2022 , *Pilot Study on Application for Analysis of Magnetic Resonance Spectroscopy Spectra*. IFAC-PapersOnLine, 55(4), pp.45-50.
- [11] KUMAR, S. & MANKAME, D. P. (2020) , *Optimization driven deep convolution neural network for brain tumor classification*. Biocybernetics and Biomedical Engineering, 40(3), 1190-1204.
- [12] D. RAMMURTHY, & P.K. MAHESH, 2022. , *Whale Harris Hawks Optimization based deep learning classifier for brain tumor detection using MRI images*. Journal of King Saud University – Computer and Information Sciences, 34, 3259-3272
- [13] D. DEB & S. ROY, 2021 , *Brain tumor detection based on hybrid deep neural network in MRI by adaptive squirrel search optimization*, *Multimed. Tools*. 80 (2) 2621–2645.
- [14] YIN, B., WANG, C., & ABZA, F. (2020) , *New brain tumor classification method based on an improved version of whale optimization algorithm*. Biomedical Signal Processing and Control, 56, 101728.
- [15] SULTAN, H. H., SALEEM, N. M., & AL-ATABANY, W. (2019), *Multi-classification of brain tumor images using deep neural network*. IEEE access, 7, 69215-69225.
- [16] PANDISELVI, T., & MAHESWARAN, R. (2019) , *Efficient framework for identifying, locating, detecting and classifying MRI brain tumor in MRI images*. Journal of medical systems, 43, 1-14.
- [17] DEEPAK, S., & AMEER, P. M. (2021) , *Automated categorization of brain tumor from mri using cnn features and svm*. Journal of Ambient Intelligence and Humanized Computing, 12, 8357-8369.
- [18] NARMATHA, C., ELJACK, S. M., TUKA, A. A. R. M., MANIMURUGAN, S., & MUSTAFA, M. (2020) , *A hybrid fuzzy brain-storm optimization algorithm for the classification of brain tumor MRI images*. Journal of ambient intelligence and humanized computing, 1-9.
- [19] POLEPAKA, S., RAO, C. S., & CHANDRA MOHAN, M. (2020) , *IDSS-based two stage classification of brain tumor using SVM*. Health and Technology, 10(1), 249-258.
- [20] MENZE, B. H., JAKAB, A., BAUER, S., KALPATHY-CRAMER, J., FARAHANI, K., KIRBY, J., ... & VAN LEEMPUT, K. (2014) , *The multimodal brain tumor image segmentation benchmark (BRATS)*. IEEE transactions on medical imaging, 34(10), 1993-2024.
- [21] CRISTIN, D. R., KUMAR, D. K. S., & ANBHAZHAGAN, D. P. (2021) , *Severity level classification of brain Tumor based on MRI images using fractional-chicken swarm optimization algorithm*. The Computer Journal, 64(10), 1514-1530.
- [22] PAUL, J. S., PLASSARD, A. J., LANDMAN, B. A., & FABBRI, D. (2017, MARCH), *Deep learning for brain tumor classification*. In Medical Imaging 2017: Biomedical Applications in Molecular, Structural, and Functional Imaging (Vol. 10137, pp. 253-268). SPIE.
- [23] CHENG, J., HUANG, W., CAO, S., YANG, R., YANG, W., YUN, Z., ... & FENG, Q. (2015) , *Enhanced performance of brain tumor classification via tumor region augmentation and partition*. PloS one, 10(10), e0140381.
- [24] SAYAH, A., BENCHEQROUN, C., BHUVANESHWAR, K., BELOUALI, A., BAKAS, S., SAKO, C., ... & GUSEV, Y. (2022). , *Enhancing the REMBRANDT MRI collection with expert segmentation labels and quantitative radiomic features*. Scientific Data, 9(1), 338.
- [25] NOREEN, N., PALANIAPPAN, S., QAYYUM, A., AHMAD, I., IMRAN, M., & SHOAI, M. (2020) , *A deep learning model based on concatenation approach for the diagnosis of brain tumor*. IEEE Access, 8, 55135-55144.

Edited by: Polinpapilinho F. Katina

Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions

Received: Jan 7, 2024

Accepted: Mar 23, 2024



IMPROVING DATA SECURITY AND SCALABILITY IN HEALTHCARE SYSTEM USING BLOCKCHAIN TECHNOLOGY

K.R.ROHINI *, DR.P.S.RAJAKUMAR † AND DR.S.GEETHA ‡

Abstract. The wearable tech revolution and the proliferation of Internet of Things (IoT) gadgets have created exciting new opportunities for remote patient monitoring. Healthcare providers are increasingly utilizing wearable technologies to expedite the process of diagnosis and treatment. The healthcare and research fields have been substantially impacted by emerging technologies. On the other hand, there are legitimate worries regarding the security of data transfers and transaction recording upon using these technologies. Healthcare departments need easy data interchange for interoperability. Protecting data security and integrity is crucial when exchanging information with authorized parties. In many existing solutions, patients' sensitive data is collected and stored in smart healthcare systems. Scalability, breach or unauthorized access to this data can compromise privacy. The adoption of blockchain technology is one way to safeguard patient privacy in healthcare. Also, by facilitating the safe and secure exchange of data, blockchain technology is revolutionizing the healthcare industry by making traditional methods of diagnosis and treatment more reliable. However, blockchain has serious problems with its highly limited scalability. This article suggests a new method "SHORTBLOCKS" depending on blockchain technology for the safe administration and analysis of data. To circumvent the security-scalability issue and achieve high throughput, this study employs the newly introduced protocol "SHORTBLOCKS", which extends blockchain concept to a direct acyclic graph of blocks. The proposed system uses both a private and a public blockchain that are created on the new protocol. In order to analyze patient health data, a system using smart contracts and a private blockchain is built. The system logs the occurrence to the public blockchain in the case that the smart contract raises an alarm due to an unusual reading. This will fix the scalability issue with initial blockchain as well as the security and privacy issues involved in remote patient monitoring. Simulation results shows the performance of the proposed system by comparing with existing solutions, SPECTRE protocol and GHOSTDAG protocol.

Key words: Healthcare, Remote patient monitoring, Security and privacy, interoperability, Private blockchain, Public blockchain, Smart contracts, Shortblocks.

1. Introduction. Clinical data is abundant in the healthcare industry due to the regular generation, access, and dissemination of enormous volumes of data. Due to privacy and security concerns, as well as the sensitive nature of the data, storing and distributing this enormous amount of data is not only necessary but also extremely challenging [1].

A wide variety of resources are available to various healthcare providers, including hospitals, pharmacies, and insurance providers [2]. The rapid development of ICT has allowed for the widespread use of wearable sensors to link many physical objects. Data processing, collecting, and creation on a massive scale are made possible by the electrical devices and software employed for this purpose. Although there are many healthcare facilities, the prevalence of lethal diseases such as pneumonia, influenza, cancer, heart diseases, etc. has risen dramatically [3]. Constant vigilance and tracking of vital signs has been made possible with the use of sensing devices that can gather, and analyze patient data, due to the technological revolution. Hospitals and other related departments receive this data for the purpose of delivering healthcare. Problems with legal interoperability are present here. Healthcare service delivery will be more efficient with a centralized system for managing and exchanging medical data [4].

In healthcare and clinical settings, safe, secure, and scalable (SSS) data-sharing is critically necessary for diagnosis and integrated clinical decision making. The ability for medical professionals to transmit patient

*Research Scholar, Dept. of CSE, Dr. M.G.R. Educational and Research Institute, Maduravoyal, Chennai, Tamil Nadu, India (rohinimay19@gmail.com)

†Professor, Dept. of CSE, , Dr. M.G.R. Educational and Research Institute, Maduravoyal, Chennai, Tamil Nadu, India. (*rajakumar.subramanian@drmgrdu.ac.in)

‡Professor, Dept. of CSE, , Dr. M.G.R. Educational and Research Institute, Maduravoyal, Chennai, Tamil Nadu, India. (*somangeetha@drmgrdu.ac.in)

medical records to the relevant authority for expedited follow-up makes sharing of data, a crucial activity. Both primary care physicians and other medical professionals should have the ability to securely and quickly transmit their patients' clinical data to one another, so that everyone involved has access to the most current and accurate information regarding their patients' health. During treatment, doctors and hospitals need access to patients' medical records, but individuals also have a right to feel safe sharing their personal information with these institutions.

Conversely, e-health and tele-medicine are two popular areas where patients can get their clinical data reviewed by a specialist located far away. A "store-and-forward technology" or online healthcare monitoring in real-time (e.g., tele-monitoring, telemetry, and the like) is used to transmit patient data in these two online clinical setups [5]. Clinical specialists are able to remotely perform patient diagnosis and treatment in these online clinical contexts by exchanging clinical data. Because patient data is unique to each individual case, ensuring its confidentiality, integrity, and availability in all such therapeutic arrangements is a top priority. In order to facilitate relevant and healthy therapeutic discussions pertaining to instances involving faraway patients, the capacity to securely and scalably transmit data is of the utmost importance. For the simple reason that better diagnosis and more efficient treatment are the outcomes of secure data transmission, which aids in clinical communication by soliciting opinions or confirmations from a group of clinical experts [6].

Additionally, several interoperability problems crop up frequently in this domain. The secure and reliable transfer of patient records between hospitals and universities, for instance, can be fraught with practical difficulties. Such exchanges of clinical data need considerable, trustworthy, and healthy collaboration between the entities involved. Difficult patient matching algorithms, ethical policies, legislation, and processes; sensitive clinical data; and data sharing agreements are all potential roadblocks. Prior to establishing a clinical data exchange, it is essential that all stakeholders agree on these and other important matters [7].

The proliferation of remote devices linked to the internet for data and access transfer has been greatly enhanced by the persistent developments in the IoT domain. Thus, from the field of education to that of supply chain management, the Internet of Things (IoT) has transformed and shaken up nearly every industry on the planet. Internet of Things (IoT) has also shown to be highly effective in the healthcare industry, streamlining diagnostic processes and effectively tracking patients' activities. In addition, the main point we want to make about the Internet of Things is that it helps with patient monitoring even when the patient isn't actively using their devices, which may be a real challenge with the current system. Additionally, there are vast opportunities for improved diagnostic and treatment efficiency that arise from remote access to data and ongoing analysis of it.

Blockchain applications have recently garnered a lot of interest for safe healthcare data transmission [8], biological data sharing [9], e-health data sharing [10], and general activities. A peer-to-peer (P2P) network is the foundation of a blockchain. In essence, it is a cryptographic, algorithmic, and mathematically expressed multi-field network architecture that aims to address the constraints of standard distributed database synchronization methods through the use of distributed consensus techniques.

With the introduction of blockchain, the technology recently assumed full control of all storage, transactions, and access. Additionally, blockchain technology has demonstrated tremendous promise in a variety of industries, that includes retail, healthcare, supply-chain management, finance, and more. The privacy and security of data is a constant concern in healthcare because many parties rely on it to carry out their own objectives. For instance, insurance companies play a crucial role in the delivery of a specific service to patients; in order to do their jobs properly, they need access to patients' data for analysis and service planning. Unfortunately, it is not uncommon for companies to manipulate or even leak this data. Consequently, this system is an authentic answer to the problem of data misuse and to the problem of keeping the many stakeholders' faith in one another. With the use of Blockchain technology, healthcare organizations may eliminate the need for reconciliation, rework, and repetition. We can find the optimal level of service intensity in a predictive healthcare setting with the help of artificial intelligence (AI). It is important to communicate with patients, professionals, and careers in a way that is appropriate for each setting.

Consider a case study pertaining to the COVID-19 case as an example. When it comes to COVID-19-safe clinical practice, blockchain technology is crucial. The integration of Blockchain technology with AI has the ability to speed up the process of diagnosing and treating COVID-19 patients, while also helping to create

therapeutic guidelines for future outbreaks that could be similar to coronavirus.

Blockchain technology allows for the secure and private sharing of data from many sources, including hospitals, primary care physicians, pediatricians, clinical laboratories, and others. Artificial intelligence solutions are utilized for data analysis.

Take the COVID-19 case study as an example. Cryptography is the backbone of Covid-19-safe medical practice. Rapid diagnosis and treatment of COVID-19 patients is possible with the use of Blockchain and AI, which also assists in developing treatment recommendations for future outbreaks (like coronavirus). With a blockchain-based system, pediatricians, primary care physicians, hospitals, and clinical laboratories may all securely share patient data. Data analysis makes use of AI solutions. The paradigm promotes research into appropriate medicines, helps with risk management, and inspires the development of new drugs.

The use of Internet of Things and Blockchain technologies for secure data transmission in smart health care systems is proved to be effective. However, with the concept of blockchain, scalability problem arises. Many existing solutions are not addressing properly the problem of scalability and secure communication in smart health care system. Hence, a sophisticated method is required to console the problem of scalability while transferring the data security using Internet of Things and Blockchain in Smart health care Systems.

This study proposes a novel approach called "SHORTBLOCKS" that incorporates the blockchain concept. Its purpose is to offer scalable and secure communication in the healthcare domain using blockchain as a paradigm. Here is the breakdown of the remaining sections of the paper: Section 2 provides a synopsis of relevant background material and related work in the healthcare domain that makes use of blockchain technology. In Section 3, the proposed method's framework is described in terms of how healthcare and the medical industry might use blockchain technology. Section 4 lays out the strategy that has been suggested. The suggested protocol is demonstrated in Section 5. In Section 6, we go over the findings from comparing the current state of blockchain technology with the suggested technique "SHORTBLOCKS" that incorporates blockchain technology. In Section 7, we lay out the final thoughts, and in Section 8, we look forward to the potential applications of blockchain technology in healthcare, and finally, we provide a list of references.

2. Related Work. The paper's primary rationale for investigating blockchain technology's potential use in healthcare is based on the work of [11], who comprehensively outlined several current tendencies in this field of study. Since Bitcoin's inception in [12], the potential applications of the underlying technology have been practically limitless, even beyond the financial sector.

Public blockchains are decentralized database systems that let anyone with an internet connection to determine who may access the data stored on them. Ivan [13] also showed how to encrypt health data using this method. This approach creates a PHR (personal health record) based on blockchain by publicly storing encrypted healthcare data. The patients were able to gain improved access to their clinical data through the way they suggested. Patients can now freely observe and manage their information, as well as take part in maintaining their data and making it available to any associated healthcare provider. Another study by Chen et al. [14] suggested a system for managing and sharing patients' confidential medical data that incorporated blockchain technology with cloud storage. Ensuring the secure storage and communication of personal medical data is a potential use case for the suggested approach. The proposed method is novel because it eliminates the need for a middleman by providing patients full access to and management of their own medical records.

In order to streamline the administration of electronic health record transactions, Dey et al. [15] suggested a blockchain-based Internet of Things model. As a principal agent for linking the bio-sensors to the IoT platform, the architecture suggested using the MQTT protocol. In addition, the design included the IPFS (InterPlanetary File System), which may identify state entries or block modifications caused by certain transactions appended to blocks in order to decrease stored transaction deduplication. In [16], Gem, a provider of enterprise blockchain solutions, announced that Philips Blockchain Lab, a research and development centre of healthcare massive Philips, would be the first major healthcare operator to join the Gem Health network. Gem Health is an Ethereum-powered network for developing healthcare applications and shared infrastructure.

To assess the current state of patient healthcare, Wang et al. [17] laid forth a blockchain architecture that uses AI healthcare systems and parallel execution instead. In order to aid in clinical decision-making, the suggested method evaluates the patient's general status, diagnosis, and treatment plan while also analyzing the related therapeutic processes carrying out computational trials in parallel. To assess the precision of diagnosis

and efficacy of therapy, the suggested approach has undergone testing on real and virtual healthcare networks.

The use of blockchain technology to verify the achievement of clinical trial endpoints was introduced in [18]. The method was put to the test by Irving and Holden using a clinical trial methodology that had previously documented result switching. Scientific research' credibility and the application of blockchain technology as an inexpensive, independently verifiable auditing tool were both validated.

A healthcare system controlled by the Internet of Things was suggested by Budida et al. [19] in an interactive environment. The suggested design is based on generative data ingestion from biosensors and smart wearables, which then provides patients with clear feedback and simple solutions. In order to provide patients with quicker and more accurate recommendations, A Smart Hospital system, proposed by Sivagami et al. [20], would combine human reaction suggestions with the efficiency of sensors. The proposal proposes a system that uses radio frequency identification (RFID), wireless sensor networks (WSN), and smart wearables to accomplish a number of goals. These goals include smart sensing of the patient's environment, ward allocation according to doctor placement requirements, movement monitoring, and report analysis based on calculated data, after data has been uploaded by the sensors.

An innovative platform for the exchange of healthcare information, BloCHIE was created by Jiang et al. [21]. Using blockchain technology embedded in many sources, the proposed platform assesses the needs for healthcare data exchange, particularly with regard to electronic medical records and personal healthcare data, but also handles a wide range of other data types. They integrated on-chain and off-chain verification procedures into the platform to ensure it met the desired standards of privacy and authenticity. All parties involved in the healthcare system, from payers to providers to patients, stand to benefit from blockchain technology's revolutionary potential in the areas highlighted by Nichol in [22]. In order to transform healthcare, Nichol amplifies the principles and applies examples that are at the forefront of a new frontier. Written in the style of his articles, blogs, and thoughts, the book lays forth the groundwork for an internal revolution in healthcare.

Discussing the foundations of all the different kinds of blockchains and how they might be utilized in the healthcare industry for data maintenance, validation, and storage. The various blockchain designs that are currently available were discussed by Zainab Alhadhrami et al. [23]. Furthermore, the consortium blockchain was found to be the solution that was essentially pre-built for healthcare data storage. To put it simply, a consortium blockchain is a valid blockchain in which the node owner and miners share control of access. Furthermore, the operation of the consortium blockchain is predicated on the consensus principle, which requires the approval of the majority of stakeholders or blockchain nodes.

Clinical professionals and healthcare entities can significantly enhance medical data sharing, privacy, and security by utilizing blockchain technology. In a similar vein, Cryan, M.A. [24] suggested a novel and methodical design based on blockchain technology to secure private patient information, resolve pressing data security concerns, and integrate blockchain software into an entire healthcare organization's infrastructure.

In order to address the privacy requirement imposed by HIPAA, Ahram et al. [25] created a healthcare blockchain that controlled the access to patients' demographic and racial information. The study also demonstrated the generative design of a blockchain network that included three distinct kinds of nodes: those for primary care physicians, referral services, and urgent care.

For the purpose of diagnosing and treating malignant tumors in distant patients, Shubbar introduced a healthcare framework based on blockchain technology in [26]. Patients' data at both specialty medical centers and their homes may be reliably and securely verified using the proposed protocol's use of smart contracts and blockchains. As a last point, Taylor unveils a new initiative in [27] that will utilize blockchain technology to strengthen the safety of the pharmaceutical distribution network. Still in its early phases of development, the initiative aims to streamline the process of tracing the origin and manufacturing date of medicines by utilizing blockchain monitoring and time stamps.

Healthcare applications serve a variety of objectives, including but not limited to pharmaceuticals, biomedical research, neurology, genomics, EHRs, clinical facts, processes, and decision-making [28]. By standardizing data and establishing communication protocols, Internet of Things (IoT) technology can improve healthcare efficiency. More effective healthcare services lead to less data interoperability, safer patient data, improved connectivity, and user interfaces. The healthcare industry and the development of trustworthy healthcare applications receive a lot of attention from researchers [29]. Ensuring that stakeholders, such as pharmacies,



Fig. 3.1: Smart Patient Health Care Monitoring

hospitals, and patients, can access healthcare records securely and without data tampering is of the utmost importance. It is possible to overcome these obstacles with blockchain technology [30].

Omar et al. (2019) [31] suggested a blockchain-based system for EMR maintenance. A patient-centric healthcare data management system was created by the authors. This system uses blockchain technology as a storage method to promote privacy. Under this system, patients will have full authority over their information. This method improves privacy, security, accountability, and integrity while increasing patient interest in EMRs.

In [32], GHOSTDAG protocol is employed which uses a public and a private blockchain. The authors proposed a method that utilizes a private blockchain, the system logs the event to the public blockchain for data transfer. The drawback of this system is privacy and security concerns surrounding remote patient monitoring. In [33], the use of blockchain technology to verify the achievement of clinical trial endpoints was introduced with SPECTRE protocol. By analyzing a clinical trial protocol that has previously documented outcome switching, Irving and Holden conducted an empirical evaluation of this strategy. They validated the validity of scientific research and the usage of blockchain technology as an inexpensive, independently verifiable auditing technique. However the security and privacy concerns remain still unanswered.

With the introduction of smart contracts, Blockchains can now store data that can be retrieved at will (Li et al., 2019) [34]. Nevertheless, smart contracts necessitate expert assistance for administration and maintenance and might be a pain to set up. Several scholarly articles present potential frameworks for health-related IoT applications. Due to the extra permission-based security they provide, private Blockchains form the basis of these concept designs. One such approach was to develop and manage a separate set of smart contracts for each Internet of Things (IoT) device.

3. Framework. Patients are finding it increasingly difficult to get an appointment with a primary care physician or other healthcare provider due to the massive growth in the patient population in many nations. The proliferation of Internet of Things (IoT) and wearable technology in the last several years has enhanced the care quality for patients using remote patient monitoring. More people can be treated by doctors because of this. Patients can be monitored and cared for outside of the typical clinical setting (such as at home) with smart patient monitoring. The primary benefit is the inherent ease of service it provides to patients. Patients have the option to remain in touch with their healthcare providers as needed. Additionally, it enhances the quality of care while decreasing medical expenses. This is the primary motivation for healthcare providers looking at ways to make remote patient monitoring accessible to everyone. A remote patient monitoring system could primarily consist of a smartphone, an app for remote patient monitoring, and a specifically developed device that can record and send health data to smart contracts as shown in Fig 3.1. When it comes to remote patient monitoring, wearable technology and the internet of things are crucial. For the purpose of health monitoring, illness diagnosis, and treatment, wearable devices transmit data collected from patients to healthcare facilities.

Smart electronic devices with microcontrollers that may be worn as accessories or implanted into garments are known as wearable gadgets in healthcare. They have sophisticated capabilities including built-in alerting systems, real-time feedback, and wireless data transmission, and they are easy to use and barely noticeable. A variety of vital signs, including respiration rate, blood sugar, and pressure, can be transmitted to medical professionals by means of these devices.

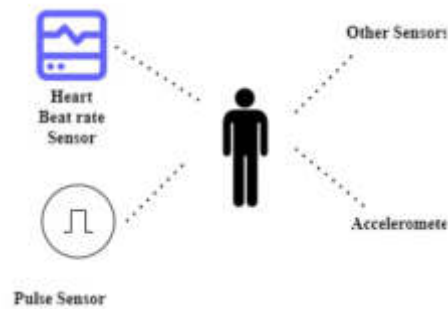


Fig. 3.2: Different Wearable Devices

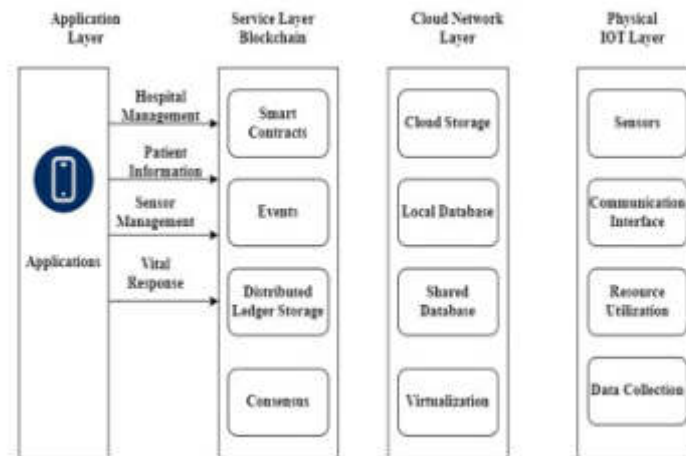


Fig. 3.3: Proposed Framework

Wearable devices are a part of the Internet of Things (IoT). These devices can connect to each other and share data through the software, electronics, sensors, actuators, and connectivity they contain as shown in Fig 3.2. Such infrastructure necessitates secure data sharing in order to manage such patient data with other institutions. The risk of exposure can increase if health data is shared, and the data is very sensitive. The second issue is that trust must be centralized in the present data sharing system because of its centralized architecture.

Blockchain technology may hold the key to ensuring the privacy and security of personal information. Blockchain technology ensures that data is secure and resistant to failures. In terms of data storage, it's a decentralized architecture. The healthcare system is the ideal fit for blockchain technology since data cannot be removed or altered from blocks. Nevertheless, blockchain's scalability and security remain unresolved. Blocks must propagate to all miners promptly for blockchain to be scalable and secure. Before the next block is formed, all honest nodes receive a message from the miner, who is responsible for maintaining the blocks, extending it to the blockchain. The scalability and processing performance issues are brought about by the proliferation of such lengthy blockchains in the existing approach. Consequently, we present a more sophisticated and extensible blockchain solution for the Remote Patient Monitoring system. For this purpose, we have adopted the SHORTBLOCKS protocol [33], a safe method of transaction confirmation that works at any network rate.

As an alternative to lengthy blockchains, SHORTBLOCKS organizes all blocks into a DAG, which is a Directed Acyclic Graph. The suggested healthcare system's design is shown in Fig. 3.3, where each layer represents the integration of multiple technologies. With the decoupled capability, developers can add or remove modules from the system as needed without impacting other modules or the system as a whole. Application, cloud-based network blockchain-based service, and Internet of Things (IoT) physical layers make up the suggested paradigm [34]. Computing healthcare equipment, data storage, and communication infrastructure make up the Internet of Things (IoT) layer. Connectivity, storage, a blockchain engine, and virtualization capabilities are all provided by the cloud-based network layer. Services such as consensus, identity management, distributed ledger technology (DLT), peer-to-peer (P2P) communication, and blockchain are provided by the blockchain-based service layer.

4. Proposed Methodology. The scalability and security of blockchain, however, remain unresolved. Fast block propagation to all miners is crucial to blockchain's scalability and security. Prior to the creation of the next block, any new blocks that are extended to the blockchain by a miner (the node responsible for maintaining them) are broadcast to all honest nodes. The issue of scalability and poor computational performance arises with the proliferation of such lengthy blockchains. As a result, we provide a blockchain system that is both more sophisticated and more scalable for use in smart health care system. The proposed method 'Smart Transfer' employs a new protocol, SHORTBLOCKS. It is a secure mechanism for transaction confirmation that can withstand any network throughput. The existing system uses traditional lengthy blocks whereas the proposed new protocol, SHORTBLOCKS integrates all blocks into a Directed Acyclic Graph.

As part of the proposed system, patients can access a variety of wearable medical equipment, including insulin pumps, blood pressure monitors, and more. A smartphone or tablet receives the health records and sends them to an application that formats and aggregates them. The finalized data, along with the specified threshold value, is transmitted to the appropriate smart contract on the private blockchain for thorough investigation. If the value is less than or equal to the threshold, the health reading is considered to be outside the normal range.

A public blockchain event will be created and alerted to smart devices and hospitals via the smart contract if the health reading is aberrant. Oracle [2] is one example of a smart contract that can talk to smartphones and other Oracle smart devices directly.

When an alert is issued, the public blockchain will only store the event. Certain Electronic Health Record (EHR) storage will receive the health data measured by wearable devices. Furthermore, electronic health record storage will receive treatment instructions from smart contracts or hospitals, and the blockchain will record the transactions. Integrating blockchain transactions with EHR ensures the integrity of patient medical records. As a result, it's easier to keep patient records in EHRs confidentially and to spot any unauthorized changes. Only designated nodes are able to execute smart contracts. Another responsibility of these selected nodes is to validate the new block. Care professionals, device manufacturers, and patients themselves are some of the potential viewers who could assist in limiting data exposure.

4.1. Components. The proposed system 'SHORTBLOCKS' consists of the following components:

- a) **Wearable Sensors:** In order to communicate medical history, people are increasingly turning to wearable health devices. Smartphones link these devices, and raw health data is sent to them. These cutting-edge gadgets will soon be able to detect heartbeats, breast cancer through an implant worn in clothing, and glucose readings without ever collecting blood.
- b) **Patient:** When a patient uses the system, it will record all of their medical history. Such information could include, among other things, heart rates, sleeping circumstances, and distance travelled. The onus for authorizing, rejecting, or revoking data access from third parties like insurance companies or healthcare providers should be on the patients themselves. After all, patients are the rightful owners of their personal data. In the event that medical attention is required, the patient will divulge their personal health information to the doctor of their choice. A patient has the right to cut off any communication with their healthcare provider, insurance company, or doctor once treatment has ended.
- c) **Health care Assistant:** Health organizations and insurance companies employ healthcare providers to conduct diagnostic tests and treatments. The healthcare provider can initiate a request for access to patient's medical records and treatment history. They treat them as soon as they receive the signal alert from the organization.

- d) Proposed SHORTBLOCKS: We employ two protocols based on blockchain technology: first, a private blockchain where all experiments involving patient health data are executed using smart contracts; second, a public blockchain where alerts are written and sent to smart devices and hospitals in the event that smart contracts issue them. These blockchains are built on proposed protocol.
- e) Healthcare organization: Any time a patient needs medical attention or would need a competitive price for future coverage, they can contact their health insurance provider. The insurance provider may request access to customer data, such as health data from wearable devices and medical treatment history, in order to offer the finest healthcare facilities. It is also possible to store insurance claim occurrences in the blockchain.

4.2. Requirements. The following are the essentials in the proposed method:

- Health or treatment data stored on the blockchain can only be accessed by authorized entities. Secure logging of Internet of Things (IoT) devices, such as smart contracts, is essential for maintaining an accurate timeline of events and safeguarding the integrity of patient care. The patient must grant access to their healthcare provider or insurance company before they can see their medical records or treatment information.
- Accurate and hacker-proof medical gadgets and health data are essential. Blockchain is the ideal technology for healthcare systems since it is impossible to remove or alter data from individual blocks. Unfortunately, blockchain's poor processing speed means it cannot be relied upon in isolation. To achieve this goal, we are utilizing the SHORTBLOCKS protocol, which outperforms the original blockchain technology in terms of speed and security.

4.3. Algorithm. Patient information is collected through wearable devices and the information is sent to smart applications for processing. The processed information is sent to smart contract for further examination using private blockchain. If the readings exceed the abnormal value, then alert is raised. The alert event is then stored in public blockchain and alert is sent to the patient.

Shorblocks method will work as

1. Patient information is collected from wearable devices implanted in patients body
2. Smart Devices placed in the hospitals will collect the patients data. This information is placed in private blockchain.
3. After full examination of the patients data, if any emergency found, an event is raised.
4. An event is written in public block chain and alert is raised to the hospital(H), patient and mobile device for information.
5. Thus Shortblocks ensures the scalability, security and privacy of smart health care system.

Algorithm SHORTBLOCKS

procedure SHORTBLOCKS (P, PrB, PuB, H)

{

// P = Patient

// PrB= Private Blockchain

// Pub= Public Blockchain

// H=Hospital

Step:1 Patient(P) information is collected using wearable smart devices.

Step:2 Information is sent to the Smart devices, smart contracts

Step:3 The formatted information from the smart contract is sent to the private blockchain (PrB) for full examination

Step:4 If the information received, after full examination is found above the threshold, then it is declared as abnormal condition of the patient, needs treatment.

Step:5 An event is written in the public block chain(Pub) and alert is raised to the hospital(H), patient and mobile device for information.

}

end procedure

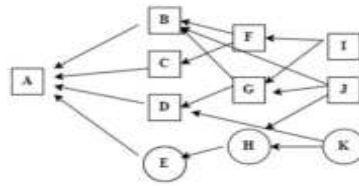


Fig. 5.1: SHORTBLOCKS cluster arrangement

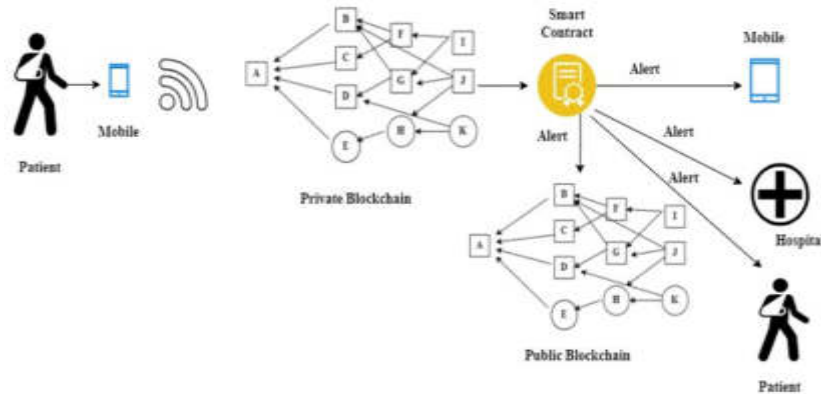


Fig. 5.2: SHORTBLOCKS process

5. Implementation. The efficiency with which each block is sent to all network miners prior to its creation determines the protocol’s security. The process of creating blocks is tedious because a proof-of-work is required for each new block. A secure blockchain will have a block propagation time that is less than or equal to the time it takes for the entire network to generate a new block. With only three to seven transactions per second, blockchain is severely limited in its throughput.

The proposed SHORTBLOCKS protocol, on the other hand, is more suited to blockchain deployment in practice. A Directed Acyclic Graph, is used to arrange blocks in the proposed new protocol instead of a lengthy chain of blocks. The blocks are organised into a k-cluster via SHORTBLOCKS protocol, with blocks outside the cluster being indicated by square and those inside the cluster being indicated by oval, as depicted in Fig 5.1.

Patients in our proposed system use a variety of wearable medical devices, including but not limited to blood pressure monitors, insulin pumps, and others. The application formats and aggregates the health information that is provided to smart devices like smartphones and tablets. After finishing, the applicable smart contract receives the formatted information and sends it to the private blockchain for full examination, together with the threshold value as shown in Fig. 5.2. If the health reading is below the threshold value, it is considered abnormal compared to standard readings. Appropriate treatment will be suggested. An event will be created on the public blockchain and alerts smart devices and hospitals if the health reading is aberrant.

When an alert is issued, the public blockchain will only store the event. The data collected from wearable health monitors will be sent to a specific Electronic Health Record, EHR system for storage. Along with the transaction event being saved on the blockchain, smart contracts or hospitals will also transmit treatment commands to EHR storage. The integration of blockchain technology with electronic health records (EHR) allows for the verification of personal health information. Protecting and identifying changes to patient records in electronic health records is made easier with this. Not all nodes can execute smart contracts. Additionally, it is the responsibility of these selected nodes to validate the new block. One way to limit the exposure of data

Table 6.1: Input Patient Data for the Experimental Analysis

Patient	A	Threshold
Heartbeat	80	>100
Blood Group	A	-
Blood Pressure	130/75	150/100
Diabetic Level	320	>200
Temperature	99	>103

Table 6.2: Avg Time Compariso

Data Type	Proposed method – ‘SHORTBLOCKS’ protocol (with Blockchain Avg Time in Sec)	SPECTRE protocol Avg Time in Sec	GHOSTDAG protocol Avg Time in Sec	Simple – Blockchain Avg Time in Sec
Block creation	0.643	0.665	0.687	0.736
Block updating	0.476	0.492	0.517	0.537
Block sharing	0.491	0.503	0.512	0.523
Block deleting	0.679	0.702	0.728	0.754

is to limit who can view it. This includes healthcare practitioners, manufacturer of the devices, and patients themselves.

6. Result Analysis. The proposed method “SHORTBLOCKS” and the existing simple blockchain method are compared in terms of performance parameters viz. Avg. time, Latency, Processing Overhead. The input data collected from the Patient A, using wearable devices for the experimental analysis is tabulated below in Table 6.1.

As the diabetic levels are above the threshold from the above input data, alert will be raised to the patient, using the public block chain for preserving the event.

a) Avg. Time. Based on the change in treatment registration requests (Transactions count) received, Fig 6.1 shows the time-to-mine comparison of two ways on existing method of using blockchain and proposed method of blockchain with SHORTBLOCKS protocol. The results are shown in seconds. On one side, we have processing time, and on the other, we have the number of treatment registration requests, transaction count. From the graph, it is clear that both methods show a trend towards increasing treatment registration requests after the first 100 or so numbers. By lowering the waiting time in the mining queue, the proposed method with SHORTBLOCKS protocol offers higher performance as compared. The Avg. time comparison between the proposed and existing methods is tabulated in Table 6.2.

b) Latency. Figure 6.2 and Table 6.3 shows the results of an investigation on the invoke transaction execution latency for the suggested system, and the existing system, broken down by user groups: 10, 20, 30, 40, and 50 users. Increasing the average latency in relation to the number of users makes change.

c) Processing Time. The results of the simulations for evaluating the systems’ performance in terms of processing overhead are shown in Figure 6.3. Many resource metrics, such as processing time, bandwidth, indirect memory, and an excess memory, may be necessary for the validation of new blocks in blockchain. The technical term for this is processing overhead. This value is discovered to be less in the suggested model when contrasted with certain current approaches. In contrast to the present models, the suggested solution outperforms them in terms of processing speed and data security. Table 6.4 depicts the statistics.

7. Conclusion. This article proposed that smart contracts built on the ‘SHORTBLOCKS’ blockchain be used to analyze patients’ health data in real-time. Using the ‘SHORTBLOCKS’ protocol, the system records the data of transactions on the blocks and utilizes smart contracts to generate notifications for the healthcare provider and patient as needed. This methodology ensures that health-related notifications are sent in a secure way in compliance with safety regulations. This method provides safe and secure patient data transmission and

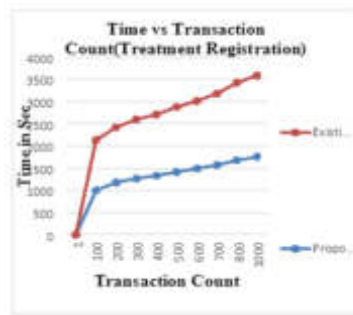


Fig. 6.1: SHORTBLOCKS avg time comparison

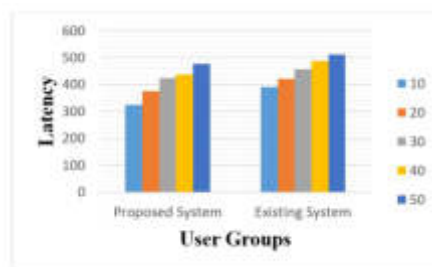


Fig. 6.2: Latency

Table 6.3: Latency Comparison

No of Pa-tients	Proposed method – ‘SHORTBLOCKS’ protocol with Blockchain	SPECTRE protocol	GHOSTDAG protocol	Simple Blockchain
10	325	345	363	390
20	375	387	403	421
30	423	438	446	456
40	436	452	467	487
50	476	489	498	512

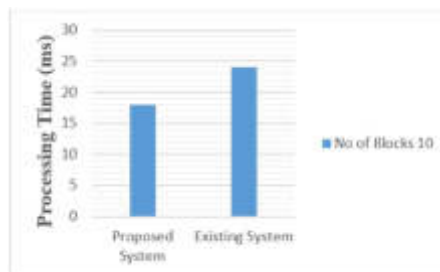


Fig. 6.3: Processing Overhead comparison

Table 6.4: Processing Overhead

Processing Time	Proposed System SHORTBLOCKS	SPECTRE protocol	GHOSTDAG protocol	Simple-Blockchain
No of Blocks (10)	18 ms	21 ms	23 ms	28 ms

communication using two blockchain viz. private and public blockchain. Due to the enormous size inflation that would result from storing complete health records on a blockchain, which would necessitate substantially more storage space at each node, we do not keep patient health information in the blockchain. Only the event raised will be facilitated in the public blockchain.

As a result, EHR (Electronic health record) gets the health data. The use of blockchain technology as a ledger and the integration of electronic health records (EHR) for the purpose of authenticating patient medical history data is only documenting the occurrence of events. Consequently, this will aid in the detection and prevention of data tampering with patient records in EHR. When energy consumption and sluggish computing are big issues, our model offers a quick, secure, and high-throughput alternative to standard blockchain-based remote patient monitoring.

Acknowledgment. The authors acknowledge the support and cooperation rendered by all the members directly and indirectly.

REFERENCES

- [1] Griebel, Lena, et al. "A scoping review of cloud computing in healthcare." *BMC medical informatics and decision making* 15.1 (2015): 1-16.
- [2] Siyal, Asad Ali, et al. "Applications of blockchain technology in medicine and healthcare: Challenges and future perspectives." *Cryptography* 3.1 (2019): 3.
- [3] Dwivedi, Ashutosh Dhar, et al. "A decentralized privacy-preserving healthcare blockchain for IoT." *Sensors* 19.2 (2019): 326.
- [4] Dubovitskaya, Alevtina, et al. "Secure and trustable electronic medical records sharing using blockchain." *AMIA annual symposium proceedings*. Vol. 2017. American Medical Informatics Association, 2017.
- [5] Bhatti, Anam, et al. "Development of cost-effective tele-monitoring system for remote area patients." 2018 International Conference on Engineering and Emerging Technologies (ICEET). IEEE, 2018.
- [6] Zhang, Peng, et al. "FHIRChain: applying blockchain to securely and scalably share clinical data." *Computational and structural biotechnology journal* 16 (2018): 267-278.
- [7] Jayasuruthi, L., A. Shalini, and V. Vinoth Kumar. "Application of rough set theory in data mining market analysis using rough sets data explorer." *Journal of Computational and Theoretical Nanoscience* 15, no. 6-7 (2018): 2126-2130..
- [8] Zhang, Jie, Nian Xue, and Xin Huang. "A secure system for pervasive social network-based healthcare." *Ieee Access* 4 (2016): 9239-9250.
- [9] Ahmed, Syed Thouheed, Vinoth Kumar, and JungYoon Kim. "AITel: eHealth Augmented Intelligence based Telemedicine Resource Recommendation Framework for IoT devices in Smart cities." *IEEE Internet of Things Journal* (2023)..
- [10] Nagaraj, J., & Leema, A. Light weight multi-branch network-based extraction and classification of myocardial infarction from 12 lead electrocardiogram images. *The Imaging Science Journal*, 71(2), 188-198 (2023).
- [11] Mettler, Matthias. "Blockchain technology in healthcare: The revolution starts here." 2016 IEEE 18th international conference on e-health networking, applications and services (Healthcom). IEEE, 2016.
- [12] Nakamoto, Satoshi. "Bitcoin: A peer-to-peer electronic cash system." *Decentralized business review* (2008).
- [13] Ivan, Drew. "Moving toward a blockchain-based method for the secure storage of patient records." *ONC/NIST Use of Blockchain for Healthcare and Research Workshop*. Gaithersburg, Maryland, United States: ONC/NIST. Vol. 1170. sn, 2016.
- [14] Chen, Yi, et al. "Blockchain-based medical records secure storage and medical service framework." *Journal of medical systems* 43 (2019): 1-9.
- [15] Natarajan, Rajesh, Gururaj Harinahallo Lokesh, Francesco Flammini, Anitha Premkumar, Vinoth Kumar Venkatesan, and Shashi Kant Gupta. "A Novel Framework on Security and Energy Enhancement Based on Internet of Medical Things for Healthcare 5.0." *Infrastructures* 8, no. 2 (2023): 22.
- [16] Prisco, Giulio. "The blockchain for healthcare: Gem launches gem health network with philips blockchain lab." *Bitcoin Magazine* 26 (2016).
- [17] Wang, Shuai, et al. "Blockchain-powered parallel healthcare systems based on the ACP approach." *IEEE Transactions on Computational Social Systems* 5.4 (2018): 942-950.
- [18] Talesh, Shauhin A. "Data breach, privacy, and cyber insurance: How insurance companies act as "compliance managers" for businesses." *Law & Social Inquiry* 43.2 (2018): 417-440.

- [19] Budida, Durga Amarnath M., and Ram S. Mangrulkar. "Design and implementation of smart HealthCare system using IoT." 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS). Ieee, 2017.
- [20] Sivagami, S., D. Revathy, and L. Nithyabharathi. "Smart health care system implemented using IoT." International Journal of Contemporary Research in Computer Science and Technology 2.3 (2016): 641-646.
- [21] Jiang, Shan, et al. "Blochie: a blockchain-based platform for healthcare information exchange." 2018 IEEE International Conference on Smart Computing (SmartComp). IEEE, 2018.
- [22] Nichol, P. B. "Blockchain applications for healthcare: Blockchain opportunities are changing healthcare globally-innovative leaders see the change." (2016).
- [23] Dhiman, Gaurav, V. Vinoth Kumar, Amandeep Kaur, and Ashutosh Sharma. "Don: deep learning and optimization-based framework for detection of novel coronavirus disease using x-ray images." Interdisciplinary Sciences: Computational Life Sciences 13 (2021): 260-272
- [24] Cyran, Marek A. "Blockchain as a foundation for sharing healthcare data." Blockchain in Healthcare Today (2018).
- [25] Ahram, Tareq, et al. "Blockchain technology innovations." 2017 IEEE technology & engineering management conference (TEMSCON). IEEE, 2017.
- [26] Shubbar, Safa. Ultrasound medical imaging systems using telemedicine and blockchain for remote monitoring of responses to neoadjuvant chemotherapy in women's breast cancer: concept and implementation. Diss. Kent State University, 2017.
- [27] Prisco, Giulio. "The blockchain for healthcare: Gem launches gem health network with philips blockchain lab." Bitcoin Magazine 26 (2016).
- [28] Haoxiang, Wang. "Trust management of communication architectures of internet of things." Journal of trends in Computer Science and Smart technology (TCSST) 1.02 (2019): 121-130.
- [29] Shakya, Prof. "Efficient security and privacy mechanism for block chain application." Journal of Information Technology and Digital World 1.2 (2019): 58-67.
- [30] Sivaganesan, Dr D. "Block chain enabled internet of things." Journal of Information Technology and Digital World 1.1 (2019): 1-8.
- [31] Maithili, K., V. Vinothkumar, and P. Latha. "Analyzing the security mechanisms to prevent unauthorized access in cloud and network security." Journal of Computational and Theoretical Nanoscience 15, no. 6-7 (2018): 2059-2063.
- [32] Li, Mengyi, Lirong Xia, and Oshani Seneviratne. "Leveraging standards based ontological concepts in distributed ledgers: a healthcare smart contract example." 2019 IEEE International Conference on Decentralized Applications and Infrastructures (DAPPCON). IEEE, 2019.
- [33] Sompolinsky, Yonatan, and Aviv Zohar. "Phantom." IACR Cryptology ePrint Archive, Report 2018/104 (2018).
- [34] Shen, Meng, et al. "Privacy-preserving image retrieval for medical IoT systems: A blockchain-based approach." IEEE Network 33.5 (2019): 27-33.

Edited by: Polinpapilinho F. Katina

Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions

Received: Jan 7, 2024

Accepted: Mar 21, 2024



ENHANCED THROTTLED LOAD BALANCING FOR VIRTUAL MACHINE ALLOCATION IN MULTIPLE DATA CENTERS

P.HANUMANTHA RAO *AND DR.P.S.RAJAKUMAR †

Abstract. "Cloud computing" hosts software and other services in remote data centers that customers can access worldwide. The user may access all the services and applications online. The IT industry has benefited greatly from the proliferation of cloud computing. On the flip side, organizations moved their operations to the cloud as a result of industrial automation. A surge in demand for cloud computing was directly correlated to the quick migration of businesses. Businesses looking to minimize expenses without sacrificing service quality will find this approach to be ideal. Considering the meteoric rise of cloud computing, service providers are delighted. Contrarily, distributing resources is a challenging task. Cloud computing overcomes some of its most fundamental obstacles, one of which is the load-balancing approach employed by load-balancers to economically optimize costs while minimizing time expenditures. Quick services for cloud customers and minimal cost for cloud providers are the goals of the optimal resource allocation method. This research suggests a novel approach to increase task processing time, which can aid in increasing cloud computing's load balancing capabilities. The proposed method Enhanced Throttled Load Balancing Algorithm (ETLBA) is an upgrade to the original Throttled Algorithm, which efficiently performs resource allocation and load balancing. The proposed ETLBA is contrasted with the existing algorithms, Round Robin, Active Monitoring Load Balancing Algorithm (AMLBA) and Throttled Load Balancing Algorithm (TLBA) to display the efficacy. Cloud Analyst tool simulates the proposed and existing methods. According on the results of the simulations, the proposed algorithm ETLBA achieves better outcomes than the popular existing algorithms in terms of processing time, request processing time, and datacenter cost. It shows 18% reduction in response time, 7% reduction in data center processing time, 16% reduction in data center request processing time and 4% less data center cost compared to the existing solutions. ETLBA performs better by selecting virtual machines using a prioritized index table and consumption index. It limits idling resources, improves response as well as reduces processing times, and cloud costs compared to conventional solutions.

Key words: Cloud Computing, Resource Allocation, Round Robin, Load Balancing, Throttled Load Balancing, Enhanced Throttled Load Balancing, Improved Response time, Data center Cost.

1. Introduction. "Cloud computing" is a system that lets people share and access large amounts of data and other resources. After using a service, users only pay for what they spent. The open environment of cloud computing allows to demonstrate how data, software packages, and distributed resources are stored.

Many multinational corporations (MNCs) offer cloud services, including Microsoft, Amazon Web Services, and many more [1]. "Software as a service," "infrastructure as a service," and "platform as a service" are the three most prevalent models of cloud computing [2]. In this context, "Infrastructure as a service" is exemplified by AWS EC2 instances. Google Apps is an example of software as a service, while Microsoft Azure is an example of a platform as a service.

Cloud computing primarily assists in the sharing of resources, data, and internet-based applications. Using web-based apps, it offers consumers on-demand services. Concerns regarding data backup and restoration are unfounded. Fig. 1.1 shows the elaborated infrastructure of the cloud computing with all the components included.

The benefits of cloud computing is its usability and cost-effectiveness of its administration. It features many desirable traits, including, but not limited to, reliability, virtualization, multitasking, improved framework cost, and referenced highlight assistance. The innovation that is stepping forward is cloud computing. A large number of startups nowadays use cloud computing. Instead of purchasing the foundation, business innovators are saving time, money, and office space by connecting the cloud benefits using PCs. In cloud computing, users

*Research Scholar, Dept. of CSE, Dr.M.G.R.Educational and Research Institute, Maduravoyal, Chennai, Tamil Nadu India. (hanumanthraovbit@gmail.com).

† Professor, Dept. of CSE, Dr.M.G.R.Educational and Research Institute, Maduravoyal, Chennai, Tamil Nadu India. (Rajakumar.subramanian@drmgrdu.ac.in).

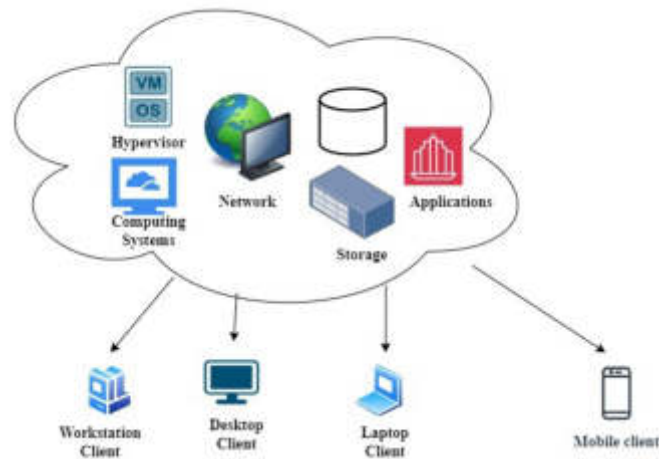


Fig. 1.1: Cloud Computing Infrastructure

can access services on an as-needed basis and pay for them in advance. This is why cloud computing services are becoming increasingly popular among customers with occasional needs.

Cloud computing aims to improve efficiency, decrease computational power requirements for thin devices, decrease operational expenses, increase security, and speed up data processing. In addition to these benefits, it makes the system more resilient to changes, faster at processing various data sets, less expensive overall (hardware, software, and maintenance costs), more energy efficient and less space hungry on discs [21].

In order to keep up with demand, organizations are distributing the burden among multiple servers. A technique known as "load balancing" is employed to ensure that no single server is overwhelmed. Delays, request drops, or even crashes may occur if the system is overloaded. In order to distribute the processing load among numerous servers, Load Balancing makes use of network connections. It reduces total response time and maximizes throughput [3]. To avoid any server going down due to overload, load balancing is essential.

When it comes to cloud environments, load balancers are efficient. Servers can be rendered inoperable due to crashes caused by extremely heavy workloads. All procedures rely on consistently fast response times and high availability of services. In addition to detecting downed servers, load balancers can reroute requests to up and operating servers. If one server is too busy, a load balancer can send requests to another. Load balancers primarily aim to ensure that servers are in better health. Servers and users are regulated by a load balancer.

It processes data packets sent by networks and applications. In order to perform multi-server request distribution, a load balancer employs a number of techniques. There are two tiers to load balancing: Level one, which entails establishing a link between the applications or services and the virtual machines that are being requested. Layer 2 entails establishing a link between physical hosts and virtual machines [4],

Load balancing is a vital component of cloud computing, for efficient resource allocation across multiple data centers. Many existing and popular load balancing techniques are used for resource allocation. But response time is the crucial factor that evaluates the efficiency of these techniques. The primary objective of this article is to present a novel approach for improvising the response time and optimizing the data center costs when compared with existing algorithms. The proposed method Enhanced Throttled Load-Balancing Algorithm (ETLBA) aims at efficient load balancing by improvising the existing throttled load balancing technique. This study presents a contrary of outcomes obtained through the use of "Cloud-Analyst Simulator". The results demonstrate that the suggested algorithm decreases the total processing time spent for the requests and response time of the datacenter.

This is the remaining structure of the paper: Section 2 presents work in this area along with the description of the several load-balancing algorithms. The proposed algorithm, along with its flowchart and pseudo-code, is presented in Section 3. The paper's section 4 details the experimental setting. While the results of the

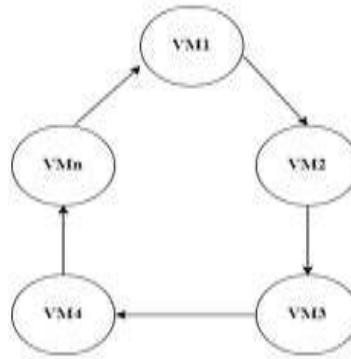


Fig. 2.1: Round Robin Algorithm Process

comparison and analysis are presented in Section 5. The paper is concluded in Section 6.

2. Literature Review.

2.1. Load-Balancing Algorithms. There are a plethora of loadbalancing techniques available in the cloud. Active Monitoring LoadBalancing (AMLB), Round Robin (RR), and Throttled (TLB) are the three most used load-balancing algorithms [5]. The algorithms that were employed are evaluated using "Cloud Analyst Simulator".

2.1.1. Round-Robin Algorithm. When it comes to time-sharing systems, the useful and most popular load-balancing method is Round-Robin. With Round-Robin, virtual machines (VMs) are distributed fairly in a circular order. Assuming the processing powers of the virtual machines are equal, this method provides benefits such as being easy to understand and implement, guaranteeing fairness in operations, and handling tasks promptly.

Distributing client requests among servers is made easy with this method. Requests from clients are received sequentially by each server. Since it is simple to both understand and apply, it ranks high among the most popular algorithms [6].

This method iteratively routes client requests to the servers that are available. When server processing and storage capacities are close to one another, it works well. This process sends requests to the node with the fewest connections, which means that at any one moment, a few of nodes can be experiencing excessive load, while others are completely unoccupied. Fig. 2.1 depicts the process of Round Robin Algorithm.

Within this algorithm, there is no famine. Round Robin's drawbacks include a lack of scalability and flexibility, the fact that some nodes may be under tremendous strain while others sit idle, and the fact that the former allocation status of the virtual machine is not preserved.

2.1.2. Weighted Round-Robin Algorithm. The Weighted Round-Robin algorithm takes its cue from the Round-Robin method and employs a weight table to allocate tasks to virtual machines according to their capacities. It executes circular distribution using this table. The algorithm's benefits: enhances the Round-Robin algorithm by incorporating a weight-table of virtual machines' processing capacities into its rotational operation; this makes the method more efficient than the original in situations when the processing powers of the virtual machines vary. The lack of scalability and flexibility, as well as the inability to restore the prior allocation status of virtual machines, are further drawbacks. Fig. 2.2 depicts the process of the algorithm.

2.1.3. Active Monitoring Load Balancing Algorithm. Because of the unpredictable nature of cloud computing, certain servers may experience excessive load during equally distributed current execution, while others may be inactive or barely touched. Similarly, by shifting resources from overburdened to underutilized servers, load distributing boosts performance. One important characteristic of cloud computing is efficiently distributing resources in the cloud and scheduling, which is used to evaluate the system's performance [7].

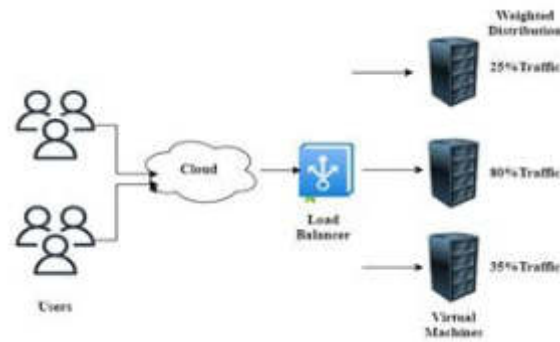


Fig. 2.2: Weighted Round Robin Algorithm Process

Optimization of costs, as a result of faster response and processing times, is affected by the attributes taken into account.

Clients submit their jobs to the computer system here. Each job that is submitted to the cloud is added to a stack and queued as it arrives. After making an estimate of the job's size, the cloud manager verifies the virtual machine's availability and capacity. The job scheduler assigns the specified resource to the queued work without delay as soon as the job size and the size of the accessible asset (virtual machine) are in agreement [8].

The major drawback of this algorithm is, in order to assign resources to a free virtual machine, it sequentially evaluates each virtual machine which is time consuming process.

2.1.4. Throttled Load Balancing Algorithm. A static load-balancing method best describes this technique. Here, we begin by verifying the values of each virtual machine's index. For system resource allocation, the a request has been forwarded to the point where the load balancer parses a table. To update the allocation policy, a particular load balancer is notified of the request, which then either returns it to the requester or processes it in reverse [9]. The complete procedure of deallocating the system begins once the allocation of the system is successful.

The increased performance and utilization are the results of this mechanism's provision of more sharing and allocation of system resources. The throttling threshold is typically set at 1. The threshold might be set to a user-specified value with little configuration.

At all times, this algorithm's fixed quantity of cloudlets are allotted to a single virtual machine. The quantity of virtual computers (VMs) available at the datacenter determines how incoming requests are dealt with if there are more request groups. Aside from that, it waits for the next available virtual machine to become available.

In comparison to the Active Monitoring Load Balancing (Optimal) technique, this algorithm represents an incremental improvement. Initially, this algorithm begins its search from the most recently allocated virtual machine all the way up to the n th virtual machine. Nevertheless, there is still a problem, and that is the fact that it does not make use of those virtual machines that become available subsequently to the execution of the request. The below Fig. 2.3 represents the procedure of Throttled Load-balancing algorithm.

2.2. Related Work. Using a throttled load balancing method in a multiple data centers, the study [10] optimized response time by distributing workloads across virtual machines. After much deliberation, they settled on the throttled load balancing algorithm as the best option for the data center in terms of processing time for requests and total response time, with minimal processing expenses.

The authors of [11] presented their idea for the Advanced Throttled Load Balancing Algorithm in the year. A priority is given to every virtual machine. Depending of the capacity of the virtual machine (VM) as well as the number and size of tasks that have been given to the operation, the priority is determined. Among the several virtual machine (VM) sets that are accessible, the enhanced tuning scheduler chooses the VM that has the highest priority. In addition, a level of priority is established in order to prevent overloading. On the other

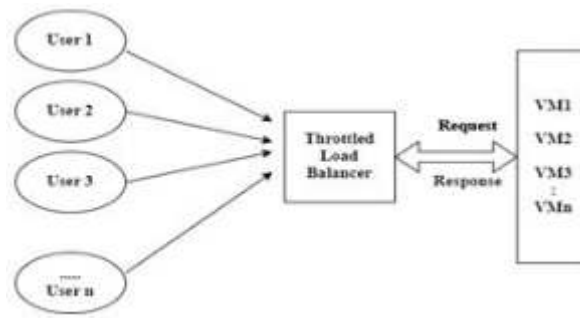


Fig. 2.3: Throttled Load-Balancing Algorithm Process

hand, if the virtual machine's priority is lower than the priority level, then the job will not be assigned to that virtual machine.

The paper [12] suggested an effective procedure for load balancing that is dependent on two distinct characteristics. The time it takes to respond to queries is the first distinguishing feature, and the second characteristic is the load distribution among the virtual machines that are already in existence. They contrasted the throttled variant with the Round Robin approach and suggested a tweaked version of the throttled algorithm.

Furthermore, it was discovered that the utilization of virtual machines (VMs) is more effective when using the Round Robin algorithm and the changed Throttled algorithm in comparison to the Throttled algorithm. In addition, out of the three algorithms tested, the Modified Throttled method yielded the best average reaction time.

The review study conducted in [13] demonstrates that load balancing is an essential technique of the cloud computation environment. It contributes to the improvement of load distribution and the efficient allocation of resources, particularly with regard to the enhancement of response time for cloud users. According to the article, there are a great deal of problems associated with LB. Some of these problems include the scheduling of activities, migration, and the utilization of resources, among other things. Research and studies on load balancing that have been conducted over the previous six years are surveyed and analyzed by the writers. This analysis's findings also show that intelligent methods, including AI and ML, have potential for cloud-based learning analytics. In particular, this study benefits researchers in identifying research topics connected to load balancing, particularly with regard to reducing the response time and avoiding breakdowns in the servers. Another benefit of this research that supports our idea is the availability of tools required for simulation and experimental contexts. This research focuses on the modelling resources.

The exceptional quality of Cloud Analyst and Cloud Sim as top-tier resources for this area of research is another thing that they demonstrate. This is the advantage of employing these tools.

A modified throttled balancing technique was proposed in [14], which they then implemented with the help of the CloudAnalyst tool of CloudSim. Following this, they checked it against other methods of load balancing and confirmed that a tweaked throttled algorithm outperformed the more traditional approaches.

The paper [15] suggested using a neural inference fuzzy system for load effective balancing. In this study, the authors also discuss the safety of virtual machines (VMs) within cloud-hosted environments. There is a correlation between load balancing and the NP-hard optimization issue. According to Forbes, they are following the news regarding the implementation of the Protection of Personal Information Act. Security of the cloud continues to be a problem, and the current system makes use of a hybrid-based fuzzy LB, but this does not satisfy those requirements. In order to improve CPU utilization and turnaround time, the authors proposed their work, which they referred to as MANFIS (Modified Adaptive Neuro Fuzzy Inference System).

In addition, they focused on increasing the level of security that their work offered. Utilizing the Fire-fly Algorithm allows for the optimization of the parameters of the MANFIS system. Utilizing the Enhanced Elliptic Curve Cryptography allows for the implementation of security measures for user authentication. Users can be authenticated using this method, which does not require a password. Based on the findings, the authors inform

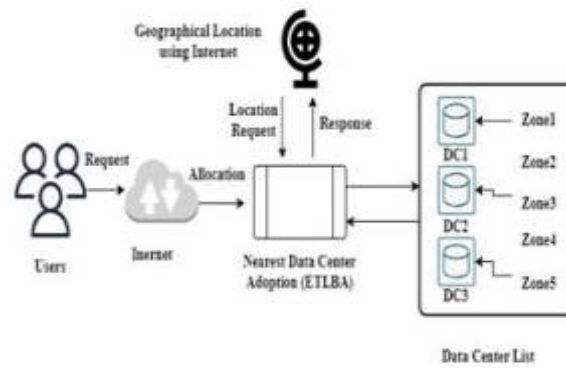


Fig. 3.1: Nearest Data Center Adoption (ETLBA)

us that cloud security has been enhanced and meets our expectations. Further, the experimental results reveal that they outperform the current system in terms of resource utilisation, cost, and execution time.

In their study, the authors of [4] introduced three algorithms for load balancing. These algorithms find use in cloud computation. Here are three algorithms: RR, AMLB, and TLB, which stand for “round-robin, active load-balancing monitoring, and throttled load-balancing”, respectively. Following an analysis of the three load balancing algorithms’ performance, it was determined that TLB outperformed the others in terms of data centre time and reaction time.

3. Proposed Method - Enhanced Thottled Load Balancing Algorithm (ETLBA). We propose an Enhanced Thottled Load Balancing Algorithm (ETLBA) as a solution to the problem that arises while utilizing the throttled load balancing algorithm. The proposed method consists of three major components:

1. Nearest data center adoption
2. Load Balancer with availability index
3. Usage index of Virtual Machines.

3.1. Nearest Data Center Adoption. The rapid development of cloud platforms has resulted in service providers maintaining many data centres spread out over the globe. The data centers that make up the cloud computing environment are dispersed throughout multiple regions.

The proposed work implements the nearest data center (DC) adoption such that to improvise the response time. The DC-1 server is hosted in zone 1, the DC-2 server is hosted in zone 3, and the DC-3 server is hosted in zone 5. The allocation of virtual machines in accordance with the proposed policy for the data center is done based on which are geographically nearest [16].

The procedures that make up the suggested algorithm for selecting the data center that is geographically nearest to the user, is listed below

1. The request is presented by the first user.
2. The database of Internet characteristics will be responsible for maintaining the table of region proximity
3. The request was forwarded to the data center that was found to be the closest.
4. In the event when multiple DC is experiencing the same network delay.
In order to maintain a balanced load,
 - a) Assign the DC in a random fashion.
 - b) Give the DC with the lowest possible network latency
5. Make sure the data center policy is keeping up to date.

The process is depicted in the Fig. 3.1.

3.2. Load Balancer with availability index. The databases of virtual machines provide the basis of a throttled algorithm. Load balancers keep track of virtual machine identifiers and the status of their services, such as whether they are available or busy [17].

The following steps depict the role and process of load balancer with availability index in the proposed ETLBA.

1. Load Balancer creates and updates an table of index for each virtual machine. Additionally, it monitors the availability and busyness of each virtual machine. When the programme began running, every single virtual computer was online.
2. User initiates the request and that will be allocated to nearest Data Centre as mentioned in part 1. Now Data center Controller is then assigned a new user request.
3. The Data Centre Controller will call the Load Balancer in order to allocate the virtual machine to the next available slot.
4. After that, when all available virtual machines are ready, the Efficient throttled load balancer will begin to build a new map.
5. Then, if the length of the available VM map is larger than zero, the efficient throttled load balancer will deconstruct the map and retrieve the first VM ID from it. After that, the process will be taken up by part 3 of the proposed method ETLBA

3.3. Usage Index of Virtual Machine. In this part, the proposed ETLBA maintains the usage list of all available VMs. The most recent task's total processing time is used to compute usage. The overall cost of a datacenter is determined by adding up the costs of data transfer and virtual machine rental.

1. The efficient load balancer monitors the usage of the VM who's ID was retrieved. If the selected one is with the lowest usage, it retrieves its ID from the "Available Index" table and returns it to the datacenter controller (DCC).
2. Otherwise the available virtual machine with lowest usage is searched again in the index table and retrieved, further for forwarding to the data center controller.
3. The Data Centre Controller receives the VM id from the Enhanced Throttled Load Balancer.
4. Based on that identifier, the Data Centre Controller will pass the call on to the appropriate virtual machine.
5. After the new VM_ ID has been allocated, the data center controller informs the enhanced throttle load balancer and removes the corresponding entry from the available virtual machine map. And keeps the status as 'Busy'.
6. After the Enhanced Throttled Load Balancer receives the request from the Controller of Data Centre, it upgrades the VM map of the available virtual machines accordingly.
7. The following procedures are executed in the event that the virtual machine (VM) in Fig. 3.2 requested does not appear in the VM Map:
 - a) The Enhanced Throttled Load Balancer returns 1.
 - b) The database administrator will then add the request to a queue.
 - c) The Enhanced Throttled Load Balancer receives a alert from the Data Centre Controller when the processing requests of each virtual machine are finished and the response has been received. It then de-allots the respective virtual machine and adds its ID to the available VM Map.
 - d) In the instant after a VM is de-allocated, the Data Centre Controller examines the request queue. If there are any calls in the pending queue, processing will start at the third phase and continue thereafter.

3.4. Procedure - ETLBA.

Procedure Enhanced Throttled Load Balancing (ETLBA)_Part 1.

- Step 1: Users sends requests from different regions for resource allocation.
 Step 2: Database of Internet characteristics will maintain the table of region proximity
 Step 3: While(New request are received by the Nearest Data Center Adoption Method)
 Step 4: Check (Nearest Data Centre =available)
 Step 5: Allocate the user request to the nearest Data Centre with low latency
 Step 6: In case of multiple nearest data centre with same latency
 Step 7: Allot the nearest data center on random basis
 end procedure



Fig. 3.2: Procedure of ETLBA

Procedure Algorithm Enhanced Throttled Load Balancing (ETLBA)_ Part 2.

- Step 1: Make sure that all virtual machines are first assigned the state of "AVAILABLE" in the VM State list; otherwise, set it to "BUSY."
- Step 2: While(New requests are received by the data center Controller)
- Step 3: If (available VMArray () >0)
- Step 4: If 'Yes' :
- Step 5: Check (Usage_VM.Selected =least)
- Step 6: Return the VMID to the data centre
- Step 7: else : Search for the next least usage available VM
- Step 8: If (available VMArray () =0) Data Centre Controller queues the request
- Step 9: Repeat Step 3
- end procedure

Procedure Enhanced Throttled Load Balancing (ETLBA)_ Part 3.

procedure Enhanced Load Balancer_Usage Index

Begin

- Step 1: VM_Id = 1; Min = 0; i=0;
- Step 2: For each VM_selected in VMList
- Step 3: usage = vm.getUsage();
- Step 4: vmId = vm.getId();
- Step 5: If i== 0 then
- Step 6: min = usage;
- Step 7: Else

```

Step 8: If min > usage then
Step 9: min = usage;
Step 10: End If
Step 11: End If
Step 12: i++;
Step 13: End For
Step 14: Return VM_Id;
Step 15: End

```

The proposed method “ETLBA” allocates resources efficiently as mentioned in the above procedure. In detail the above procedure works as: User sends request from different regions for resource allocation. All these requests are maintained region wise by the internet database. If the nearest data center is free, then it is allocated to the user request based on priority. All the user requests are queued by the data center controller. If more than one nearest data center is free, then the nearest one is allotted on random basis. All the data centers are tagged with index of availability and busy status. The data center with index available are chosen for allocation in response to user request. If two or more datacenters nearer to the user region are available with availability index then their usage is checked to choose for allocation. Once data center is allocated to the user, then its status is marked as busy and if freed, as available by data center controller.

4. Experimental Setup. Data center load balancing techniques can be cost-effectively tested using simulation. A simulated data center environment can be modelled using a variety of tool sets. Cloud Analyst simulation makes it easier and faster to examine the data. Cloud Analyst Simulator makes use of a number of technical phrases, so let’s review them.

4.1. Simulator - Cloud Analyst . The cloud computing environment ‘simulator’ that is most often used is Cloud Analyst [18]. It is an event driven simulator [19].

The major considerations are:

1. There are six distinct regions on the Earth map in the Cloud Simulator. Data centers and user bases can be located in any of the six zones.
2. The internet as it exists in the real world serves as the model in Cloud Analyst Simulation. The transmission delay and latency are caused by the traffic on the internet.
3. In the Cloud Simulator, a collection of users is referred to as a User Base. Hundreds of users could make up a single user base at times. One tool for producing load is the user base.
4. An Digital cloud-let is a user’s collection of requests. Execution commands, input files, and output files are all carried over the Internet cloud
5. Cloud Analyst Simulator revolves around its controller. It will oversee operations in the data-center setting, like the creation and deletion of virtual machines.
6. The Controller determines which virtual machine (VM) should execute the cloud task by using a virtual machine load balancer (VmLoadBalancer).

4.2. Simulation Settings. Zone1, Zone2, Zone3, Zone4, Zone5, and Zone 6 were labelled on the global map [20]. The social media platform X, which boasts over 300 million active users globally, is worth considering. The same is depicted in Fig. 4.1 [20] and Table 4.1.

For cloud usage and load balancing simulation, 5% of ‘X’ users are used. Additional assumptions include 10% of users being online at the busiest times and 10% being inactive during peak hours. We will define six in total, user databases for the six available zones, using the specifications in Table 4.2.

Fig. 4.2 displays the sample settings of the primary configuration section, which includes the following: user bases, application deployment configuration, and simulation duration.

The user requests are generated by each userbase and delivered to the processing facility for data storage. The various configurations of data centers are detailed in Table 4.3, which includes information such as the operating system, hardware type, and the cost of individual operations.

5. Results and Performance Analysis. After running simulations with the aforementioned parameters, we compared the results based on Data processing time, Overall response time, Data Center request servicing time, and Data Centre cost for the following load balancing algorithms: Round Robin, AMLBA - Active



Fig. 4.1: DC1, DC2, DC3 cloud analyst simulation

Table 4.1: X Users

Zonal Name	Zonal-Id	Registered -U sers	Sample-data (5%)
North America	1	95 Million	35,00,000
South America	2	25 Million	15,00,000
Europe	3	75 Million	25,00,000
Asia	4	35 Million	18,50,000
Africa	5	6 Million	3,50,000
Australia	6	10 Million	6,00,000

Monitoring Load Balancing Algorithm, TLBA-Throttled Load Balancing Algorithm, and the proposed method, ETLBA-Enhanced Throttled Load Balancing Algorithm.

In terms of the performance parameters mentioned earlier, the results based on simulation show that the suggested method Enhanced Throttled Load Balancing Algorithm (ETLBA) performed better than the existing methods Round Robin, Active Monitoring Load Balancing Algorithm (AMLBA), and Throttled Load Balancing Algorithm (TLBA). Below, we'll go over the same.

5.1. Overall Response Time. The amount of time it takes for the desired process to finish processing is known as response time. Round robin has an average reaction time of 179.34 milliseconds, AMLBA is 177.43 milliseconds, throttled is 171.61 milliseconds and the proposed ETLBA is 164.53 milliseconds. Based on our findings, the load balancing algorithm is not well-suited to Round Robin, AMLBA, TLBA, due to the increased response time it requires. Consequently, the proposed Enhanced throttled load balancing algorithm ETLBA outperform the other three existing in terms of response time. Fig. 5.1 graph express the performance of proposed method ETLBA in terms of Avg. Overall Response Time.

5.2. Data Center Processing Time. A load balancer's processing time in a data center is the amount of time it takes to handle all of the necessary requests. The proposed method ETLBA consumes 16.32 milliseconds, TBLA takes 23.89 milliseconds, AMLBA takes 41.12 milliseconds, and Round-robin takes 43.35 milliseconds. When compared to the RR, AMLBA and TBLA algorithms with respect to processing time in the data center, ETLBA is appropriate. The same is shown in the below graph in Fig. 5.2.

5.3. Data Center Request Servicing Time. Service time for request is also determined by the datacenter's serving time, which does not include the time it takes to transport data to the user. The proposed method ETLBA consumes 216.28 milliseconds, TBLA takes 37.17 milliseconds, AMLBA takes 45.36 milliseconds, and Round-robin takes 52.25 milliseconds. When compared to the RR, AMLBA and TBLA algorithms with respect

Table 4.2: User Database Configuration

User Database	Zonal Id	Requests per user per Hx	Data Size per request (byte)	Peak Hour (GMT)	Avg. Users in peakbss	Avg. Users nonpeals hrs
UB 1	1	60	100	13: 00 15: 00	5,00,000	50,000
UB 2	2	60	100	15: 00 17: 00	2,00,000	20,000
UB 3	3	60	100	20: 00 22: 00	3,00,000	30,000
UB 4	4	60	100	01: 00 03: 00	2,35,000	23,500
UB 5	5	60	100	21: 00 23: 00	35,000	3,500
UB 6	6	60	100	09: 00 11: 00	80,000	8,000

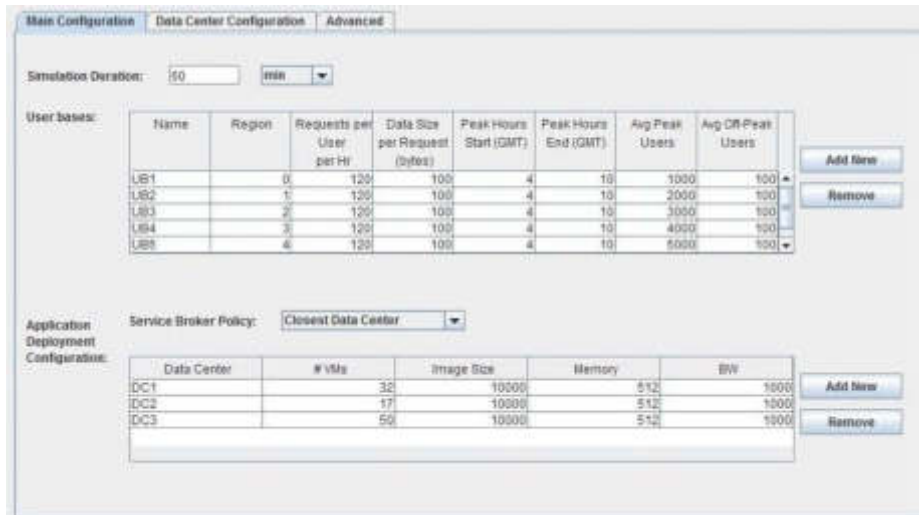


Fig. 4.2: Sample Settings in Cloud Simulator

Table 4.3: User Database Configuration

Name	Data Center1	Data Center2	Data Center3
Region-ID	1	3	S
Configuration	X86	X86	X86
Operating System	Linux	Linux	Linux
Virtual Machine	Xen	Xen	Xen
Cost of VM	0.2 \$ / hr	0.2 \$ / hr	0.2 \$ / hr
Cost in terms of Mernory	0.10 \$ / sec	0.10 \$ / sec	0.10 \$ / sec
Cost of Data Transfer	0.2 \$ / GB	0.2 \$ / GB	0.2 \$ / GB

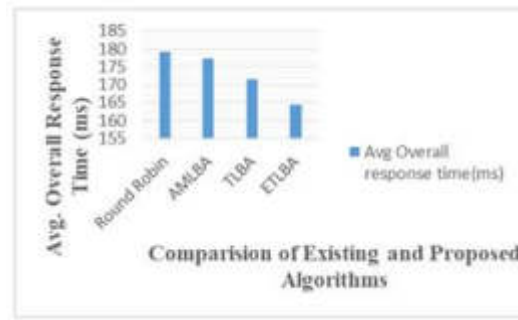


Fig. 5.1: Avg. Overall Response Time Comparison

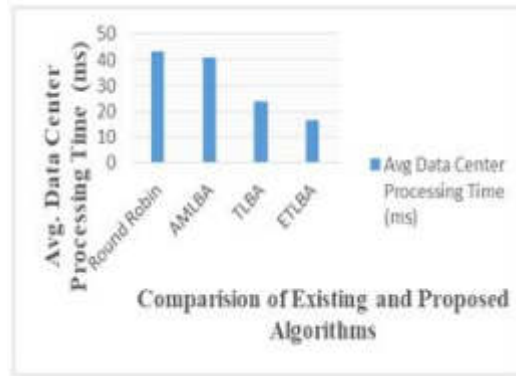


Fig. 5.2: Avg. Data Center Processing Time Comparison

Table 5.1: Overall Response Time

Algorithm	Overall resp onsetime(ms)		
	Avg(ms)	Min(ms)	Max (ms)
Round Robin	179.34	46.43	671.56
AMLBA	177.43	45.24	669.46
TLBA	171.67	43.63	663.45
ETLBA	164.53	41.28	651.43

to requests processing time in the data center, ETLBA outperforms the existing algorithms and the same is depicted in the below graph in Fig. 5.3.

5.4. Data Center Request Servicing Time. The datacenter cost is determined by adding up all the works performed by the virtual machines. These works can include diverse tasks including ALU operations, CU operations, storage operations, and more. Round robin method has incurred cost of 29.03 \$, while 26.82 \$, 23.93 \$ for AMLBA and TLBA respectively, whereas the proposed ETLBA is 21.78 \$. Based on simulation findings, the Round Robin, AMLBA, TLBA, are not cost effective, compared with the proposed Enhanced throttled load balancing algorithm ETLBA. The graph in Fig. 5.4 shows the data center costs of all the existing and proposed algorithms.

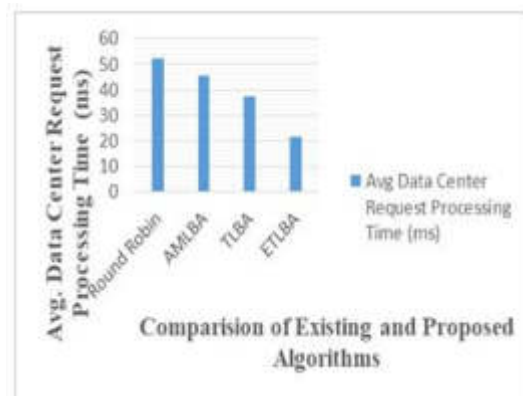


Fig. 5.3: Avg. Data Center Request Processing Time comparison

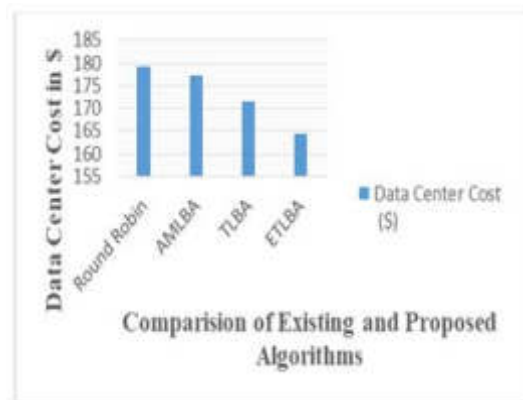


Fig. 5.4: Data Center Cost comparison

Table 5.2: Data Center Processing Time

Algorithm	Data Center Processing Time		
	Avg(ms)	Min(ms)	Max (ms)
Round Robin	43.35	8.89	251.33
AMLBA	41.12	8.89	252.15
TLBA	23.89	7.92	242.28
ETLBA	16.32	7.14	233.64

5.5. Analysis and Discussion. Table 5.1, 5.2, 5.3, 5.4 shows simulation results of all the four algorithms, Round robin, AMLBA, TLBA and the proposed ETLBA in terms of Overall response time, Data Center Request servicing time, Data Center Cost, and Data processing time. Out of the four methods under comparison, the ETLBA algorithm consistently produces the lowest level of outcomes. Since the four performance scenarios chosen are significantly distinct from one another, we can observe that ETLBA remains stable despite the variable inputs and performs well in all cases.

Table 5.3: Data Center Request Processing Time

Algorithm	Data Center Request Processing Time		
	Avg(ms)	Min(ms)	Max (ms)
Round Robin	52.25	10.19	263.27
AMLBA	45.36	10.09	243.27
TLBA	37.17	9.14	232.58
ETLBA	21.28	8.26	223.24

Table 5.4: Data Center Cost

Algorithm	Data Center Cost		
	VM Cost (\$)	Data Transfer Cost (\$)	Total (\$)
Round Robin	179.34	46.43	671.56
AMLBA	177.43	45.24	669.46
TLBA	171.67	43.63	663.45
ETLBA	164.53	41.28	651.43

6. Conclusion. A novel method ‘Enhanced Throttled Load Balancing Algorithm’ for effective load balancing is provided in this article. This method implements the strategy of allocating the resources to the priority user requests based on nearest data center, index and usage of the datacenter and its virtual machines. It delves into the strategy of enhancing load balancing to boost cloud computing performance in terms of response time, data center costs. Various load balancing approaches are simulated along with the proposed method for showcasing the performance, using the Cloud Analyst tool. These techniques include Round Robin, Active Monitoring AMLBA, TLBA (Throttle Load Balancing Algorithm), and the proposed ETLBA (Enhanced Throttled Load Balancing algorithm). Better time response, less resource starvation, more powerful virtual machines to handle more requests, and cost savings are all achieved by using the suggested algorithm (ETLBA). Consistently decreasing datacenter costs with this approach demonstrate the practical applicability and future development prospects of the ETLBA algorithm. ETLBA has the possibility of using and implementing the results of future research and testing. ETLBA can further be enhanced by including the region proximity based on the data centers. Real-world testing and research can be conducted with the proposed ETLBA in a datacentre.

REFERENCES

- [1] SINGH, ARCHANA, AND RAKESH KUMAR. , *Performance evaluation of load balancing algorithms using cloud analyst*. 10th International Conference on Cloud Computing, Data Science and Engineering (Confluence). IEEE, 2020.
- [2] NARALE, SNEHAL A., AND P. K. BUTEY. , *Throttled load balancing scheduling policy assist to reduce grand total cost and data center processing time in cloud environment using cloud analyst*." 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). IEEE, 2018.
- [3] ROY, S., HOSSAIN, D. M. A., SEN, S. K., HOSSAIN, N., & AL ASIF, M. R. , *Measuring the performance on load balancing algorithms*." Global Journal of Computer Science and Technology, 19, 41-49, 2019.
- [4] VOLKOVA, V. N., CHEMENKAYA, L. V., DESYATIRIKOVA, E. N., HAJALI, M., KHODAR, A., & OSAMA, A. , *Load balancing in cloud computing*." 2018 IEEE conference of russian young researchers in electrical and electronic engineering (EIconRus). IEEE, 2018.
- [5] MAITHILI, K., V. VINOTHKUMAR, AND P. LATHA. , *Analyzing the security mechanisms to prevent unauthorized access in cloud and network security*" Journal of Computational and Theoretical Nanoscience, 2018.
- [6] SINGH, ARCHANA, AND RAKESH KUMAR. , *Performance evaluation of load balancing algorithms using cloud analyst*." 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2020.
- [7] HOTA, ARUNIMA, SUBASISH MOHAPATRA, AND SUBHADARSHINI MOHANTY., *Survey of different load balancing approach-based algorithms in cloud computing: a comprehensive review*." Computational Intelligence in Data Mining: Proceedings of the International Conference on CIDM 2017. Springer Singapore, 2019.
- [8] MOHAPATRA, SUBASISH, SUBHADARSHINI MOHANTY, AND K. SMRUTI REKHA., *Analysis of different variants in round robin*

- algorithms for load balancing in cloud computing.*" International Journal of Computer Applications 69.22, 17-21, 2013.
- [9] GAO, G., HU, H., WEN, Y., & WESTPHAL, C. , "*Resource provisioning and profit maximization for transcoding in clouds: A two-timescale approach.*" IEEE Transactions on Multimedia 19.4, 836-848, 2016.
- [10] KARTHICK RAGHUNATH, K. M., MANJULA SANJAY KOTI, R. SIVAKAMI, V. VINOTH KUMAR, GRANDE NAGAJYOTHI, AND V. MUTHUKUMARAN, "*Utilization of IoT-assisted computational strategies in wireless sensor networks for smart infrastructure management*" International Journal of System Assurance Engineering and Management, 1-7, 2022.
- [11] PHI, NGUYEN XUAN, AND TRAN CONG HUNG. , "*Load balancing algorithm to improve response time on cloud computing.*" International Journal on Cloud Computing: Services and Architecture 7.6, 1-12, 2017.
- [12] NARALE, SNEHAL A., AND P. K. BUTEY. , "*Throttled load balancing scheduling policy assist to reduce grand total cost and data center processing time in cloud environment using cloud analyst.*" 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). IEEE, 2018.
- [13] SHAFIQ, DALIA ABDULKAREEM, N. Z. JHANJHI, AND AZWEEN ABDULLAH. , "*Load balancing techniques in cloud computing environment: A review.*" Journal of King Saud University-Computer and Information Sciences 34.7, 3910-3933, 2022.
- [14] BHANDARI, ANMOL, AND KIRANBIR KAUR., "*An enhanced post-migration algorithm for dynamic load balancing in cloud computing environment.* Proceedings of International Ethical Hacking Conference 2018: eHaCON 2018, Kolkata, India. Singapore: Springer Singapore, 2018.
- [15] LE, HIEU N., AND HUNG C. TRAN., *Ita: The Improved Throttled Algorithm of Load Balancing On Cloud Computing.* International Journal of Computer Networks & Communications (IJCNC), 14, 2022.
- [16] PATEL, HETAL, AND RITESH PATEL , "*Design and Evaluation of Wi-Fi Offloading Mechanism in Heterogeneous Networks,* International Journal of e-Collaboration, 2021, 60-70.
- [17] VINOTH KUMAR, V., S. RAMAMOORTHY, V. DHILIP KUMAR, M. PRABU, AND J. M. BALAJEE , "*Cloud analyst: an insight of service broker policy.*", International Journal of Advanced Research in Computer and Communication Engineering, 2015, 122-127.
- [18] CALHEIROS, RODRIGO N., RAJIV RANJAN, ANTON BELOGLAZOV, CÉSAR AF DE ROSE, AND RAJKUMAR BUYYA, "*CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms.*" Software: Practice and Experience, 23-50, 2011.
- [19] BUYYA, RAJKUMAR, AND MANZUR MURSHED. , "*Gridsim: A toolkit for the modeling and simulation of distributed resource management and scheduling for grid computing.*" Concurrency and computation: practice and experience 14.13-15, 1175-1220 2020.
- [20] MUTHUKUMARAN, V., V. VINOTH KUMAR, ROSE BINDU JOSEPH, MERAM MUNIRATHNAM, I. S. BESCHI, AND V. R. NIVEDITHA , "*Efficient Authenticated Key Agreement Protocol for Cloud-Based Internet of Things.*" In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2022, 365-373. Singapore: Springer Nature Singapore, 2022.

Edited by: Polinpapilinho F. Katina

Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions

Received: Jan 7, 2023

Accepted: Mar 23, 2024



DESIGN AND DEVELOPMENT OF AN UNMANNED/AUTONOMOUS OCEAN SURFACE VEHICLE USING SELF-SUSTAINING DUAL RENEWABLE ENERGY HARVESTING SYSTEM

KAMALAHASAN M ^{*}, MANIVANNAN S [†] AND SWAPNA B [‡]

Abstract. This study pioneers a breakthrough in sustainable energy solutions by developing a cutting-edge system for powering autonomous Smart Ocean surface vehicles. The research delves into the exploration of renewable energy harvesting techniques, specifically focusing on solar and hydro flow energy systems, with the aim of creating a self-sustaining power infrastructure. Through rigorous experimentation and modeling, we design and implement a versatile test rig setup to analyze the efficacy of these techniques under varying surface water conditions. Additionally, we investigate and assess distributed solar power systems ranging from 100W to 700W, as well as hydro flow power systems within the same power range, to ascertain their viability for real-world applications. Furthermore, we engineer and optimize the necessary electronic hardware utilizing IoT and industry-grade components, enabling efficient harnessing of dual renewable energy sources to power propulsion systems for our autonomous vehicles. This research introduces novel approaches to energy sustainability for autonomous ocean surface vehicles. The study disregards traditional methods and instead aims to unveil unconventional solutions for powering these vehicles. Through a series of experimental investigations, we seek to redefine the boundaries of renewable energy utilization in marine environments. By leveraging cutting-edge technologies and industry-grade components, our work aims to establish a new paradigm in the propulsion systems of autonomous oceanic vehicles.

Key words: Autonomous, vehicle, unmanned, underwater, renewable energy, IoT.

1. Introduction. The exploration and utilization of the world’s oceans have long been of paramount importance for scientific research, commercial ventures, and national security interests. In recent years, technological advancements have revolutionized our ability to navigate and understand the vast expanses of the underwater realm [1]. Among the most groundbreaking innovations are Unmanned Underwater Vehicles (UUVs) and Autonomous Underwater Vehicles (AUVs), which have emerged as indispensable tools for a wide range of maritime applications [2].

UUVs and AUVs are equipped with sophisticated sensor payloads, enabling them to conduct observation, surveillance, monitoring, and inspection tasks with unparalleled precision and efficiency. From assessing marine ecosystems and monitoring underwater infrastructure to conducting reconnaissance missions and detecting underwater threats, these vehicles play a pivotal role in expanding our understanding of the oceans and safeguarding maritime interests.

However, despite their undeniable utility, UUVs and AUVs face significant operational challenges that limit their effectiveness and autonomy [3]. Chief among these challenges are the constraints imposed by their reliance on battery power. The limited energy storage capacity of onboard batteries restricts the range and duration of autonomous missions, necessitating frequent recovery for recharging and data offloading. This reliance on support vessels not only introduces logistical complexities but also incurs substantial operational costs, particularly in remote or inaccessible marine environments [4].

Moreover, the data storage capabilities of UUVs and AUVs are often insufficient to accommodate the vast amounts of information collected during extended missions. This limitation hampers the vehicles’ ability to

^{*}Department of Electrical and Electronics Engineering, Dr. MGR Educational and Research Institute, Chennai, Tamil Nadu, India.

[†]Department of Electrical and Electronics Engineering, Dr. MGR Educational and Research Institute, Chennai, Tamil Nadu, India.

[‡]Department of Electronics and Communication Engineering, Dr. MGR Educational and Research Institute, Chennai, Tamil Nadu, India.

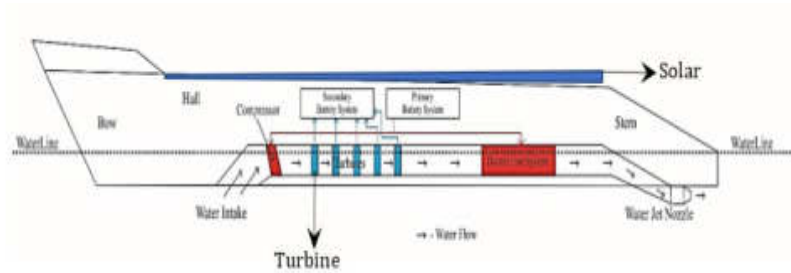


Fig. 3.1: Schematic Conceptual Design

conduct prolonged monitoring or surveillance operations autonomously, necessitating frequent data transfer and processing, further exacerbating the reliance on support infrastructure [5].

In light of these challenges, there is a pressing need to develop innovative solutions that enhance the autonomy, endurance, and operational efficiency of UUVs and AUVs. By addressing the limitations associated with battery capacity, data storage, and operational autonomy, researchers and engineers aim to unlock the full potential of unmanned underwater vehicles, paving the way for transformative advancements in ocean exploration, environmental monitoring, and maritime security.

This paper explores the current state of UUV and AUV technology, identifies key challenges and limitations, and proposes novel approaches to overcome these obstacles [6]. Through interdisciplinary collaboration and technological innovation, the goal is to usher in a new era of autonomous underwater exploration and observation, facilitating advancements in scientific research, commercial applications, and national defense strategies.

2. Related Work. A remodeled acoustic energy decay model preserved relative in acoustic energy attenuation inverse of distance square is used to generate training data. Multilayer perceptron (MLP) is the model to train these data and predicts accurate relative 3D space coordinates [7].

In depth discussion about AI chips and AI hardware. An improved controller design method based on echo state network with delay output (DO- ESN) is proposed for designing the controller of a class of nonlinear system [8]. The survey provides a comprehensive literature review on combined MBC-ANN techniques that are suitable for UAV flight control, i.e., low-level control [9].

The objective is to pave the way and establish a foundation for efficient controller designs with performance guarantee. The application of artificial intelligence (AI) in unmanned aerial vehicle (UAV) is discussed [10].

The basic connotation of AI technology is introduced. Review the history and classification of machine learning, and talk about the most recent applications of machine learning to UAVs for autonomous flight [11].

Paper reviews AI-enabled routing protocols designed primarily for aerial networks, including topology predictive and self-adaptive learning-based routing algorithms, with an emphasis on accommodating highly dynamic network topology [12]. An algorithm of a model reference adaptive controller for nonlinear systems based on Radial Basis Function (RBF) Neural Networks (NN) is proposed [13].

A nonlinear adaptive controller for an unmanned aerial vehicle (UAV) has been developed using Echo State Network (ESN), which is a form of three-layered recurrent neural network (RNN) [14]. Application of echo state network (ESN) for the nonlinear control of a fixed-wing unmanned aerial vehicle (UAV) is presented. The data required for the network training is generated using a validated flight dynamics model of the UAV [15].

3. Methodology. The project describes to harvest dual renewable energy from solar and hydro-turbine, the fig.3.1.a shows the schematic figure of the conceptualized design.

The proposed design integrated with solar panel and whereas the hydro flow duct turbine will be mounted at water-line area at the bottom portion [16]. The energy generated from solar power pack will be stored in the solar back-up battery, whereas the platform cruises the water flows in to duct chamber and the water streams cuts the turbine fin blades and rotational flow of the shaft coupled with the generator will produce the hydro energy and stores in the battery [17]. Functional flow of the system describes the solar power pack

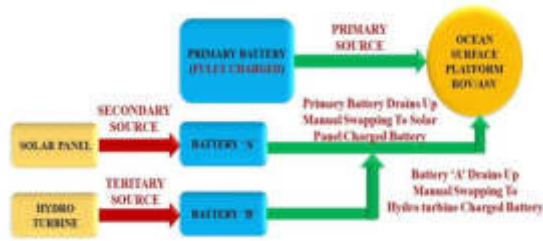


Fig. 3.2: Functional Lay-out



Fig. 3.3: Indigenously developed DC generator

system and hydro-duct flow turbine coupled to the ocean surface platform integrated with Battery Management System (BMS) and Switch Over Circuit (SOC) will enables the system to monitor the battery management and switching of the battery from one source to another [18].

3.1. Battery Charging. The battery needed to be charged in volts of 1.5 times so 24V. The amp needed must be at least 50% of the current rating of the battery, so we choose 18A since this can only be made in our hull. To increase the amp, we need to increase the solar cells [19].

3.2. Selection of DC Generator. The DC generator selected is practically constant speed, regardless of the load, which would produce consistent power generation [20]. The generator load calculations are arrived to the speed significant to produce the power output required, the fig.3.2.a. Shows the indigenously developed DC generator, which produces 300 watt with 15 VDC. The generator is a PMDC (Permanent Magnet Direct Current Generator) [21].

3.3. Selection of Battery Pack. The battery pack unit will be a three-unit system, where the primary battery will be initially full charged and battery-A & Battery-B is charging through solar panel and hydro duct turbine. The lithium-ion battery was selected for efficient charging and customized indigenously developed with 14.8V/18Ah rating. The system architect was conceptualized from above parametrical studies and flow of the working was schematized.

The system architecture consists of 4 major packs as shown in Fig 3.5:

- Battery Management pack
- Energy (solar and hydro) harvesting pack
- Control and communication pack
- Propulsion pack

The functional working of the system is that the microcontroller-processing unit will be charging from solar panel and hydro-turbine through generator to the battery sources initially assigned will be charging. The Battery Management System (BMS) will be monitoring the battery source, as initial primary battery drains out the Switch over Circuit (SOC) will be switching the primary battery source to battery-A and battery-B



Fig. 3.4: Battery pack unit

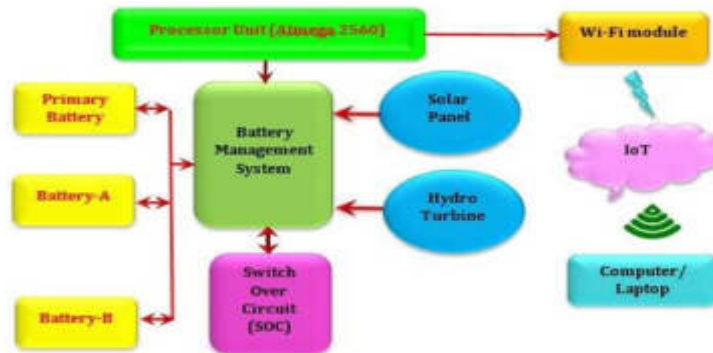


Fig. 3.5: System Architecture

and vice-versa the charging will be relayed from one source battery to other. The wireless Wi- Fi module will be transmitting the data to shore computer/ laptop.

3.4. Solar Energy Harvesting System Design. The connection of solar arrays selected with 156*156 mm monocrystalline cell is follows:

- Connecting 48 cells in series with 4 cells in an array gives 24V/9A. -12 Arrays
- Connecting another 48 cells in parallel with 4 cells in array gives 24V/18A-24 Arrays.

In the existing hull the 24 arrays can be placed in both sides and in middle bars. Each array would be in a square configuration and also the system would be engaged in a straight line of arrays on the bar with angled rotating system as per the sun movement in the sky.

The basic idea developed was to create a set of arrays which would be able to independent and connected in series or parallel which ever fitted the role accordingly. It was devised to make sure that in an active war zone, if any one array also gets damaged the remaining arrays would support the recharging system. This concept is meant to increase the reliability and endurance and not by using a single panel which would be a liability.

The mount for the solar system will have 5 parallel cylinder pipes and 2 perpendiculars main structure. The structure is strategically mounted on to the three bars by means of adjustable clamps and tightly mounted. On top of each pipe, one end of the pipe will have a bearing housing to hold the shaft that is clamped with the solar array. The other end is mounted with a servo motor housing where it can be hinged with the rotating shaft. The length of the occupied solar area in the middle section is 1600mm and breadth is 630 mm approximately. The servomotor is connected to the micro controller with input feed data according to the azimuth angle of the sun's rays the solar panels will be aligned according to the highest efficiency angle.

4. Results and Discussion. The circuit designed, were developed shown in fig 4.1 which is significant function for switching between input renewable energy sources and batteries, based on operating conditions. The Battery Management System (BMS) with Switch Over Circuit (SOC) manages to monitor the operation of the Battery with high voltage will be utilized for load and low voltage for charging and the other battery in idle.

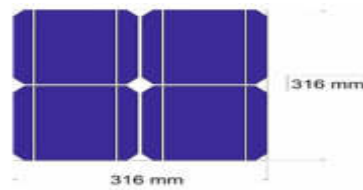


Fig. 3.6: Solar array layout of single panel

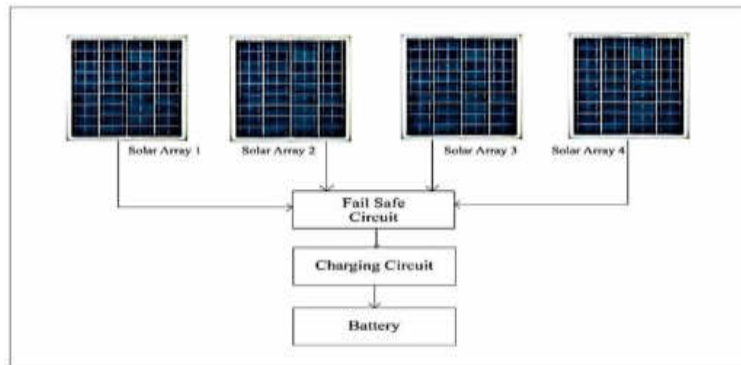


Fig. 3.7: Basic block diagram of the distributed solar array concept

The developed circuit focus on the power produced from solar energy and hydropower which stored in batteries. Since battery is the primary component for energy supply, we should have an alternate source for emergency. Hence, a second battery is used in case of low power or a failure in first battery. Thus, the SOC circuit switch between two batteries, after switching, the disconnected battery will be charged using the power from DC Generator.

The fully charged battery will be automatically cut-off from DC Generator using transistor and voltage regulator. Fig. 4.1 indicates the test evaluation carried out on the electronic hardware unit developed, which the Visual Basics (VB) displays the BMS and SOC working and data stored.

The figure 4.2 displays data from hardware testing of the Battery Management System (BMS) with the Switch Over Circuit (SOC). The plotted data likely includes voltage levels, current measurements, or other relevant parameters monitored during the testing phase. The trends in the data could provide insights into the performance and efficiency of the BMS and SOC in managing the battery system under different operating conditions.

Fig. 4.3 illustrates the voltage levels of the solar power system over time. The curve might show how the voltage fluctuates throughout the day as solar energy is harvested and stored in the battery. The trend could reveal patterns in solar power generation and help assess the effectiveness of the solar panel array in charging the battery under varying sunlight conditions.

The web-page data display (Fig 4.4) presents real-time or logged data from the IoT monitoring system. It includes parameters such as battery voltage, energy consumption, or system status, transmitted wirelessly to a web interface for remote monitoring and analysis. The displayed data provides valuable insights into the performance and operation of the renewable energy harvesting system, facilitating remote monitoring and management.

The trial fetches the solar panel data and performance of BMS with SOC were tested, and stored data were evaluated with IoT monitoring were checked.

- Developed web page performance for getting values from the boat while running were interpreted.
- The Values will be automatically stored in an Excel sheet



Fig. 4.1: BMS and SOC connected to Primary Battery, Solar battery, and Turbine battery

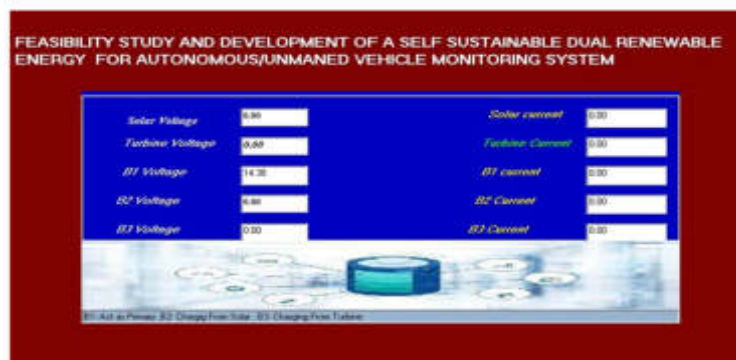


Fig. 4.2: BMS with SOC hardware testing data

- Whenever we need the data, the data sheet can be downloaded.

The performance of hydro-duct turbine was not evaluated in the phase I trial, since the floating platform (existing platform within the University) were not compatible, due to that performance of the hydro-duct turbine study were not carried out. Feasibility study on dual-renewable energy harvesting were carried out successfully. As Dual renewable energy source emphasis, as solar power system drops, the hydro-turbine energy charged system utilization will fetch the Ocean surface platform further.

5. Conclusion. Earlier the existing floating platform available with university has been utilized, where non- compatibility of the platform was studied. Further, a 50W platform was designed. Overall achievement

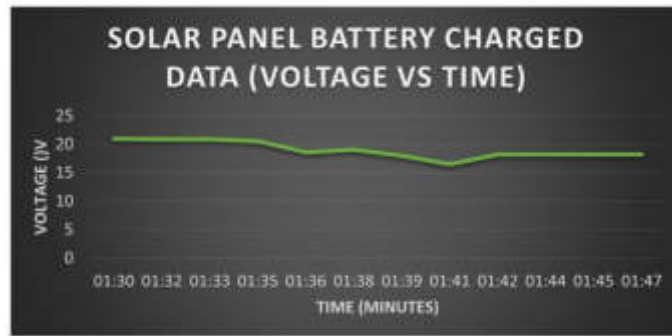


Fig. 4.3: Solar Power Charged Graph Plotted Voltage vs Time

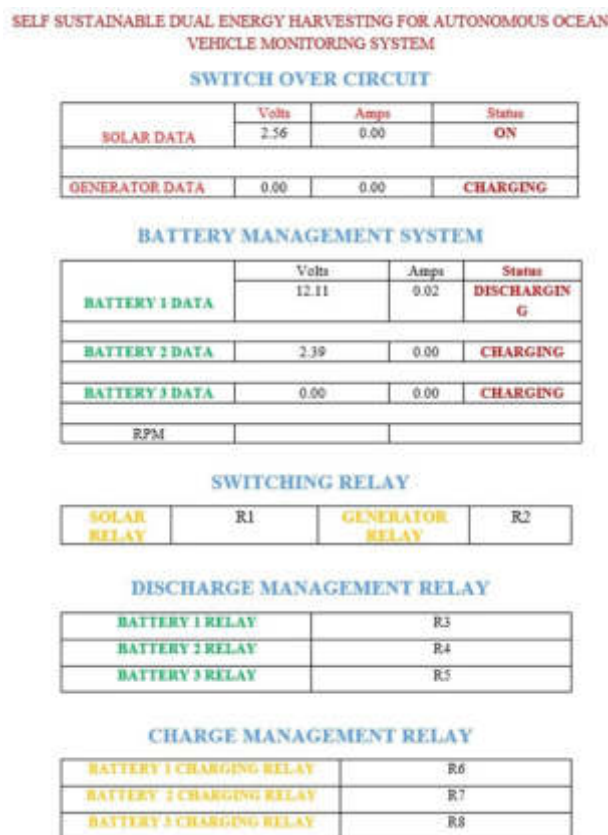


Fig. 4.4: IoT Web-Page Data Display (IoT web link: <https://in.000webhost.com/cpanel-login>)

of the project is in case of absence of solar power, the power generated during running of the boat, the power stored in the battery-B through turbine rotation will provide power after sunset. Solar system was developed and integrated on the boat. Feasibility study on dual-renewable energy harvesting was carried out. Technology demonstrated can be adopted to a bigger ocean surface vehicle will enhance the more efficient result. Higher rate of power harvest feasible with larger platform. Incorporating an ocean surface platform with ducted Hull at the bottom, for hydro turbine integration will increase flow more laminar and better energy produce can be achieved.

REFERENCES

- [1] Rowley, Jack. "Autonomous Unmanned Surface Vehicles (USV): A Paradigm Shift for Harbor Security and Underwater Bathymetric Imaging." OCEANS 2018 MTS/IEEE Charleston. IEEE, 2018.
- [2] Wang, Peng, Xinliang Tian, Wenyue Lu, Zhihuan Hu, and Yong Luo. "Dynamic Modeling and Simulations of the Wave Glider." Applied Mathematical Modelling 66 (February 2019): 77–96.
- [3] Liu, Zhixiang, Youmin Zhang, Xiang Yu, and Chi Yuan. "Unmanned Surface Vehicles: An Overview of Developments and Challenges." Annual Reviews in Control 41 (2016): 71–93. <https://doi.org/10.1016/j.arcontrol.2016.04.018>.
- [4] Blaich, Michael, Stefan Wirtensohn, Markus Oswald, Oliver Hamburger, and Johannes Reuter. "Design of a Twin Hull Based USV with Enhanced Maneuverability." IFAC Proceedings Volumes 46, no. 33 (2013): 1–6.
- [5] Johnston, Phil, and Mike Poole. "Marine Surveillance Capabilities of the AutoNaut Wave-Propelled Unmanned Surface Vessel (USV)." OCEANS 2017 - Aberdeen, June 2017.
- [6] Wang, Jianhua, Wei Gu, and Jianxin Zhu. "Design of an autonomous surface vehicle used for marine environment monitoring." 2009 International Conference on Advanced Computer Control. IEEE, 2009.
- [7] Baseer, Mohammad Abdul, Venkatesan Vinoth Kumar, Ivan Izonin, Ivanna Dronyuk, Athyoor Kannan Velmurugan, and Babu Swapna. "Novel Hybrid Optimization Techniques to Enhance Reliability from Reverse Osmosis Desalination Process." Energies 16, no. 2 (2023): 713.
- [8] Rice, J., L. A. Gish, J. Barney, Z. Gawboy, B. Mays, L. Moore, and A. Nickell. "Design and Analysis of an Improved Wave Glider Recovery System." OCEANS 2016 MTS/IEEE Monterey, September 2016.
- [9] Jia, Li Juan, Xuan Ming Zhang, Zhan Feng Qi, Yu Feng Qin, and Xiu Jun Sun. "Hydrodynamic Analysis of Submarine of the Wave Glider." Advanced Materials Research 834–836 (October 2013): 1505–11.
- [10] Singh, Yogang, S. K. Bhattacharyya, and V. G. Idichandy. "CFD approach to modelling, hydrodynamic analysis and motion characteristics of a laboratory underwater glider with experimental results." Journal of Ocean Engineering and Science 2.2 (2017): 90-119.
- [11] Shalini, A., L. Jayasuruthi, and V. VinothKumar. "Voice recognition robot control using android device." Journal of Computational and Theoretical Nanoscience 15, no. 6-7 (2018): 2197-2201.
- [12] Wang, Huan, Xiaoxu Du, and Baoshou Zhang. "Propulsive Performance Analysis of Underwater Flapping Multi-foil." OCEANS 2019-Marseille. IEEE, 2019.
- [13] Moeller, Dietmar P. F. "Parameter Identification of Dynamic Systems." Mathematical and Computational Modeling and Simulation, 2004, 257–310.
- [14] V Yuanbo, TIAN Xinliang, LI Xin, SONG Chunhui. Two-dimensional numerical simulation of NACA 0012 flapping foil hydrodynamics[J]. Chinese Journal of Ship Research, 2018, 13(2): 7-15.
- [15] King, Marlene Judith Taranger. Passive Oscillating Foils for Additional Propulsion under Calm Conditions. MS thesis. NTNU, 2019.
- [16] Liu, Jialun, and Robert Hekkenberg. "Sixty years of research on ship rudders: effects of design choices on rudder performance." Ships and offshore structures 12.4 (2017): 495-512.
- [17] Reddy, K. Hemant K., Ashish K. Luhach, V. Vinoth Kumar, Sanjoy Pratihar, Deepak Kumar, and Diptendu S. Roy. "Towards energy efficient Smart city services: A software defined resource management scheme for data centers." Sustainable Computing: Informatics and Systems 35 (2022): 100776.
- [18] Dennler, G., S. Bereznev, D. Fichou, K. Holl, D. Ilic, R. Koeppe, M. Krebs, et al. "A Self-Rechargeable and Flexible Polymer Solar Battery." Solar Energy 81, no. 8 (August 2007): 947–57.
- [19] Sollesnes, Erik, Ole Martin Brokstad, Rolf Kla boe, Bendik Vagen, Alfredo Carella, Alex Alcocer, Artur Piotr Zolich, and Tor Arne Johansen. "Towards Autonomous Ocean Observing Systems Using Miniature Underwater Gliders with UAV Deployment and Recovery Capabilities." 2018 IEEE/OES Autonomous Underwater Vehicle Workshop (AUV), November 2018
- [20] Sadhasivam, Jayakumar, V. Muthukumaran, J. Thimmia Raja, V. Vinothkumar, R. Deepa, and V. Nivedita. "Applying data mining technique to predict trends in air pollution in Mumbai." In Journal of Physics: Conference Series, vol. 1964, no. 4, p. 042055. IOP Publishing, 2021.
- [21] H.R Cao, Y. Liu, X.J Yue and W.J. Zhu, "Cloud-Assisted UAV Data Collection for Multiple Emerging Events in Distributed WSNs", Sensors, vol. 17, no. 8, pp. 1818, 2017.

Edited by: Mahesh T R

Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions

Received: Jan 8, 2024

Accepted: Apr 28, 2024



HYBRID ELECTRIC VEHICLE ENERGY MANAGEMENT STRATEGY BASED ON GENETIC ALGORITHM

YINGZHE LUO* AND CHAOXIONG FAN†

Abstract. A genetically optimized fuzzy control algorithm is used to realize the parallel hybrid system's real-time energy distribution between the engine and the electric motor. This paper first adopts fuzzy control to improve the robustness and real-time performance of the entire system. Then, a membership function optimization method based on a genetic algorithm is proposed by simulating the working state of CYCUDDS. Simulation experiments show that compared with unoptimized fuzzy control, the improved fuzzy controller can improve the vehicle's fuel economy and prevent continuous battery discharge, thereby improving battery endurance.

Key words: Genetic Algorithm; Hybrid electric vehicle; Battery optimization; Energy management; Fuzzy control

1. Introduction. Hybrid vehicles use a design that combines an internal combustion engine with an electric transmission, making the structure of the entire vehicle more complex. As the core of new energy vehicle control, its energy management method is the focus of its research. Adjusting and operating the motor within the effective range can save energy and reduce emissions. Literature [1] studies the operation mode of hybrid drive military hybrid vehicles. Improve vehicle fuel economy by building logical connections and state transitions between operating modes. However, its theoretical basis is based on experience. Literature [2] established a regular energy consumption control model for HEV and used algorithms such as genetic algorithm and quadratic programming to optimize the model's threshold value to improve the vehicle's fuel economy. However, the author only optimized it under local working conditions. Literature [3] uses wavelet analysis to separate high-frequency and low-frequency signals in load forecasting. Use wavelet transform to establish load forecasting model. using neural networks to train and predict high-frequency and low-frequency signals. The project achieved a 2.37% reduction in fuel consumption. Literature [4] combines variable prediction area RBF network and Q-learning to construct a battery state-of-charge energy management strategy under dynamic operating conditions. The project results achieved improvements in fuel economy. Therefore, this project plans to take the single-row star hybrid vehicle system as the research object and use the working condition information obtained from the actual vehicle test. Working condition information trains, the vehicle speed prediction model [5]. Genetic algorithm fuzzy control and other methods are used to establish the optimal optimization control problem in the short-term forecast area. A hybrid power system's energy consumption management method is studied using working condition speed prediction and differential evolution methods. The hybrid vehicle system's energy-saving and emission-reduction goals can be achieved.

2. Vehicle parameter matching and modeling. The main parameters of the vehicle in the hybrid system are directly related to the working status of the system, so how to select appropriate parameters is a critical link in the development of the entire system.

2.1. Maximum speed is used as the basis to calculate the total power of the system. When half the vehicle's weight is $m_a = 1750$ kg, the maximum speed V_{\max} of the vehicle is expected to reach 200 km/h, and the total electric power $P_{v \max}$ with speed as the target can be determined by formula (2.1).

$$P_{v \max} = \frac{1}{\zeta_T} \left(\frac{m_a c h V_{\max}}{3600} + \frac{Z_\beta A V_{\max}^3}{76140} \right)$$

*School of Mechanical and Electrical Engineering, Shijiazhuang University of Applied Technology, Shijiazhuang, Hebei, 050081, China (Corresponding author, pig1e1987@163.com)

†Great Wall Motor Company Limited, HAVAL R&D Center, Baoding, Hebei, 071000, China

ζ_T is the overall efficiency of the entire power system, and 0.9 is selected as the calculation parameter. c represents the gravitational factor, $N/kg.h$ is the rolling resistance factor. Z_β is the drag factor. A is the headwind side of the car, m^2 .

2.2. Determine the total power based on the maximum gradient. The obstacle-crossing performance of a car is the maximum slope it can climb on a good road. Due to the total weight $m_h = 2000$ kg, its maximum climbing slope can reach $\theta_{\max} = 35\%$ when the speed is $V_\delta = 30$ km/h, and the total power $P_{\delta \max}$ taking the slope as the target is determined by the formula (2.2).

$$P_{\delta \max} = \frac{1}{\zeta_T} \left(\frac{m_a c h V_\delta}{3600} + \frac{Z_\beta A V_\delta^3}{76140} + \frac{m_a c \theta_{\max} V_\delta}{3600} \right)$$

The total power $P_{\delta \max} = 67$ kW with the slope as the target can be calculated from formula (2.2). The maximum power of the power supply must meet the following dynamic index:

$$P_{\max} > \max(P_{v \max}, P_{\delta \max})$$

3. Battery model. A lithium-ion battery energy storage system is proposed using the constant rate method. If it is driven at a constant speed at a speed of $V_e = 60$ km/h in a pure electric state, it is expected to cover 60 kilometers. In this process, the electric motor supplies all the power required by the vehicle [6]. The total capacity required by the battery at this time is obtained through formula (2.4).

$$\begin{cases} W_\varphi \zeta_\varphi = \int_0^T P_m dT \\ P_m = \frac{V_e}{3600 \zeta_T} \left(m_a c h + \frac{Z_\beta A V_e^2}{21.15} \right) \\ T = 3600 \frac{S}{V_e} \\ Z = \frac{1000 W_\varphi}{U} \end{cases}$$

ζ_φ is the battery discharge efficiency. T is the driving time. P_m is the motor working power, kW. W_φ is the total energy required by the battery, kW · h. Z is the battery capacity. U is the voltage level, V.

3.1. Transmission system model. The maximum power point also corresponds to the maximum rotational speed, so when the maximum vehicle speed $V_{\max} = 200$ km/h, the rotational speed n is 6000 .

$$V_{\max} = 0.377 \frac{nd}{\theta_c \theta_0}$$

d is the radius of the wheel rotation. θ_c is the speed of the transmission system. θ_0 is the transmission ratio of the main reduction gear transmission. The maximum reduction ratio calculation formula is based on the maximum climbing slope.

$$\frac{T_{\max} \theta_c \theta_0 \zeta_T}{d} = m_h c h \cos \delta + m_h c \sin \delta$$

4. Calculation of vehicle torque. The torque required by the vehicle is calculated based on the vehicle operating equilibrium equation of the vehicle power system [7]. This provides a basis for the formulation of energy control plans.

$$T_{tq} = \frac{d}{\theta_c \theta_0 \zeta_T} \left(Gh + \frac{Z_d A V_e^2}{21.15} + Gi + \delta m \frac{du}{dt} \right)$$

T_{tq} is the torque required by the entire vehicle. θ_c is the transmission ratio. θ_0 is the primary reducer speed ratio. ζ_T is the drive train efficiency. G is the gravity of the vehicle. h is the rolling resistance coefficient. Z_d is the air resistance coefficient. A is the windward area of the vehicle. u is the vehicle speed. θ is the slope. δ is the car rotation mass conversion coefficient. m is the vehicle mass.

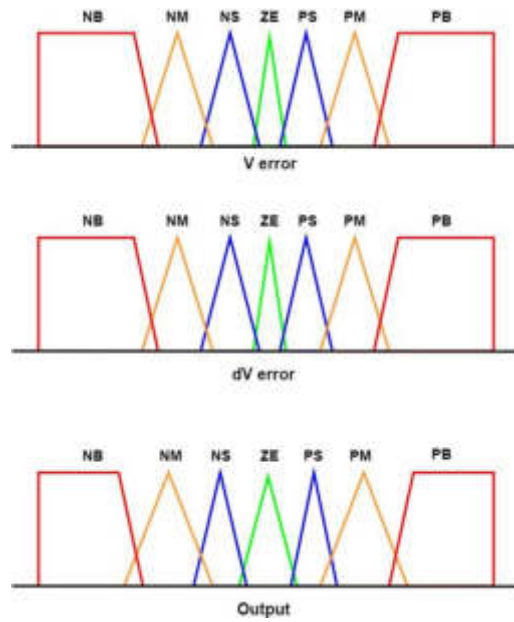


Fig. 5.1: *Input variable membership function.*

5. Fuzzy control algorithm. The energy flow distribution rules of internal combustion engines and electric motors under different working conditions are studied to achieve adequate control of energy saving and emission reduction [8]. Therefore, this paper proposes a method based on fuzzy control.

5.1. Input and output of variables. The fuzzy control method can ensure that the engine operates within the efficient range and the balance of the battery charge state [9]. This paper uses the torque S_d required by the car and the battery charging status as input quantities and carries out fuzzy control based on this. Take the engine torque S_a as the output variable.

5.2. Membership function. In designing the fuzzy controller, we must first determine its degree of belonging and, secondly, determine the segmentation point appropriately and try to select the appropriate fuzzy subset [10]. The membership degrees of input and output are partitioned based on the actual working conditions of parallel hybrid vehicles. The demand torque S_d is divided into five fuzzy subgroups $\{\alpha, \beta, \gamma, \delta, \varepsilon\}$, whose domains are $[-60, 60]$, respectively. The adaptive battery state of charge of the battery pack is divided into five fuzzy subsets $\{\zeta, \eta, \theta, l, \kappa\}$, whose domains are $[0, 1]$, respectively. The engine's range output depends on its accurate output torque, and its range is $\{40, 43, 45, 48, 51, 53, 56, 58, 60\}$. According to the characteristics of the fuzzy sub-region, the ladder-like membership function is selected to obtain accurate control output. Figure 3.1 shows the input variable membership function.

5.3. Fuzzy control rules. It is necessary to fully use the control system's actual work experience and fully consider the input and output characteristics of the controlled system. One is that the engine drives the car independently while the battery is charging. When the required torque exceeds the maximum torque the engine can carry, the motor will assist the engine in driving the car. Second, when the battery charge is small, and the power characteristics of the electric vehicle are maintained, the engine distributes appropriate torque to the battery for charging. When the maximum torque equals the required torque, the engine's power characteristics should first be considered [11]. Third, the engine will automatically stop when the speed is low. In this case, the engine runs independently, reducing fuel consumption. Fourth, the engine and motor drive the car together to achieve high-load operation. An input/output relationship mapping diagram (Figure 3.2) was developed for the abovementioned situation.

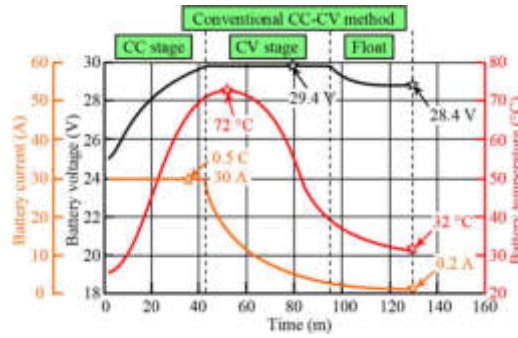


Fig. 5.2: *Input/output relationship map diagram.*

6. Membership function optimization based on fuzzy control. Due to the traditional manual judgment method, selecting member weights in fuzzy control systems is often subjective, and it isn't easy to achieve the overall optimal. Therefore, it must be optimized to achieve the system's optimal control effect [12]. It is often difficult to obtain optimal results using classical optimization methods for more complex problems, such as parameter optimization of membership functions. This paper proposes a multi-objective optimization method based on genetic algorithms at this time.

6.1. Initialize the overall. This paper proposes a fuzzy controller design method based on a genetic algorithm. Since the decimal chromosome has a short length, high accuracy, fast operation speed, and good stability to the population of mutation operations, we use the decimal encoding method. As can be seen from Figure 4.1, X1, X2, X3, and X4 each represent subdivisions corresponding to their respective functions. Since all lines of input and output need to be encoded, a one-dimensional decimal matrix with a length of 49 is finally generated [13]. In this matrix, X1-X20 is the code of the required torque membership function, and X21-X40 is the membership of the battery. Degree function codes, X41-X49, are engine output torque membership function codes.

6.2. Selection of fitness function. Optimum fuel consumption and minimum exhaust emissions conflict with each other in most operating ranges and are consistent in only a few. The engine must be kept within this range as much as possible to achieve optimal control [14]. For this purpose, the exhaust gas, fuel consumption and other indicators under actual working conditions are selected, and the optimized objective function is obtained by weighting each indicator.

$$G(x) = \frac{1}{\psi_1 + \psi_2 + \psi_3 + \psi_4} \times \left(\psi_1 \int_0^s \frac{Q}{\bar{Q}} dt + \psi_2 \frac{HC}{\bar{HC}} dt + \psi_3 \int_0^s \frac{NO_x}{\bar{NO}_x} dt + \psi_4 \int_0^s \frac{CO}{\bar{CO}} dt \right)$$

x represents a marker that corresponds to the chromosome number. $\psi_1, \psi_2, \psi_3, \psi_4$ represents the corresponding weighted value. Q is fuel consumption. HC stands for hydrocarbon emissions. NO_x stands for the release of nitrogen oxides. CO stands for the release of carbon monoxide. $\bar{Q}, \bar{HC}, \bar{NO}_x, \bar{CO}$ is the optimal reference value corresponding to each parameter. $\psi_1 = 0.7, \psi_2 = \psi_3 = \psi_4 = 0.1$ is used to determine the weight of each optimization index when performing optimization.

6.3. Operation parameter setting. The group size is directly related to species diversity and production efficiency. If it is too small, the species diversity of the species will decrease, and if it is too large, it will be detrimental to production. The roulette method is used for personal selection. When performing crosses, two cut-point crosses were selected, and the probability of crosses was 0.9; because the mutation rate of this variety is relatively low, it was set at 0.01; the maximum number of gene generations was set at 50.

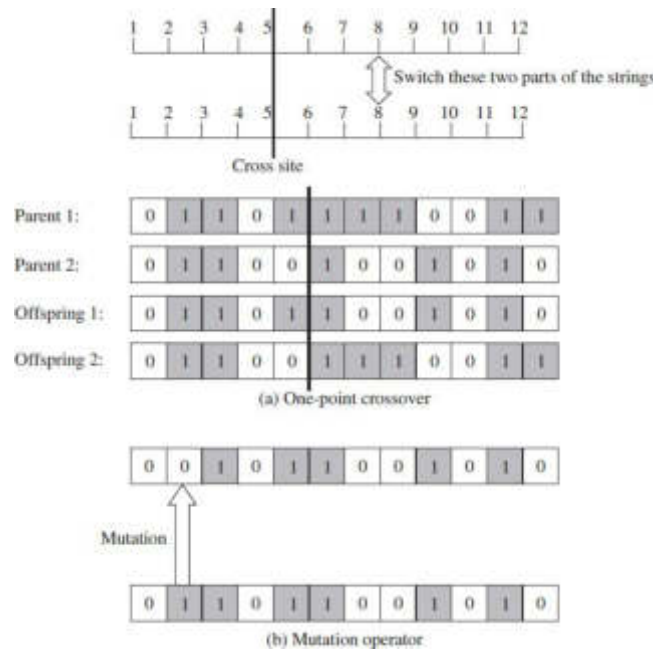


Fig. 6.1: *Decimal encoding method.*

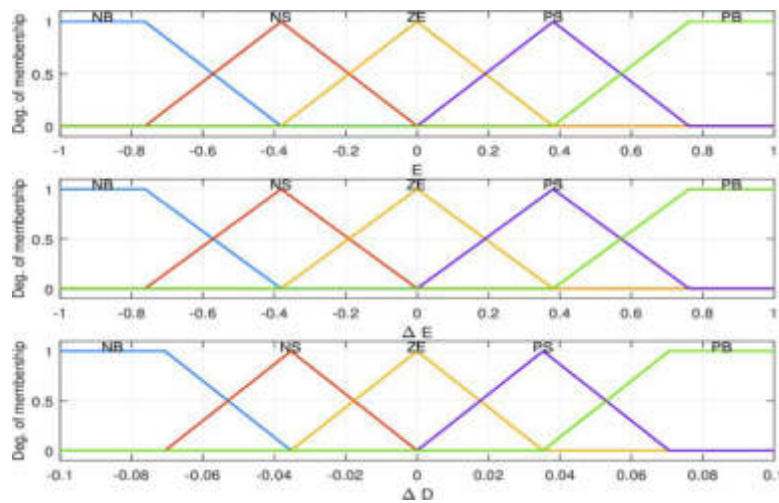


Fig. 6.2: *Optimized input variable membership function.*

6.4. Optimization results. Genetic algorithms optimize and adjust it continuously, taking parallel hybrid power as the object. Finally, the optimal membership function is given, as shown in Figure 4.2.

7. Simulation analysis. This project plans to use Intel(R) Core (TM) i7-9700CPU@3.00GHz computer, use Python language to implement the rate prediction of GRU-NN and implement rapid prediction of the model through Python programming [15]. Call the data of the genetic algorithm speed prediction model for simulation. The primary performance indicators of the car and the relevant parameters of the differential evolution algorithm are shown in Table 5.1 and Table 5.2.

This project plans to use a combination of genetic and differential evolution algorithms to compare fuel

Table 7.1: *Vehicle parameters.*

Part	Parameter	Numerical value
	Vehicle mass/kg	1425
Vehicle parameters	Windward area/m ²	1.819
	rolling resistance coefficient	0.014
	drag coefficient	0.313
	Tire radius/m	0.299
	Main reducer transmission ratio	4.094
	Maximum speed/(r · min ⁻¹)	6250
Engine	Maximum output power/kW	58
	Maximum speed/(r · min ⁻¹)	6250
Motor MG1	Maximum torque/(N · m)	318
Motor MG2	Maximum speed/(r · min ⁻¹)	5729
	Maximum torque/(N · m)	57
	Capacity/(A · h)	6.8
Power Battery	SOC	0.45 0.75
	Sun gear teeth	31
Planetary gear	Number of teeth of the ring gear	81

Table 7.2: *A-DE algorithm parameters.*

Parameter	Numerical value
Initial population size	21
Number of iterations	100
Intrinsic coefficient of variation	0.52
Crossover probability	0.31

consumption with three optimization methods: optimal system dynamics optimization and minimum equivalent fuel consumption, in which the adaptive battery state of charge is taken as 0.55. This project plans to use a reverse algorithm to analyze the optimal driving parameters of each stage and the minimum fuel consumption value of each stage from the final state. A forward iteration method obtains the optimal driving mode under each operating condition. And optimize each operating condition of this mode to achieve the optimal configuration of the best operating mode for each working condition [16]. This enables effective control of the startup MG1 and MG2 operating modes. Fuel consumption is shown in Table 5.1. The power distribution of the engine, motor MG1 and motor MG2 is shown in Figure 5.1, and the motion trajectory of the adaptive battery state of charge is shown in Figure 5.2.

The final value of the adaptive battery state of charge controlled by the genetic algorithm fuzzy control is 0.5448, the fuel consumption is 3.4510 L/100 km, and the fuel economy is 93.04%, which is 4.55% lower than the differential evolution and dynamics planning (Table 5.3). Figure 5.1 shows that the starting power of the differential evolution and dynamic planning strategies is maximum in the first 300 seconds, while the genetic algorithm fuzzy control maintains the engine power's stability, and power planning reduces the engine starting time. The MG1 motor using differential evolution produces large oscillations during operation, but using genetic algorithm fuzzy control can keep the output of MG1 stable. Therefore, the adaptive battery state of charge under fuzzy control in Figure 5.2 will experience more excellent attenuation in the early stages, and the power planning will undergo more significant changes. At the same time, the differential evolution algorithm maintains a stable, highly adaptive battery charge—electrical state. In the subsequent high-speed range, the genetic algorithm fuzzy control is used to maintain a more stable number of starts and a more appropriate engine power. At the same time, differential evolution has problems with the high number of starts, high starting power consumption, and adaptive battery state-of-charge changes during working conditions—big questions. During the entire driving process, the output power of the engine when starting and stopping the dynamic planning

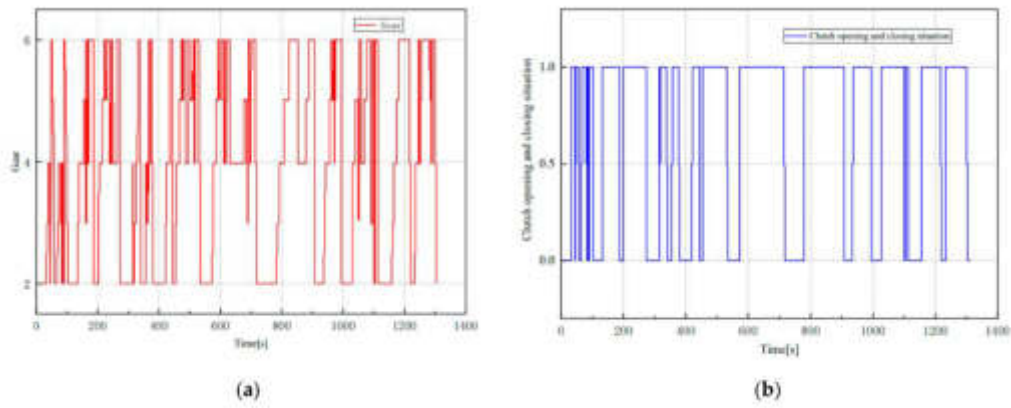


Fig. 7.1: Power distribution under different methods and strategies.

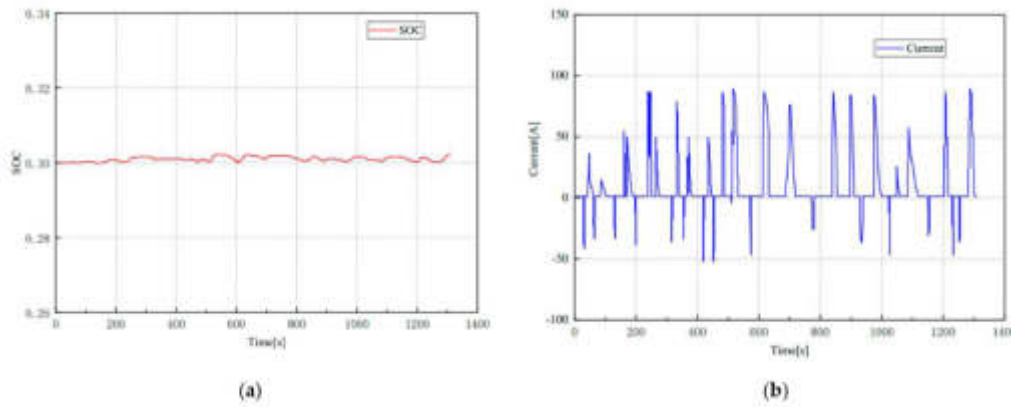


Fig. 7.2: SOC trajectory.

Table 7.3: Experimental results.

Method strategy	Final value SOC (T)	Fuel consumption/[L · (100km) ⁻¹]	Fuel economy/%
Dynamic programming	0.573	3.361	100
Genetic Algorithm Fuzzy Control	0.568	3.595	96.917
Differential Evolution Algorithm	0.563	3.766	88.490

strategy is greater than the other three control methods, resulting in high fuel consumption, while the genetic algorithm fuzzy control always maintains a small adaptive battery state-of-charge change interval, thereby preventing the harm to the battery caused by changes in the adaptive battery state of charge.

8. Conclusion. It was found that all three solutions can keep the battery’s state of charge at equilibrium. Before and after optimization, the SOC value decreased by 0.0427 and 0.0063. By improving this method, the battery’s state of charge changes significantly, thereby improving the battery’s endurance. In addition, the battery’s state of charge before optimization was finally 0.6865. After optimization, the battery state of charge showed a trend from low to high over time and finally stabilized around 0.7260, indicating that the optimized

electric vehicle braking system performs better. The working condition distribution of the optimized engine is relatively scattered, making it difficult to always be in an efficient working range when facing various working conditions. Each working point can work in a relatively dense area, and most of the working points are located in the high-efficiency area using this method.

REFERENCES

- [1] Guzs, D., Utans, A., Sauhats, A., Junghans, G., & Silinevics, J. (2022). Resilience of the Baltic power system when operating in island mode. *IEEE Transactions on Industry Applications*, 58(3), 3175-3183.
- [2] Ding, N., Prasad, K., & Lie, T. T. (2021). Design of a hybrid energy management system using designed rule-based control strategy and genetic algorithm for the series-parallel plug-in hybrid electric vehicle. *International Journal of Energy Research*, 45(2), 1627-1644.
- [3] Sidharthan Panaparambil, V., Kashyap, Y., & Vijay Castelino, R. (2021). A review on hybrid source energy management strategies for electric vehicle. *International Journal of Energy Research*, 45(14), 19819-19850.
- [4] Li, J., Zhou, Q., Williams, H., Xu, H., & Du, C. (2021). Cyber-physical data fusion in surrogate-assisted strength pareto evolutionary algorithm for PHEV energy management optimization. *IEEE Transactions on Industrial Informatics*, 18(6), 4107-4117.
- [5] Wang, L., Li, M., Wang, Y., & Chen, Z. (2021). Energy management strategy and optimal sizing for hybrid energy storage systems using an evolutionary algorithm. *IEEE Transactions on Intelligent Transportation Systems*, 23(9), 14283-14293.
- [6] Yuan, H. B., Zou, W. J., Jung, S., & Kim, Y. B. (2022). Optimized rule-based energy management for a polymer electrolyte membrane fuel cell/battery hybrid power system using a genetic algorithm. *International Journal of Hydrogen Energy*, 47(12), 7932-7948.
- [7] Raboaca, M. S., Bizon, N., & Grosu, O. V. (2021). Optimal energy management strategies for the electric vehicles compiling bibliometric maps. *International Journal of Energy Research*, 45(7), 10129-10172.
- [8] Zhang, J., Roumeliotis, I., & Zolotas, A. (2021). Nonlinear model predictive control-based optimal energy management for hybrid electric aircraft considering aerodynamics-propulsion coupling effects. *IEEE Transactions on Transportation Electrification*, 8(2), 2640-2653.
- [9] Ray, P., Bhattacharjee, C., & Dhenuvakonda, K. R. (2022). Swarm intelligence-based energy management of electric vehicle charging station integrated with renewable energy sources. *International Journal of Energy Research*, 46(15), 21598-21618.
- [10] Kandidayeni, M., Trovão, J. P., Soleymani, M., & Boulon, L. (2022). Towards health-aware energy management strategies in fuel cell hybrid electric vehicles: A review. *International Journal of Hydrogen Energy*, 47(17), 10021-10043.
- [11] Gharibeh, H. F., Yazdankhah, A. S., Azizian, M. R., & Farrokhifar, M. (2021). Online energy management strategy for fuel cell hybrid electric vehicles with installed PV on roof. *IEEE Transactions on Industry Applications*, 57(3), 2859-2869.
- [12] Król, A., & Sierpiński, G. (2021). Application of a genetic algorithm with a fuzzy objective function for optimized siting of electric vehicle charging devices in urban road networks. *IEEE Transactions on Intelligent Transportation Systems*, 23(7), 8680-8691.
- [13] Wang, X., Wang, R., Shu, G., Tian, H., & Zhang, X. (2022). Energy management strategy for hybrid electric vehicle integrated with waste heat recovery system based on deep reinforcement learning. *Science China Technological Sciences*, 65(3), 713-725.
- [14] Boukoberine, M. N., Donato, T., & Benbouzid, M. (2022). Optimized energy management strategy for hybrid fuel cell powered drones in persistent missions using real flight test data. *IEEE Transactions on Energy Conversion*, 37(3), 2080-2091.
- [15] Liu, Q., Lanfermann, F., Rodemann, T., Olhofer, M., & Jin, Y. (2023). Surrogate-assisted many-objective optimization of building energy management. *IEEE Computational Intelligence Magazine*, 18(4), 14-28.
- [16] Yang, C., Zha, M., Wang, W., Yang, L., You, S., & Xiang, C. (2021). Motor-temperature-aware predictive energy management strategy for plug-in hybrid electric vehicles using rolling game optimization. *IEEE Transactions on Transportation Electrification*, 7(4), 2209-2223.
- [17] Hou, Z., Guo, J., Xing, J., Guo, C., & Zhang, Y. (2021). Machine learning and whale optimization algorithm based design of energy management strategy for plug-in hybrid electric vehicle. *IET Intelligent Transport Systems*, 15(8), 1076-1091.

Edited by: Zhigao Zheng

Special issue on: Graph Powered Big Aerospace Data Processing

Received: Dec 12, 2023

Accepted: Dec 29, 2023



DATA PROTECTION AND PRIVACY PROTECTION OF ADVERTISING BASED ON CLOUD COMPUTING PLATFORM

ZHISHE CHEN*

Abstract. This paper uses the hybrid leapfrog algorithm to mine user information in encrypted advertisements effectively and intelligently. This method handles the nonlinearity of the original data by mapping it into kernel space. The representation of the original ciphertext in the kernel space is obtained by sparsely reconstructing the encrypted original advertising data. Build a corresponding scoring mechanism and select the best advertising data characteristics. The selected data were clustered using the data fuzzy clustering method based on the improved hybrid leapfrog. Set the adjustment coefficient to improve the local optimization performance of hybrid frog leaping. This algorithm uses the tightness and separation in genetic algorithms and constructs a fitness function to determine the clustering critical value. This enables the practical, intelligent mining of homomorphic passwords with privacy protection. Experimental results show that the method proposed in this article can effectively improve the convergence speed and accuracy of clustering. Improve Blowfish by combining multi-threading, sharing encryption and other methods. This enables encryption and decryption of large amounts of model data. The research of this project has very important research value in improving the security performance and effectiveness of cryptographic algorithms.

Key words: Privacy-preserving deep learning; Precise advertising; Secure computing; Privacy protection; Data mining

1. Introduction. Advertising data mining technology has become an emerging research direction. Advertising data mining technology has been widely used in many fields, such as engineering, scientific research, etc. This provides users with revenue. But when this method is used for ad mining, it can also have adverse effects, like the security of personal information. Technicians can speculate that it may contain private or sensitive information. There is a potential risk of leakage when mining advertising data with homomorphic passwords, so it must be protected. It is necessary to conduct data mining while ensuring the security of ciphertext information. Especially with the rapid development of network technology, scientific researchers can obtain massive amounts of information from various web pages. Traditional advertising user data mining methods to mine privacy protection data based on ciphertext may cause data leakage and affect the accuracy of data mining. Literature [1] uses homomorphic cryptography technology on the lattice to mine data. It implements data mining for privacy-preserving data cluster analysis. The user encrypts the data before transmitting it to the provider. After using Blowfish-based confidentiality and data mining technology, network service providers cannot obtain user data. Compared with existing data publishing methods, Geji data mining technology has significant advantages in information security. This algorithm ensures the accuracy of the spacing of integer encrypted ciphertext. Experiments have shown that the computing speed of the lattice mining algorithm is significantly higher than that of other algorithms, but there are also problems with low accuracy. Literature [2] proposed a time series data mining algorithm for differential privacy—screening of sequential patterns from candidate templates. Geometric principles are used to add noise to selected mode support values to interfere with their generation. The simulation results show that the proposed algorithm meets the differential confidentiality requirements. Although it has a good mining effect, the accuracy is not high [3]. This paper proposes homomorphic encryption data privacy protection technology and model privacy protection technology based on the Blowfish algorithm.

2. Homomorphic encryption technology. Homomorphic encryption is an encryption method based on computational complexity. Perform corresponding operations on the homomorphically encrypted data and then decode it. A similar conclusion was made for the unencrypted raw material [4]. Homomorphic encryption is a

*School of Information Science and Engineering, Wuchang Shouyi University, Wuhan, Hubei, 430064, China (Corresponding author, czs_works@163.com)

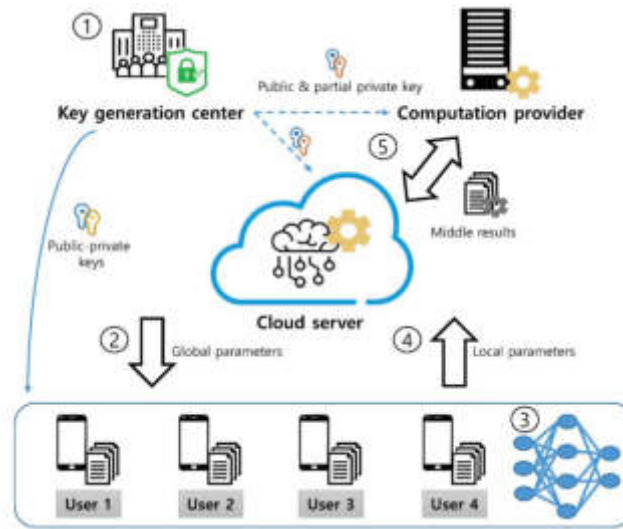


Fig. 2.1: Homomorphic encryption intelligent reasoning architecture.

new computing method that has emerged in recent years. It can effectively ensure the security of information transmission and calculation [5]. Homomorphic ciphers can be used to encode ciphertext directly. Ciphertext calculation eliminates the cumbersome process of decrypting large-scale ciphertext, then calculating, and then encrypting. This algorithm saves the amount of calculation while ensuring safety. This method can provide ideas for research on privacy protection issues of advertising users in the network environment. Any individual server can hide trained models. These models can be secretly exchanged between multiple servers to ensure model security. Figure 2.1 shows the intelligent inference architecture based on homomorphic cryptography.

The implementation process of homomorphic encryption advertising user data privacy protection technology for intelligent algorithms is as follows: (1) Use methods such as homomorphic addition and homomorphic multiplication to establish convolution and pool and approximate different excitation functions; (2) Based on security, Verify the security of ciphertext through comprehensive multi-party computation; (3) Study the security and privacy protection mechanism based on ciphertext. The above process applies to 6 different machine learning algorithm scenarios: convolutional neural network, BP neural network, logistic regression algorithm, SVM, linear regression algorithm and multi-layer perceptron [6]. The inference engine consists of three servers, which can automatically encrypt user information locally without decrypting it. It only needs to perform cryptographic operations on it and then feedback the results of its operations to the user. The user decrypts it to obtain inferred conclusions [7]. The user has no sense of the above inference. Only the user can encrypt it, ensuring the security of the entire computing process. This paper studies cryptographic intelligence algorithms based on Parlier's cryptography ideas. The algorithm of Parlier homomorphic cipher is:

$$z = \text{Enc}(u, s) = h^u s^m \bmod m^2$$

h and m are the password public keys; $s \in M.s$ and m are mutually prime. h is a random number much smaller than m^2 , and is generally more than 4000 digits, so its operation cost is very huge [8]. The calculation formula of Parlier homomorphic addition is:

$$\begin{aligned} \text{Enc}(u, s) \text{Enc}(v, \varepsilon) &= (h^u s^m \bmod m^2) (h^v \varepsilon^m \bmod m^2) = \\ &h^{(u+v)} (s\varepsilon)^m \bmod m^2 \text{Enc}(u+v, s\varepsilon) \end{aligned}$$

ε is the same random number as s . s is the random number for pairing u , and ε is the random number for pairing v . The formula for Parlier homomorphic multiplication is:

$$\text{Enc}(u, s)^\kappa = (h^u s^m \bmod m^2)^\kappa = h^{u\kappa} (s^\kappa)^m \bmod m^2 = \text{Enc}(u\kappa, s^\kappa)$$

κ is a sub-square number. Information was exchanged based on confidential sharing. The confidential multi-party computation method is used in this paper [9]. This algorithm allows multiple parties to jointly keep their value secret without destroying computing power [10]. This method divides a piece of data into several segments and does not display the original data when sharing. Two operation participants perform the same operation on an encryption key group and then reassemble it. On the user side, the private information u is decomposed into two parts, u_0 and $u_1, u = u_0 + u_1$. Then the two variables u_0 and u_1 are sent to the two servers D_0 and D_1 respectively [11]. Mere possession of u_0 and u_1 does not jeopardize the privacy of data u . The formula for decomposing tensor u is this:

$$u_0 = \text{share}(u, S) = S \bmod N$$

$$u_1 = \text{share} 2(u, S) = u - S \bmod N$$

S and N are both arbitrary numbers. Two servers Q_0 and Q_1 and one secondary server Q_2 are set up, which form a computing cluster. When creating a shared section, you should know how it is distributed [12]. To generate a secret share, you only need to separate the numbers that need to be converted into two values. People Q_0 and Q_1 first exchange half of the shares, and then use the parts they hold to perform calculations and transactions, and finally get the final result. Q_0 will pass α_1 to Q_1 , and Q_1 will pass β_0 to Q_0 . Because Q_0 cannot enter β_1 in Q_1 , it cannot determine the value of β . Adding is the most straightforward operation that can be performed through confidential sharing [13]. It can look like this:

$$\alpha + \beta = (\alpha_0 + \alpha_1) + (\beta_0 + \beta_1)$$

Equation (2.6) is rearranged by adding exchange rules and adding combination rules to:

$$\alpha + \beta = (\alpha_0 + \beta_0) + (\alpha_1 + \beta_1)$$

Q_0 solves for $\alpha_0 + \beta_0$ · Q_1 solves $\alpha_1 + \beta_1$ to ensure that Q_0 gets only part of β . Q_1 gets only part of α . The two parties doing the multiplication need to communicate with each other during the operation [14]. The notation above is used to define multiplication using secret sharing in the following way:

$$\alpha\beta = (\alpha_0 + \alpha_1)(\beta_0 + \beta_1) = \alpha_0\beta_0 + \alpha_0\beta_1 + \alpha_1\beta_0 + \alpha_1\beta_1$$

Q_0 may be responsible for $\alpha_0\beta_0$ and Q_1 may be responsible for $\alpha_1\beta_1$. Giving intermediate project $(\alpha_0\beta_1 + \alpha_1\beta_0)$ to Q_0 or Q_1 will cause certain security risks. This is because either of the two items can be added together, revealing the original α or β . For example, if Q_0 wants to solve $\alpha_0\beta_1$, then Q_0 must have a β_1 . Since Q_0 already contains β_0 , we can find the value of β . Shielding can be used for concealment. When partial concealment is required, new unknowns can be introduced into the parties so that the mask can be eliminated without affecting the final result when all are calculated. The third party Q_2 generates information that it is unwilling to share with the other party for confidentiality purposes [15]. That is, masking $Q_0\beta_1$ and $Q_1\alpha_0$. In this article, the shields are called d and k , and ζ and η are called the values of the shield. Multiplying $\alpha\beta$ by Q_0 is:

$$f_0 = dk_0 + d_0\eta + \zeta k_0 + \zeta\eta$$

The multiplication of Q_1 by $\alpha\beta$ is equal to:

$$f_1 = dk_1 + d_1\eta + \zeta k_1$$

Builds the mask coefficient from Q_2 · Q_2 generates 3 new numbers and divides them into different parts [16]. The first two digits are any number, and the third digit is the product of the first two digits. These values are obtained by subtracting the mask from the original data in the following way:

$$\zeta = (\alpha_0 - d_0) + (\alpha_1 - d_1)$$

$$\eta = (\beta_0 - k_0) + (\beta_1 - k_1)$$

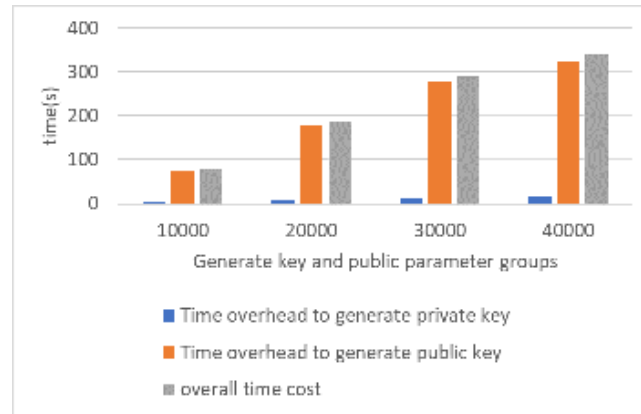


Fig. 3.1: *Key generation time cost.*

Q_2 transmits the values d_0 and k_0 to Q_0 , and d_1 and k_1 to Q_1 . Here the values of d_0 and k_0 take any integer, $d_1 = d - d_0, k_1 = k - k_0 \cdot Q_0$ then produces the $(\alpha_0 - d_0)$ part of ζ and the $(\beta_0 - k_0)$ part of $\eta \cdot Q_1$ produces $(\alpha_1 - d_1)$ regions of ζ and $(\beta_1 - k_1)$ regions of η ; then D and G exchange the ζ and η shares without revealing any information about α or β ; finally, the values are added to the corresponding equations. Combining (2.9) and (2.10) becomes:

$$f_0 + f_1 = \alpha\beta$$

Confidential sharing enables secure interaction with data. All expressions can be simulated using addition and homomorphic multiplication methods. Based on the two primitives of homomorphic addition and homomorphic multiplication, functions such as convolution, pooling, and approximation activation can be implemented, respectively. It can be applied to machine vision and linear or logistic regression fields. In practical applications, it is necessary to rewrite functions such as CNN, BP, MLP, logistic regression, etc., and use Taylor expansion approximation [17]. At the same time, the format of additive and multiplicative operations in homomorphic cryptography is maintained. This method is compatible with various mechanisms such as convolution, ReLU activation function, Maxpool and normalization, and can be combined with different inference methods to form corresponding safe computing solutions.

3. Analysis of experimental results.

3.1. The impact of Blowfish distance preservation on cryptography. The effect on the key is shown in Figure 3.1. The time it takes for a user to generate a key is mainly the time it takes to generate a key. The time when the private key was generated is not taken into account. The generation of a public-private key pair takes approximately eight milliseconds. Find the value of the transformation matrix during initialization keyword processing. Experiments have shown that only 30 fundamental transformations are needed. You can see from Figure 3.2 that encryption takes longer than decryption [18]. This is because generating random numbers that meet the requirements during cryptographic processing takes some time. The average time for each encoding is 0.4 milliseconds. Compared with existing algorithms, the algorithm proposed in this paper has better results. This is mainly because multiple bytes of data can be encrypted or decrypted simultaneously. Figure 3.3 shows that the calculation of 40,000 passwords only takes 67 seconds, and the calculation of 40,000 intermediate points takes only 67 seconds [19]. The amount of encryption required for clustering in a cloud computing environment is acceptable. Since cloud services have higher performance in practice, they can achieve higher privacy mining efficiency in practical applications.

3.2. The performance and accuracy of privacy-based cluster mining algorithms. The research content mainly involves the preprocessing time overhead of some users, part of the time overhead of cloud service providers, and problems compared with existing algorithms [20]. It can be seen from Figure 3.4 that



Fig. 3.2: *Encryption and decryption time.*

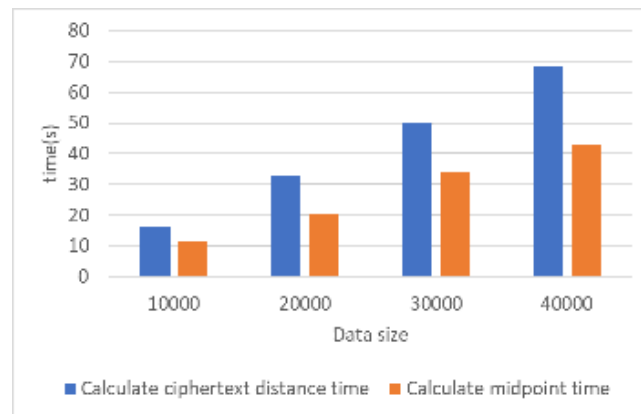


Fig. 3.3: *Computing time cost of ciphertext in the cloud.*

the data published by PPCM is more time-consuming than the existing algorithm, but it is still within the allowed range. The fundamental problem is that it is based on the Blowfish problem and increases the time cost required to keep the data confidential.

Figure 3.5 shows that the time complexity of the ciphertext k-means method is $O(n \log(n))$. The computational time complexity of the ciphertext hierarchical clustering algorithm is $O(n^2)$. The calculation time complexity of ciphertext DBSCAN is $O(n)$. Research has found that in a natural environment, the calculation process of k-mean does not fully comply with the theoretical calculation complexity, and due to differences in sample type and initial value selection, The time consumption of the PPk-mean method may deviate from the analytical results to a certain extent. Experiments show that the privacy-based hierarchical mining method performs weakest, while the one based on PPDBscan is the best. This has a lot to do with the selected materials [21]. The other three methods also have advantages and disadvantages when processing various data types. In many situations, users need to perform multiple mining methods to discover potential clusters in a data set.

As shown in Figure 3.6, this method has the same accuracy as the current most correct RBT method but is safer. In exceptional circumstances, k unknown data sets may lead to inconsistent values of k. This project plans to use three methods to compare the k-means analysis results of the original data, and the PPk-means analysis results under the ciphertext to obtain the accuracy rate.

This algorithm can improve its security, accuracy, and mining efficiency. At the expense of this, the accuracy and security of users' private information are guaranteed. The accuracy of mining is very critical for users. If

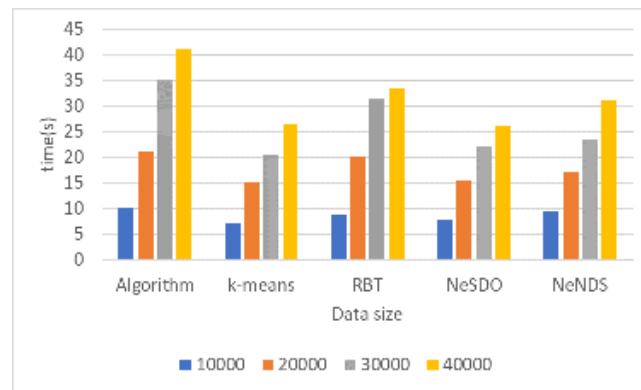


Fig. 3.4: User preprocessing time overhead.

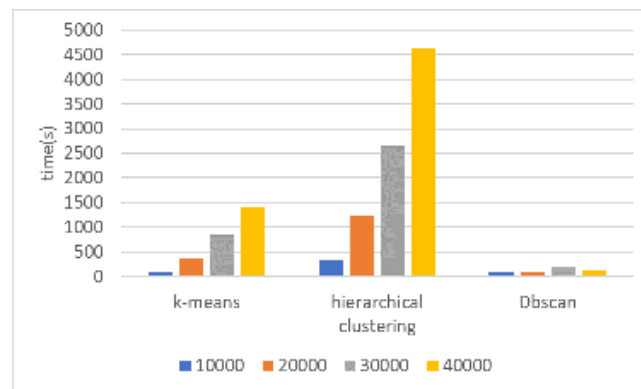


Fig. 3.5: Time cost of privacy-preserving cluster mining.

there is a problem with the accuracy of discovery, it will likely bring irreparable consequences to the user.

4. Conclusion. The confidentiality of user information in ciphertext advertisements based on intelligent algorithms mainly focuses on the confidentiality of deduction data owned by users. This algorithm can keep the deduced data confidential throughout the entire reasoning process. Because encryption can only be performed by the data holder, confidentiality of the data and instant inference results can be considered simultaneously. And it also avoids the cost of repeated encryption and decryption. A model security method based on Blowfish is proposed. It can effectively store and read advertising mode documents while ensuring ciphertext security.

REFERENCES

- [1] Boerman, S. C., Kruijemeier, S., & Zuiderveen Borgesius, F. J. (2021). Exploring motivations for online privacy protection behavior: Insights from panel data. *Communication Research*, 48(7), 953-977.
- [2] Bleier, A., Goldfarb, A., & Tucker, C. (2020). Consumer privacy and the future of data-based innovation and marketing. *International Journal of Research in Marketing*, 37(3), 466-480.
- [3] Shahabi, H. G., & Soni, S. (2023). SECURITY AND PRIVACY CHALLENGES IN VEHICULAR AD-HOC NETWORKS: THREATS, COUNTERMEASURES. *Eigenpub Review of Science and Technology*, 7(1), 22-38.
- [4] Sharma, T., & Bashir, M. (2020). Use of apps in the COVID-19 response and the loss of privacy protection. *Nature Medicine*, 26(8), 1165-1167.
- [5] Quan-Haase, A., & Ho, D. (2020). Online privacy concerns and privacy protection strategies among older adults in East York, Canada. *Journal of the Association for Information Science and Technology*, 71(9), 1089-1102.

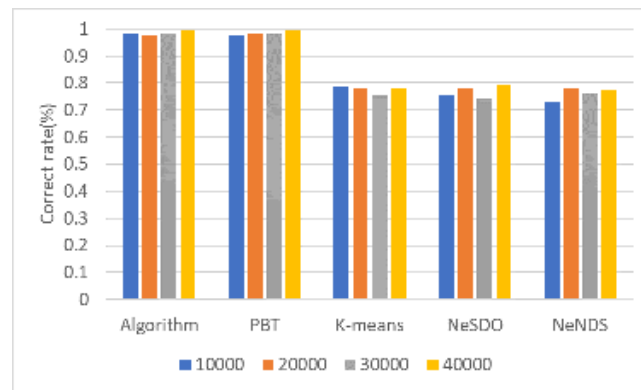


Fig. 3.6: *Privacy-preserving mining algorithm accuracy.*

- [6] Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., & Lin, Z. (2021). When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)*, 54(2), 1-36.
- [7] Jiang, H., Li, J., Zhao, P., Zeng, F., Xiao, Z., & Iyengar, A. (2021). Location privacy-preserving mechanisms in location-based services: A comprehensive survey. *ACM Computing Surveys (CSUR)*, 54(1), 1-36.
- [8] Zhou, X., Liang, W., She, J., Yan, Z., Kevin, I., & Wang, K. (2021). Two-layer federated learning with heterogeneous model aggregation for 6g supported internet of vehicles. *IEEE Transactions on Vehicular Technology*, 70(6), 5308-5317.
- [9] Qu, Y., Pokhrel, S. R., Garg, S., Gao, L., & Xiang, Y. (2020). A blockchained federated learning framework for cognitive computing in industry 4.0 networks. *IEEE Transactions on Industrial Informatics*, 17(4), 2964-2973.
- [10] Rafeian, O., & Yoganarasimhan, H. (2021). Targeting and privacy in mobile advertising. *Marketing Science*, 40(2), 193-218.
- [11] Qiu, H., Qiu, M., Liu, M., & Memmi, G. (2020). Secure health data sharing for medical cyber-physical systems for the healthcare 4.0. *IEEE journal of biomedical and health informatics*, 24(9), 2499-2505.
- [12] Duan, W., Gu, J., Wen, M., Zhang, G., Ji, Y., & Mumtaz, S. (2020). Emerging technologies for 5G-IoV networks: applications, trends and opportunities. *IEEE Network*, 34(5), 283-289.
- [13] Du, M., Chen, Q., Xiao, J., Yang, H., & Ma, X. (2020). Supply chain finance innovation using blockchain. *IEEE Transactions on Engineering Management*, 67(4), 1045-1058.
- [14] Siriwardhana, Y., Gür, G., Ylianttila, M., & Liyanage, M. (2021). The role of 5G for digital healthcare against COVID-19 pandemic: Opportunities and challenges. *Ict Express*, 7(2), 244-252.
- [15] Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6), 305-311.
- [16] Acemoglu, D., Makhdoumi, A., Malekian, A., & Ozdaglar, A. (2022). Too much data: Prices and inefficiencies in data markets. *American Economic Journal: Microeconomics*, 14(4), 218-256.
- [17] Ali, M., Naem, F., Tariq, M., & Kaddoum, G. (2022). Federated learning for privacy preservation in smart healthcare systems: A comprehensive survey. *IEEE journal of biomedical and health informatics*, 27(2), 778-789.
- [18] Qi, L., Hu, C., Zhang, X., Khosravi, M. R., Sharma, S., Pang, S., & Wang, T. (2020). Privacy-aware data fusion and prediction with spatial-temporal context for smart city industrial environment. *IEEE Transactions on Industrial Informatics*, 17(6), 4159-4167.
- [19] Alazab, M., RM, S. P., Parimala, M., Maddikunta, P. K. R., Gadekallu, T. R., & Pham, Q. V. (2021). Federated learning for cybersecurity: Concepts, challenges, and future directions. *IEEE Transactions on Industrial Informatics*, 18(5), 3501-3509.
- [20] Tan, T. M., & Saraniemi, S. (2023). Trust in blockchain-enabled exchanges: Future directions in blockchain marketing. *Journal of the Academy of marketing Science*, 51(4), 914-939.
- [21] Stoilova, M., Livingstone, S., & Nandagiri, R. (2020). Digital by default: Children's capacity to understand and manage online data and privacy. *Media and Communication*, 8(4), 197-207.

Edited by: Zhigao Zheng

Special issue on: Graph Powered Big Aerospace Data Processing

Received: Dec 12, 2023

Accepted: Dec 29, 2023



DESIGN OF 0-DAY VULNERABILITY MONITORING AND DEFENSE ARCHITECTURE BASED ON ARTIFICIAL INTELLIGENCE TECHNOLOGY

JIAN HU*, ZHENHONG ZHANG[†], FEILU HANG[‡] AND LINJIANG XIE[§]

Abstract. In response to the difficulty in detecting attacks caused by the unknown nature of 0-day vulnerabilities, the author proposes a knowledge graph based 0-day attack path prediction method. By extracting concepts and entities related to attacks from existing research on the ontology of network security and network security databases, a network defense knowledge graph is constructed to extract discrete security data such as threats, vulnerabilities, and assets into interrelated security knowledge. Using a knowledge graph reasoning method based on path sorting algorithm to explore possible 0-day attacks in the target system. Experimental results have shown that the proposed method can rely on the knowledge system provided by the knowledge graph to provide comprehensive knowledge support for attack prediction, reduce the dependence of prediction analysis on expert models, and effectively overcome the adverse effects of unknown 0-day vulnerabilities on prediction analysis. It improves the accuracy of 0-day attack prediction and utilizes the path sorting algorithm to infer based on the explicit feature of graph structure, being able to effectively backtrack the reasons behind the formation of reasoning results, this to some extent improves the interpretability of attack prediction analysis results.

Key words: 0day attack, Attack path prediction, Artificial intelligence, Defense architecture

1. Introduction. With the continuous development of computer and communication technologies, network security has increasingly become an important issue affecting network performance and data security. The proliferation of network attacks and viruses poses a serious threat to network and application systems. For enterprise level users, whenever they encounter these threats, they often cause data damage, system abnormalities, network paralysis, information theft, decreased work efficiency, and significant direct or indirect economic losses. Since 2006, the US Security Training and Research Institute (SANS) has included zero day attacks as one of the top 20 global internet security threats annually [1]. Enterprises that have not yet patched zero day vulnerabilities have become the preferred targets for hackers, and such attacks are developing rapidly. Zero day attack refers to an attack launched by malicious software that exploits certain vulnerabilities in the operating system or application software that are not known to developers or have not been patched in a timely manner. Those vulnerabilities that are not known to developers or have not been patched in a timely manner are also known as "zero day vulnerabilities". As a type of attack, the difference between zero day attacks and traditional hacker attacks is that the target of zero day attacks is some potential unknown or publicly disclosed but not patched vulnerabilities. According to authoritative institutions, there may be 4-5 coding vulnerabilities in every 1000 lines of code in operating systems and applications currently in use[2]. With the continuous emergence of various computer vulnerabilities, the situation of zero day attacks is also constantly changing: From a single point to a desktop type, then gradually transitioning to a network type, and even currently there is a trend towards a full network type development. In this complex form, defense against zero day attacks is also constantly evolving. Traditionally, regular updates of system patches, firewalls, intrusion detection systems, and antivirus software are commonly used to protect critical business and IT infrastructure. These systems provide good first level protection, but still cannot avoid zero day vulnerability attacks. Faced with the increasing

*Network Security Management Center of Information Center of Yunnan Power Grid Co., LTD., Kunming, Yunnan, China, 650000 (Corresponding author, hjiang2023@126.com)

[†]Network Security Management Center of Information Center of Yunnan Power Grid Co., LTD., Kunming, Yunnan, China, 650000

[‡]Information Security Operation and Maintenance Center of Information Center of Yunnan Power Grid Co., LTD., Kunming, Yunnan, China, 650000

[§]Information Security Operation and Maintenance Center of Information Center of Yunnan Power Grid Co., LTD., Kunming, Yunnan, China, 650000

zero day threat, both system, network, and security vendors are loudly calling for the importance of real-time updates [3,4]. For manufacturers, the purpose of instant updates is to respond to the increasingly short attack time gap of hackers. However, users often lack manpower, resources, and time to effectively perform the work of instant updates and repairs.

The main reason why many enterprises are unable to immediately complete the repair work of system or software and hardware device vulnerabilities is that they do not have time to discover the vulnerabilities, do not further evaluate and diagnose the vulnerabilities, and are unable to patch and update all computers or endpoint devices. Another more important reason is the inability to conduct compatibility testing on patches, which often leads to system instability and even crashes.

Attack prediction technology is the key to research on 0day attack detection. However, research on 0-day attack prediction generally relies on hypothetical conditions, attack models constructed by expert knowledge, and the pre - and post attack dependencies in the same attack path to address the impact of unknown 0-day vulnerabilities, there are three shortcomings in this process: Firstly, the conditional assumption lacks effective constraints, which can easily lead to a large scale of prediction results for 0-day attacks, reducing the significance of prediction; The second is the attack model constructed by expert knowledge, which is easily constrained by the subjective knowledge of experts; Thirdly, the prediction method is difficult to apply when the known attack path is incomplete. In response to the above shortcomings, the author proposes a 0-day attack path prediction method based on a knowledge graph [5]. By using a network defense knowledge graph, attack related threats, assets, vulnerabilities, and other data are fused into a security knowledge base that is interrelated and covers a wide range of knowledge. Based on the integrated vulnerability data, attack intent, and other knowledge, reasonable constraints are imposed on the assumptions of unknown attributes of 0-day vulnerabilities; Secondly, using path sorting algorithms, the relationship path between the attacker entity and the target entity in the knowledge graph is used as a feature to predict the 0-day attack from a more comprehensive perspective, overcoming the limitations of expert knowledge construction models [6]; Finally, using historical attack data as samples, a logistic binary classifier is designed and trained to implement single step attack prediction, thereby breaking away from dependence on known attack paths. By reusing the single step attack probability output by the logistic binary classifier, the comprehensive utilization rate of the attack path is calculated to predict the 0-day attack path most likely to be exploited by the attacker against the target asset, thereby supporting defense decisions [7,8].

2. Methods.

2.1. Preparatory knowledge. In order to make the expression clear and accurate, the relevant concepts in the text are defined as follows.

Definition of Network Defense Knowledge Graph (CKG). CKG is represented by triplets (CSO, FACT, T), where CSO=(C, R, P) is the network security ontology, C is the class set, R is the relationship type set, P is the attribute type set, FACT is the set of data knowledge represented in RDF (resourcdescriptionframework) triplet format, T is the set of type dependency relationships between classes in CSO and entity objects in FACT [9].

Define a 20day vulnerability. 0day vulnerability refers to a general term for system vulnerabilities that have not been discovered by security vendors but may be mastered by hacker organizations. In order to make the research more targeted, the following 0day vulnerabilities only refer to technical vulnerabilities that have not been discovered by security vendors.

Define a 30 day attack. A 0-day attack refers to a single step attack initiated by an attacker using a known 0-day vulnerability, denoted as a^0 . Relatively, known attack a^k is a single step attack initiated by an attacker exploiting a known vulnerability.

Define 40 day attack path zap. Zap refers to an acyclic attack sequence consisting of a set of single step attacks with 0 day dependencies, represented by (A, E), where A is the set of single step attacks and E is the directed edge set of linked single step attacks [10].

Define 50 day attack graph ZAG. ZAG refers to an attack graph containing 0-day attacks, represented as (A, Priv, L, Prob), among them, $A = \{a^0\} \cup \{a^k\}$ is a single step attack set consisting of 0 day attacks and known attacks. Single step attack a is represented as a binary (host, Vul), where the host is the target device, Vul is the vulnerability exploited, and Priv is the pre - and post permission set for single step attacks,

Table 2.1: Main Symbols and Their Description

symbol	describe	symbol	describe
CKG	Network Defense Knowledge Graph	vul	Vulnerabilities exploited by single step attacks
CSO	Network Security Ontology	rp	Relationship Path
zap	0day attack path	c	Classes in ontology
ZAG	0day attack graph	r	Relationship types in ontology
A	Single Step Attack Set	Domain(r)	Definition domain of relationship types
a	Single step attack	Range(r)	The range of values for relationship types
Priv	Node permission set	s	Source entity
L	Directed Link Set between Single Step Attack and Permissions	d	Target entity
E	Directed Edge Set Between Single Step Attacks in Attack Paths	h	Relationship Path Eigenvalues
Prob	Probability set of single step attack occurrence	H	Relationship path feature vector
host	Devices that have been attacked	θ	Relationship path weight vector

$L=\{AxPriv\}U\{PrivxA\}$ is the link between a single step attack and permissions, representing the pre - and post relationship between them. Prob is the set of probabilities of a single step attack occurring.

Define 6 relationship paths (rp, relationpath). rp refers to a sequence composed of a set of relationship types in a knowledge graph, written as $rp : c_0 \xrightarrow{r_1} c_1 \xrightarrow{r_2} \dots \xrightarrow{r_l} c_l, r \in R$. Among them, $c_l \equiv Range(rp), l = |rp|$ represents the length of the relationship path, which is the total number of relationship types included in the relationship path, $0 \leq i \leq l$.

The relationship between knowledge graph and attack graph is as follows: CKG is the input of attack prediction algorithm, serving as a knowledge base to provide the necessary knowledge for attack prediction; 0day attack graph ZAG is a graphical representation of attack prediction results. The difference between relational paths and attack paths is as follows: rp is used as a feature in the logistic regression model in path sorting algorithms to perform attack prediction; The 0-day attack path zap is a prediction result of the extracted attack path based on the 0-day attack graph, combined with the probability of multi-step attacks occurring. The main symbols that appear are described in Table 2.1 [11,12].

2.2. Network Defense Knowledge Graph. Knowledge graph is an effective technical method that uses graph models to describe knowledge and model the relationships between things. Applying this technology to the field of network security and constructing a network defense knowledge graph can integrate heterogeneous and fragmented network security data into a unified and interrelated security knowledge format, providing support for attack prediction and testing the required knowledge. The 0-day attack graph ZAG is a graphical representation of the attack prediction results. The difference between relational paths and attack paths is as follows: RP is used as a logistic regression model in path sorting algorithms to perform attack prediction, which is beneficial for implementing targeted defense. The following is a detailed introduction to the construction of the network defense knowledge graph architecture and the design of the network security ontology.

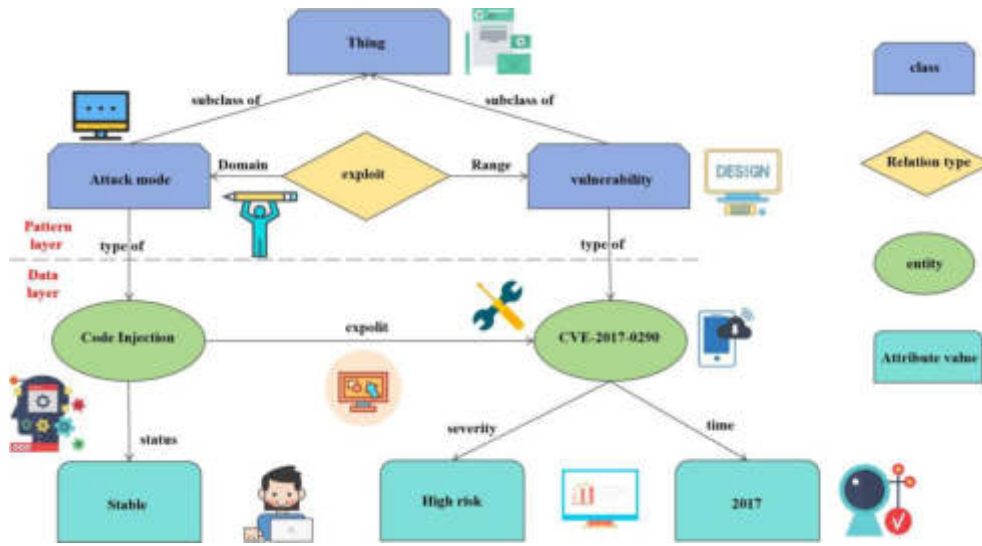


Fig. 2.1: Example of the Relationship between Pattern Layer and Data Layer

(1) *Architecture construction.* According to Definition 1, the network defense knowledge graph can be divided into two parts: The pattern layer and the data layer. Among them, the pattern layer is the core of the knowledge graph, and the basic concept system of network defense is defined by the Cybersecurity ontology (CSO), providing pattern definitions for modeling data layer knowledge[13]. The data layer is the main body of the knowledge graph, which is a collection of data knowledge obtained through knowledge extraction, knowledge fusion, and other steps, modeled under the pattern definition. Data knowledge is represented in the form of RDF triplets as (subject, predicate, object). The example of a two-layer relationship is shown in Figure 2.1. The pattern layer of this example defines attack pattern classes, vulnerability classes, and the relationship type exploit with the two as defined cities and value ranges, respectively. The data layer models data knowledge based on this pattern (CVE-2017-0290, CodeInjection, exploit), indicating that code injection can exploit vulnerability CVE-2017-0290.

According to the hierarchical structure of the knowledge graph, there are mainly two methods for its construction: top-down and bottom-up. Due to the mature research on the conceptual system in the field of network security, the network defense knowledge graph is suitable for a top-down knowledge graph construction method, which first constructs a pattern layer based on the network security ontology, and then integrates multiple knowledge extraction and fusion technologies based on the pattern layer to extract and model data knowledge from heterogeneous data sources and construct a data layer [14,15].

(2) *Ontology design.* From the top-down sequence of knowledge graph construction, it can be seen that CSO is the key to determining the quality of network defense knowledge graph. The author uses ontology integration to construct CSO, integrating existing mature network security ontologies into a unified ontology, drawing on current research achievements in this field, and achieving complementary advantages among different achievements[16]. Due to NSSEKB_0 (ontology of network security situation element knowledge base) systematically sorts out the network security knowledge system from three levels: domain ontology, application ontology, and atomic ontology, and has good knowledge completeness. AFACSDO (asset infrastructure security domain ontology) reuses multiple representative network security ontologies, which is the latest achievement in network security ontology research and has good timeliness. Therefore, the author selected the above two research results as integration objects to implement ontology integration.

2.3. 0-day attack path prediction. The network defense knowledge graph defines the concept of attacks as a type of relationship with attacker classes as the domain and device classes as the value domain. The inference problem of specific attack behaviors is transformed into the prediction problem of attack relationships

between attacker entities and device entities in the data layer of the knowledge graph. The path ranking algorithm (PRA) is an effective method for predicting knowledge graph links. Its prediction results not only have high accuracy, but also strong interpretability, making it easy for defenders to mine knowledge about attack causes and other factors after obtaining the prediction results, therefore, the author chooses the PRA algorithm to perform one-step attack prediction and constructs a 0-day attack graph. On this basis, by analyzing and comparing the comprehensive utilization rates of different attack paths, predict the most likely 0-day attack path that attackers are likely to utilize [17].

(1) Analysis of Unknown Properties of 0day Vulnerability. Before implementing the prediction, the basis of prediction analysis is to make assumptions and implement constraints on the missing 0day vulnerability attributes through unknown attribute analysis. The unknown attributes of the 0day vulnerability mainly include the location of the vulnerability, exploitation conditions, and impact information. Current research mainly supplements unknown 0-day vulnerability information through conditional assumptions. In this process, how to reasonably constrain the assumed vulnerability information is the key to determining the quality of prediction results. Based on the knowledge graph integration, the author proposes hypotheses about the existence, availability, and harm of 0-day vulnerabilities, and uses statistical analysis, sample training, and intention analysis to constrain the relevant hypotheses.

Existence assumption: The 0day vulnerability may exist in any component of the device. The vulnerability data provided by databases such as NVD and CNNVD show significant differences in the number of vulnerabilities exposed by different components, indicating that the existence of vulnerabilities is related to the components that serve as their carriers. Therefore, for this assumption, component features can be used to constrain and consider the 0-day vulnerability as a potential vulnerability that the component may expose in the future. By statistically analyzing the vulnerability exposure history data of different components, the possibility of 0-day vulnerabilities in different components can be quantified[18].

Availability Assumption: The 0day vulnerability may be triggered by arbitrary permissions on the target device. Permissions exist in the form of permission entities in the knowledge graph, and triggering different vulnerabilities requires varying degrees of permission. The higher the permission, the easier it is for attackers to trigger vulnerabilities. Set the attribute "tri_prob" of the relationship "trigger" in the knowledge graph to indicate the probability of permission triggering a 0-day vulnerability, and assign values based on the level of permission. The relationship "trigger" is contained in the relationship path between the attacker entity and the target device entity, so "triprob" can be reflected in the sample features by participating in the calculation of path features. In the process of training a Logistic binary classifier using attack samples, the exploitation conditions of the 0-day vulnerability can be constrained to the scenarios in the samples based on the difference between positive and negative samples. When the classifier determines that the 0-day attack relationship is valid, the permissions used to trigger the 0-day vulnerability can be determined by querying the permission entities that the attacker entity has obtained on the device entity in the knowledge graph.

Harmful assumption: The harm generated by exploiting a 0-day vulnerability can only meet the minimum requirement for attackers to achieve their attack intent on the target device.

The harm caused by vulnerability exploitation is represented in the form of entities in the knowledge graph, and at the same time, the knowledge graph integrates the attacker's attack intent by threatening entities. Due to the fact that attackers exploit vulnerabilities to launch attacks with the aim of achieving the attack intent, if the consequences of exploiting vulnerabilities are not sufficient to support the implementation of the attack intent, then the attacker is not necessary to exploit the vulnerability. When the consequences exceed the need to achieve the attack intent, the excess is not significant to the attacker, therefore, the above assumptions about using attack intent to constrain the harm caused by exploiting 0day vulnerabilities are feasible and reasonable. For example, the attacker entity APT-28 mainly engages in theft activities, and the target device entity stores confidential information. A threat entity "stealing secrets" is constructed in the knowledge graph as APT-28's attack intention against the device. If it is predicted that APT-28 has launched an attack on the device using the 0day vulnerability, the resulting consequences are constrained by the threat entity as "confidentiality damage", without further impact on integrity, availability, and other aspects.

(2) Attack path prediction. Using the historical attack data of a given system to construct attack samples, a classifier LC is trained to predict whether an attack has occurred. Based on this, 0 day attacks and known

attacks are distinguished from the positive attack samples, and 0 day attack samples are constructed. The classifier LCz is trained to determine whether a single step attack occurred as an 0 day attack. After completing the attack prediction, utilize the query function of the graph database to mine the vulnerability and pre - and post conditions of attack exploitation based on the starting and ending entities and relationship paths, construct a single step attack, and generate a 0-day attack graph. Based on the 0-day attack graph, with the attacker's initial permissions as the starting point and the target permissions as the endpoint, extract the 0-day attack path, and calculate the comprehensive utilization rate of different attack paths by reusing the probability of a single step attack output by the LCA classifier. Based on this, predict the most likely 0-day attack path that the attacker is likely to use. The calculation of comprehensive utilization rate is shown in Equation 2.1.

$$Exploit(zap) = \prod_{a \in zap} prob(a) \quad (2.1)$$

3. Results and Analysis. In order to verify the effectiveness of the method, the experimental environment consists of three subnets, with firewalls deployed between the subnets to achieve access control. Among them, the web server and email server deployed in the DMZ region respectively provide external application service interfaces and internal email services: subnet 1 is the office area, where two hosts and one file server are deployed, and the file server stores enterprise confidential files; subnet 2 is the business area, where application servers are deployed to provide application business support for the web server. The distribution of vulnerabilities in the system is divided into 10 training scenarios and 1 experimental scenario, respectively, for training classifiers and implementing predictions. Among them, the classifier was trained using 9 sets of attack samples generated from training scenarios, and the remaining 1 set was used to generate a test set. The performance parameters of the classifier were tested. The experimental scenario was mainly used to generate 0-day attack maps and predict 0-day attack paths. Compared with existing research results, the advantages of this method were tested.

3.1. Sample Training. Based on the attack data of CTF teams simulating attackers on target systems in different training scenarios, relying on knowledge graphs, the successfully attacked devices are used as target entities to calculate path characteristics and construct attack positive samples $\{(H_j, y = 0)\}$. The failed and unselected devices are used as target entities to construct attack negative samples $\{(H_j, y = 0)\}$. Using the model `sklearn.linear_model` in the Python 3.5 environment `_LogisticRegression` constructs a binary classifier and trains LCA using attack samples generated from training scenarios 1-9. On this basis, in the attack positive samples, distinguish between 0 day attacks and known attacks, construct 0 day attack positive samples and negative samples respectively, and train LCz. Use 5-fold cross validation to obtain the learning curve of the classifier, as shown in Figures 3.1 and 3.2[19].

Using the samples generated from training scenario 10 as the test set, the classifier was tested and the predicted results were compared with the actual attack situation. The accuracy of the classifier's LCA was 0.875, the recall was 0.917, and the harmonic mean F1 was 0.883. The classifier LCz's prediction results for the test set are consistent with the actual situation. It can be seen that the classifier has good recognition ability against 0-day attacks.

3.2. Experimental Results. Use the trained classifiers LCA and LCz to predict possible attacks in the target system in the experimental scenario. The predicted attack process is as follows: the attacker first utilizes remote access privileges to access `firewalls_1`. Initiate a 0-day attack, break through access control, and obtain remote access to the `E-MailServer`. Utilize the CVE-2018-18772 vulnerability in its operating system CentOS to launch a cross site request attack, obtain access to `Host_1` and `firewall2`, and launch code injection attacks using the existing CVE-2018-12714 and CVE-2017-17156 vulnerabilities in both, through `Host_1`. Obtain access to `File_server` and access to `Host2` through `firewall2`. At this time, the access permission can be directly used to launch a 0-day attack on `File_server` or `Host2`, obtaining user permissions for `File_Server` and triggering known vulnerabilities such as CVE-2018-8169. The probability of a single step attack occurring in LCA output is shown in Table 3.2.

With the ultimate goal of obtaining `root_File_Server`, the 0-day attack path is extracted, including `zap1: a1 → a2 → a3 → a4` and `zap2: a1 → a2 → a5 → a6 → a7`. The comprehensive utilization rates of the two paths are calculated using Equation 2.1 to be 0.18 and 0.21, respectively. From this, it can be seen that

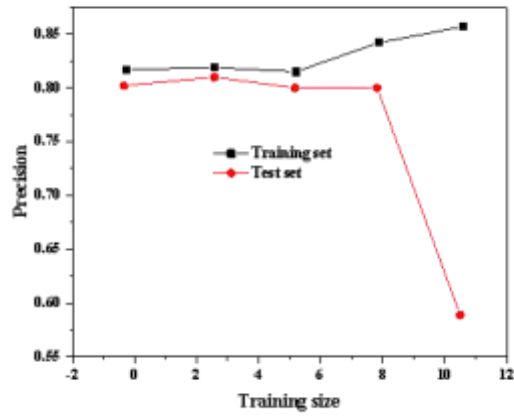


Fig. 3.1: LCA Learning Curve

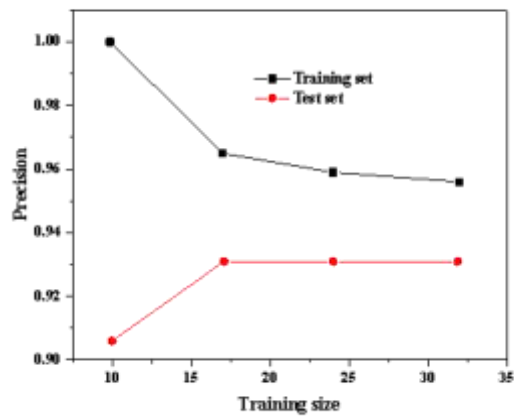


Fig. 3.2: LCZ Learning Curve

Table 3.1: Probability of attack occurrence

code	Single step attack	Probability of attack occurrence
a1	(Firewall_1 0day-001)	0.51654178
a2	(Email_Server CVE-2018-18772)	0.77063787
a3	(Host_1 CVE-2018-12714)	0.78201284
a4	(File_Server 0day-003)	0.58126118
a5	(Firewall_2 CVE-2017-17156)	0.76765849
a6	(Host_2 0day-005)	0.77269452
a7	(File_Server CVE-2018-8169)	0.9277652

Table 3.2: Comparative Analysis

method	accuracy	Convenience	Applicability	Comprehensive knowledge	Interpretability
Method A	Lower	Lower	higher	Lower	Not supported
Method B	higher	Lower	Lower	Lower	Not supported
Author's method	higher	higher	higher	higher	support

although the attack path zap2 involves 5 attacks, more than zap1's 4 attacks, its comprehensive utilization rate is higher and more likely to be exploited by the attacker.

3.3. Experimental analysis. Based on the experimental results, by comparing the author with the other two A and B methods, analyze the advantages of this method in terms of accuracy, convenience, applicability, comprehensiveness of relevant knowledge, and interpretability of prediction results, in order to verify its effectiveness. The specific comparative analysis is shown in Table 3.2.

In terms of accuracy, A's simple prediction of attack path zap1 based on the number of 0-day attacks is more likely to become the attack path chosen by the attacker. The author, by constraining the assumption of the existence of 0-day vulnerabilities (the most likely component in Web_Server to have 0-day vulnerabilities is the Apache webserver component, with a probability of 0.05), and using the trained classifier LCA to calculate the probability of an attack on Web_Server (with a probability of 0.3), determined that the attack relationship did not hold, and reasonably excluded the 0-day attack here, reducing the size of the 0-day attack prediction results, at the same time, based on the comprehensive utilization rate, Zap2 can more reasonably predict the attack path chosen by the attacker, improving the accuracy of prediction.

In terms of convenience, A and B not only need to collect relevant knowledge before implementing prediction, but also need to construct specialized 0-day attack rules as the basis for prediction analysis, which causes certain expenses. However, the author relies on network defense knowledge graph and knowledge graph inference methods to implement prediction, without the need for specialized 0-day attack rules, saving expenses and improving the convenience of the method.

In terms of applicability, method B relies on a relatively complete known attack path to implement attack inference. However, in this experimental environment, Web_server does not have a known vulnerability, and firewall_1's known vulnerability, CVE-2019-1934, requires user level permission to trigger. Therefore, it is unable to generate a known attack path under initial permission conditions, and method B is not suitable for such scenarios.

In terms of knowledge comprehensiveness, the author builds a network defense knowledge graph based on the mature conceptual knowledge in the field of network security, from three aspects: threat, asset, and vulnerability, and extracts relationship paths as features to apply to attack prediction. The prediction process not only uses vulnerability knowledge such as the existence, availability, and impact of vulnerabilities involved in A and B, but also combines knowledge of attack intent and asset types, making prediction analysis more comprehensive and improving the rationality of prediction results[20].

4. Conclusion. The author proposes a 0-day attack path prediction method based on network defense knowledge graph to address the difficulty of detecting 0-day attacks caused by the unknown nature of 0-day vulnerabilities, as well as the shortcomings of existing research in using conditional assumptions and correlation before and after attacks to overcome the impact of unknowns. By utilizing the mature knowledge of network security ontology in current research, a comprehensive network defense knowledge graph has been constructed, integrating discrete threat, vulnerability, and asset knowledge into a highly correlated knowledge system, providing comprehensive knowledge support for attack prediction. On this basis, the attack prediction problem is transformed into a link prediction problem. A path sorting algorithm with high prediction accuracy and strong interpretability of prediction results is selected to extract the relationship paths between the attacker entity and the target device entity as features, and more comprehensively predict the attack. This effectively over-

comes the influence of unknown 0-day vulnerabilities and one-sided expert knowledge, and improves prediction accuracy, and provided support for the interpretability of the predicted results. The next step is to expand the knowledge module, improve the accuracy of attack prediction, and explore the traceability problem of network attackers based on knowledge graphs in the presence of multiple attackers simultaneously.

5. Acknowledgement. Yunnan Power Grid CO., LTD. Science and Technology Project "Web Application Protection Based on RBI Remote Browser Isolation Technology" (NO.:059300KK52220011)

REFERENCES

- [1] Khaoula, T., Abdelouahid, R. A., Ezzahoui, I., & Marzak, A. (2021). Architecture design of monitoring and controlling of iot-based aquaponics system powered by solar energy - sciencedirect. *Procedia Computer Science*, 191(33), 493-498.
- [2] FAN Xue-wei, XIE Feng, WANG Xiao-wu, TANG Nan. (2023). Design of remote monitoring system for electric torque wrench based on b/s and c/s fusion architecture. *Manufacturing Automation*, 45(2), 175-178.
- [3] Yusuf, W. (2022). The design of integrated fire spot monitoring system for industrial plantation forest using enterprise architecture approach. *EDPACS: The EDP audit, control and security newsletter*44(2), 66.
- [4] Qian, H. (2021). Design of tunnel automatic monitoring system based on bim and iot. *Journal of Physics Conference Series*, 1982(1), 012073.
- [5] Sangeetha, M., Thejaswini, G., Shoba, A., Gaikwad, S. S., Amretasre, R. T., & Nivedita, S. (2021). Design and development of a crop quality monitoring and classification system using iot and blockchain. *Journal of Physics: Conference Series*, 1964(6), 062011 (14pp).
- [6] Katpatal, Y. B., & Singh, C. K. (2023). Conjunctive use of flow modelling, entropy, and gis to design the groundwater monitoring network in the complex aquifer system. *International Journal of Hydrology Science and Technology*, 15(1), 78-.
- [7] Zhang, H., Ge, D., Yang, N., Jia, P., & Yang, Y. (2021). Study on internet of things architecture of substation online monitoring equipment. *MATEC Web of Conferences*, 336(5), 05024.
- [8] Sun, C., Hao, H. U., Yang, Y., & Zhang, H. (2022). Prediction method of 0day attack path based on cyber defense knowledge graph. *Chinese Journal of Network and Information Security*, 8(1), 151-166.
- [9] Lindberg, L., Vinnars, B., & Lalander, C. (2022). Process efficiency in relation to enzyme pre-treatment duration in black soldier fly larvae composting. *Waste Management*, 137(45), 121-127.
- [10] Chen, D., Yan, Q., Wu, C., & Zhao, J. (2021). Sql injection attack detection and prevention techniques using deep learning. *Journal of Physics: Conference Series*, 1757(1), 012055 (7pp).
- [11] (2021). Elective cesarean delivery at term and its effects on respiratory distress at birth in japan: the japan environment and children's study. *Health Science Reports*, 4(4),46.
- [12] Chen, J., Kong, Q., Sun, Z., & Liu, J. (2021). Freshness analysis based on lipidomics for farmed atlantic salmon (*salmo salar* l.) stored at different times. *Food chemistry*, 67(7),131564.
- [13] Luo, S., Wang, Z., Li, X., Onchari, M. M., & Jin, S. (2021). Feed deprivation over 16 days followed by refeeding until 75 days fails to elicit full compensation of *procambarus clarkii*. *Aquaculture*, 547(34), 737490.
- [14] Meng, B., Smith, W., & Durling, M. (2021). Security threat modeling and automated analysis for system design. *SAE International Journal of Transportation Cybersecurity and Privacy*765(1), 4.
- [15] Chondamrongkul, N., Sun, J., & Warren, I. (2021). Formal security analysis for software architecture design: an expressive framework to emerging architectural styles. *Science of Computer Programming*, 206(27), 102631.
- [16] Petrillo, A., Murino, T., Piccirillo, G., Santini, S., & Caiazzo, B. (2023). An iot-based and cloud-assisted ai-driven monitoring platform for smart manufacturing: design architecture and experimental validation. *Journal of Manufacturing Technology Management*, 34(4), 507-534.
- [17] Behmel, S., Damour, M., Ludwig, R., Rodriguez, M. J. (2021). Intelligent decision-support system to plan, manage and optimize water quality monitoring programs: design of a conceptual framework. *Journal of Environmental Planning and Management*, 64(3a4),43.
- [18] Dutta, A., & Kumar, A. (2022). The imperative relationship between architecture, urban design and development and disaster management. *ECS transactions*35(1), 107.
- [19] Bortsova, G., Cristina González-Gonzalo, Wetstein, S. C., Dubost, F., & Bruijne, M. D. (2021). Adversarial attack vulnerability of medical image analysis systems: unexplored factors. *Medical Image Analysis*, 73(1), 102141.
- [20] Yi, J., & Bo, W. (2021). Architecture design of an intelligent monitoring system for turbine filtration device. *Journal of Physics: Conference Series*, 1944(1), 012040 (6pp).

Edited by: Zhigao Zheng

Special issue on: Graph Powered Big Aerospace Data Processing

Received: Dec 12, 2023

Accepted: Dec 29, 2023



A HYBRID IMAGE FUSION AND DENOISING ALGORITHM BASED ON MULTI-SCALE TRANSFORMATION AND SIGNAL SPARSE REPRESENTATION

DAJUN SHENG*

Abstract. In response to the problem of denoising in image fusion, the author proposes a hybrid image fusion and denoising algorithm based on multi-scale transformation (MLT) and signal sparse representation (SRS). A hybrid model is constructed for shear transformation, and the coefficients after MLT decomposition are thresholded. Sliding window technology and translation invariance are used to form sparse representation for image fusion, and SRS algorithm is used to remove noise from the source image. The experimental results show that the algorithm reduces the contrast and spectral information distortion of the fused image, displays high-quality visual fusion effects, maintains high PSNR values under different noise levels, can provide a more complete description of the features in the image, accurately judge the focus area, maintain the structural correlation of the image, and strengthen the description of fusion edges and details in the fused image. It has been proven that the methods of multi-scale transformation and sparse signal representation can fuse and denoise images.

Key words: Multiscale transformation, Signal sparsity, Image fusion, Denoising algorithm

1. Introduction. In the real world, 20% of human perception information comes from hearing, about 70% of information comes from vision, and about 10% of information comes from taste, smell, touch, and other pathways. From the perspective of biological visual information perception, visual information such as images and videos has become the most important means for humans to perceive and recognize information [1]. In 2015, data showed that the total number of uploaded photos on social networking site Facebook reached 600 billion, with a growth rate of 500 million photos uploaded daily; The average daily video views on the video sharing website YouTube are as high as 8 billion times. With the continuous development of science and technology, the demand for visual information by humans is increasing day by day, and the amount of visual information data is rapidly increasing. The acquisition of multi-source heterogeneous visual information has brought unprecedented development opportunities to visual information processing technology [2]. However, in the process of visual information perception, many factors such as data acquisition, compression, transmission, and storage, as well as hardware device limitations and human operation errors, result in image quality problems such as data loss, noise introduction, and motion blur, which also bring huge challenges to theoretical research and engineering practice.

Natural image quality plays a crucial role in communication and visual perception. High quality images have richer content and information, providing users with a better interactive experience; Poor quality images can lose important information and even cause discomfort to users. Although improving the performance of imaging hardware can improve image quality to a certain extent, equipment costs will significantly increase. The blurring effect caused by the shaking of the shooting equipment, as well as the Gaussian, pulse, and quantization noise introduced during the shooting, storage, and compression processes, cannot be avoided by improving hardware facilities due to the degradation and distortion effects on image quality caused by the computational processing process itself or network packet loss and noise interference [3]. Therefore, utilizing computer theoretical technologies such as image processing, machine vision, and numerical analysis to analyze and process multimodal or noisy images, in order to better understand and perceive target objects, has significant theoretical and practical significance. Image denoising and fusion technology has emerged with the aim of removing or weakening image quality issues during the process of acquiring, transmitting, or storing images. Compared with hardware methods, image denoising and fusion technology has obvious characteristics such as

*College of Big Data and Artificial Intelligence, Xinyang University, Xinyang, Henan 464000, China (Corresponding author, x0376y@163.com)

low cost, high flexibility, and wide applicability. However, it involves the understanding, representation, and modeling of image degradation, noise characteristics, and its own characteristics, and there are many difficult problems and challenges. Image denoising and fusion technology is essentially a fundamental research in the fields of image processing and computer vision, and has received widespread attention [4]. The research in the field of image fusion and denoising began in the 1950s and 1960s. In the practical process, the multi-source images obtained through a large amount of manpower and financial resources often have a certain degree of blurriness. Due to the limited technical conditions at the time, blurred images did not have practical value. Fortunately, through the unremitting efforts of experts and scholars to accurately reconstruct real original images, there is currently a relatively mature and widely used image denoising and fusion technology. A typical successful example is in 1964, NASA's Jet Propulsion Laboratory captured images of the moon on a spacecraft using television cameras, which contained information on noise and interference [5]. Computer processing was used to remove interference and noise, correct geometric distortion and contrast loss, and greatly improve image quality. In recent years, image denoising and fusion techniques based on sparse representation theory have effectively represented and approximated the original image by constructing a dictionary using linear combinations of a few atoms, mining the relationship between representation coefficients and corresponding atoms to reveal the inherent nature of visual information, and obtaining images that conform to human visual perception characteristics. At the same time, due to its superior performance such as simple model, easy implementation, noise resistance, interpretability, and ability to process high-dimensional data, it has sparked a wave of research on image denoising and fusion technology under sparse representation frameworks in the academic and engineering fields. With the continuous development of information technology such as images and videos, image denoising and fusion technology has been widely applied in many scientific and technological fields such as military remote sensing, security monitoring, medical imaging, and consumer electronics.

It can be foreseen that with the comprehensive arrival of social media, the Internet, and the era of big data, the vigorous development of information technology mainly based on images and videos will inevitably give rise to a large number of emerging applications and new demands for image denoising and fusion technology. The large and widespread application demands in the industrial sector will also drive the continuous development of image fusion and denoising research fields. As mentioned earlier, research on image denoising and fusion is also of great significance for the development of theoretical technologies in fields such as image processing and computer vision. On the one hand, image denoising and fusion technology can serve as the underlying technical support for other high-level image processing techniques, thereby improving the efficiency, accuracy, and stability of subsequent image processing tasks; On the other hand, research on image self problems not only involves modeling, representing, and understanding the attributes of images themselves, but also involves interdisciplinary research on human visual mechanisms and psychological perception. It is a fundamental research in the fields of image processing and machine vision, and has great research value for disciplines such as machine vision, pattern recognition, and image understanding[6].

Therefore, in order to solve the noise problem in image fusion, the author proposes a hybrid image fusion and denoising algorithm based on multi-scale transformation (MLT) and signal sparse representation (SRS). The algorithm process is as follows:

- Step 1: Perform shear transformation under the mixed model, thresholding the values of various coefficients after MLT decomposition;
- Step 2: Utilizing sliding window technology and translation invariance to form sparse representations for image fusion;
- Step 3: SRS global processing image denoising algorithm removes noise from the source image.

The experimental results show that the proposed algorithm reduces the contrast and spectral information distortion of the fused image, and has good image fusion and denoising effects.

2. A hybrid image fusion and denoising algorithm based on MLT and SRS. As shown in Figure 2.1, the fusion algorithm proposed by the author consists of the following three steps:

- a) Use cartoon texture decomposition to decompose the original image into cartoon and texture parts. The cartoon part mainly includes the structural and geometric parts of the image, while the texture part mainly includes the oscillation and noise parts of the image.
- b) The cartoon and texture parts of the image are fused separately. The cartoon part is fused using convolu-

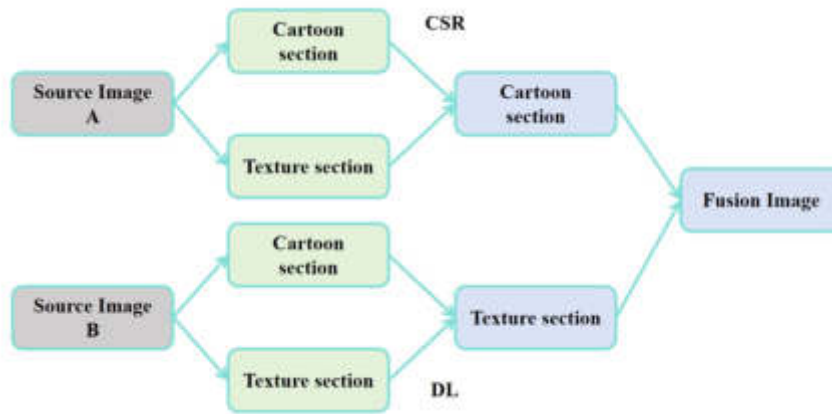


Fig. 2.1: Fusion algorithm framework

tional sparse representation method, while the texture part is fused using dictionary learning method[7].
 c) The fused cartoon and texture parts are fused to obtain a fully focused image.

According to the above algorithm, the high-frequency information in the source image needs to be extracted first[8,9,10]. Based on the MLT algorithm to obtain the required data, and according to the composite wavelet and affine system theory, when the dimension $n=2$, the affine system for composite expansion is defined as follows:

$$A_{AS(\psi)} = \{\psi_{j,l,k}(x) = |\det(A)|^{\frac{1}{2}} \psi(S^l A^j x - k)\}_{j,l \in Z, k \in Z} \quad (2.1)$$

In the formula, A and S are both 2×2 non singular matrices, $\Psi \in l^2(R^2)$ is a composite wavelet, $|\det S| = 1$. Let A be the parabolic scaling matrix, and S represent the shear matrix for $\forall_a > 0, s \in R$. Among them, $\widehat{\psi}_1 \in C^\infty(R)$ is a wavelet.

$$\text{supp} \widehat{\psi}_1 \subset [-\frac{1}{2}, \frac{1}{16}] \cup [-\frac{1}{16}, \frac{1}{2}] \quad (2.2)$$

And $\text{supp} \widehat{\psi}_1 \subset [-1, 1]$, therefore $\widehat{\psi}(0) \in C^\infty(R)$ and $\widehat{\psi}(0) \subset [-\frac{1}{2}, \frac{1}{2}]^2$ assume: $\sum_{j \geq 0} |\widehat{\psi}_1(2^{-2j}\omega)|^2 = 1, |\omega| \geq \frac{1}{8}$, and for $\forall_j \geq 0, \sum_{L=-2^j}^{2^j-1} |\widehat{\psi}_2(2^j\omega - l)|^2 = 1, |\omega| \leq 1$. From this, it can be concluded that:

$$\sum_{j \geq 0} \sum_{L=-2^j}^{2^j-1} |\widehat{\psi}(0)(\xi A_0^{-j} S_0^{-1})|^2 = \sum_{j \geq 0} \sum_{L=-2^j}^{2^j-1} |\widehat{\psi}_1(2^{-2^j} \xi_1)|^2 |\widehat{\psi}_2(2^{-2^j} \frac{\xi_2}{\xi_1} - 1)|^2 = 1 \quad (2.3)$$

For $\xi = R^2$, where x^D is the indicator function of D, $\xi \in [-\frac{1}{8}, \frac{1}{8}]^2, \widehat{\psi} \in [-\frac{1}{8}, \frac{1}{8}]^2$ and $\widehat{\psi} = 1$, set $\{\psi(x - k) : k \in Z^2\}$ is a framework of $L^2([-\frac{1}{16}, \frac{1}{16}]^2)^V$, one attribute of $\psi^d, d=0.1, G(\xi) = (\xi_1, \xi_2)$, changes continuously along the straight line $\xi_2 = \pm \xi_1$, and is used to establish a shear transformation hybrid model.

$F \in R$ is a real value sample signal, SRS is a linear combination of dictionary based prototype signals, forming sparse representation theory based on $D \in R^{n \times m}$ dictionary, among them, there are m prototype signals, and in dictionary D, there is a linear combination of prototype signals indicating $\forall x \in f, \exists s \in R^T$, such as $x \approx Ds$, where s is the sparse coefficient in D. It is usually assumed that the dictionary follows a restricted isometric attribute and is redundant, which solves the problem of reconstructing signals using optimization problems to find the non-zero component with the smallest s:

$$\min_s \|s\|_0 \text{subto} \|D_s - x\| < \varepsilon \quad (2.4)$$

Column based composite system testing solves the problem of the number of non-zero coefficients in sparse matrices, where sparse representation globally processes images, depending on the local information of the source image[11]. The general source image is divided into small blocks with a fixed dictionary D , and the sparse representation of the image is fused using sliding window technology and translation invariance. The source image I is divided into j small blocks of size $n \times n$, represented in dictionary order as vectors $v.V^j$, which can be represented as: $v = \sum_{t=1}^T S^j(t)d_t$, where j is the number of image blocks, d_t is the prototype from D and $D = [d_1 \dots d_t \dots d_T]$, it contains T prototype vectors, and $S^j = [s^1(1) \dots, s^j(t), \dots, s^j(T)]$ is a sparse representation, therefore, the image block of I is used to reconstruct a matrix v , $v=DS$, where S is a sparse matrix.

Due to the impact of noise on image fusion, in order to improve the effectiveness of image fusion, threshold processing is performed on the coefficients after MLT decomposition. The threshold is defined as:

$$\widehat{T}(\widehat{\sigma_I}) = \widehat{\sigma}_n / \widehat{\sigma_I} \quad (2.5)$$

Among them, σ_I and $\widehat{\sigma}_n$ is the standard differentiation and noise of the image source, respectively, assuming that source image I and source noise n are independent of each other, and the noise model of image z is represented as $x = I+n$. Therefore, the average noise and signal source calculations are as follows: $\sigma_x^2 = \sigma_I^2 + \sigma_n^2$. Among them, σ_x^2 is the variance of the observed signal, σ_n^2 is the noise density of the source image, and the output noise power of the source image σ_I^2 is: $\widehat{\sigma_I} = \sqrt{\max((\widehat{\sigma}_s^2 - \widehat{\sigma}_s^2), 0)}$.

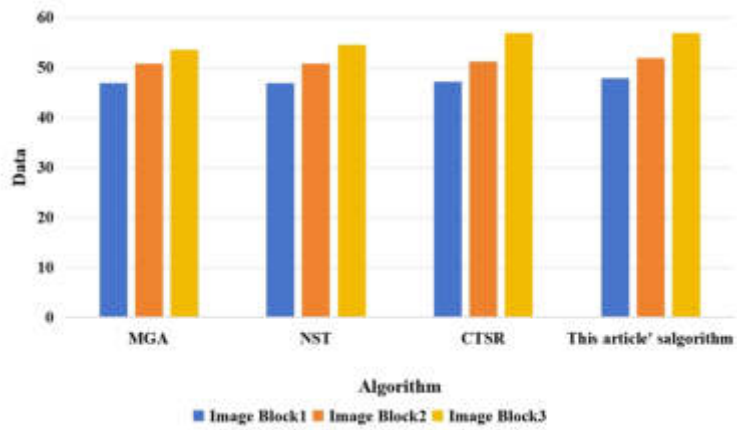
The image fusion algorithm based on MLT and SRS proposed by the author can prevent and reduce the contrast and spectral information distortion of the fused image. When some noise is detected in the source image, a given threshold is applied for filtering. The steps are as follows:

1. Perform MLT decomposition on two source images $\langle I_A, I_B \rangle$, and apply MLT to obtain their low pass band $\langle L_A, L_B \rangle$ and high pass band $\langle H_A, H_B \rangle$.
2. Perform threshold processing on low-pass and high pass using the threshold obtained from equation (5) to remove unnecessary coefficients from the decomposition.
3. Perform low-pass fusion by applying sliding window technology to $\langle I_A, I_B \rangle$, dividing image I into image blocks of size $\sqrt{n} \times \sqrt{n}$, and dividing them by step size pixels from top left to bottom right, due to the presence of $\{P_A^i\}_{i=1}^T$ and $\{P_B^i\}_{i=1}^T$ in L_A and L_B respectively, rearrange $\langle P_A^i, P_B^i \rangle$ into column vectors and rearrange $\langle \widehat{V}_A^i, \widehat{V}_B^i \rangle$ in each iteration.
4. Perform high-throughput fusion and filter using the threshold rule of formula (5) to ensure that the fused image contains the source image.
5. Perform image reconstruction and perform corresponding inverse MLT on L_F and H_F to reconstruct the final fused image I_F .

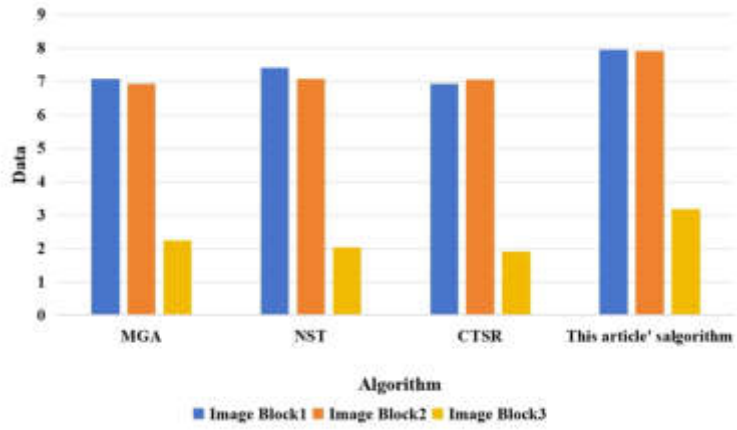
3. Experiments and Results. The experiment randomly assigns values from dictionary D with size 128×512 from image blocks in the training dataset, and then performs sparse encoding to obtain the sparse matrix of the signal. Estimated 160000 training data and 16×16 patches, randomly sampled into images, with a dictionary size set to 128. The experiment uses three commonly used metrics to evaluate the quality of fused images, namely mutual information (MI), standard deviation (SD), and entropy. The proposed algorithm is compared and analyzed with MGA, NST, and CTSR algorithms. As shown in Figure 3.1(a)-(c), the proposed algorithm achieved the best results in SD, MI, and entropy metrics, outperforming MGA, NST, and CTSR algorithms. The proposed algorithm displayed high-quality visual fusion images [12,13,14].

Experimental analysis of different noise levels, that is different standard deviations The application threshold of Equation 2.1 after MLT is used to verify the effect of removing noise and obtaining high-quality denoised images. The proposed algorithm is compared with MGA, NST, and CTSR algorithms, as shown in Figure 3.2. Figure 3.2 shows adding different levels of noise to different images σ the PSNR value shows that the proposed algorithm has higher PSNR values than MGA, NST, and CTSR algorithms in all cases of noise levels[15].

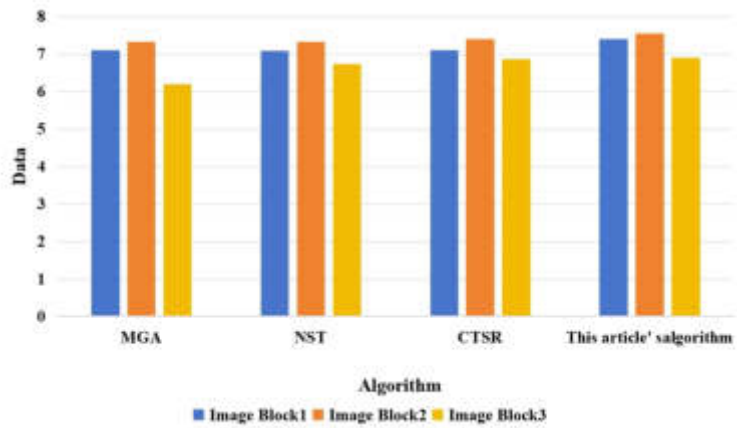
From Figure 3.3, it can be seen that the algorithm proposed by the author has the highest values in desk, Pepsi, and book. Although the comprehensive evaluation criteria in the images of flower, lab, and plane did not reach the maximum value, it can be observed that algorithms such as MGA all exhibit varying degrees of blurring of focus area judgment, block effects, and distortion of fusion boundaries during fusion. The



(a) MI



(b) SD



(c) Entropy

Fig. 3.1: Performance evaluation of the different image fusion algorithms

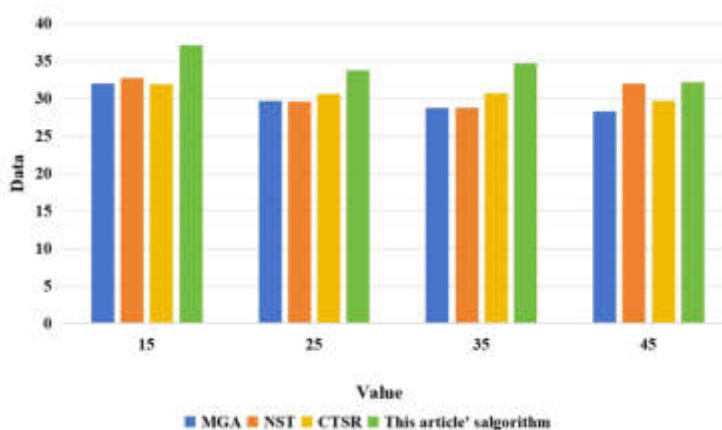


Fig. 3.2: The performance evaluation at different noise levels

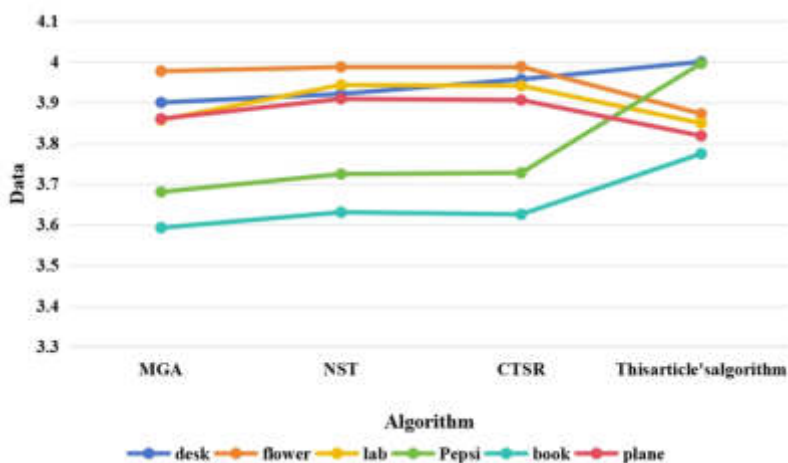


Fig. 3.3: Comparison of algorithm evaluation

emergence of these fusion results is due to the inability of the image decomposition algorithm used to describe all features in the image, inaccurate judgment of the focus area of the source image, and neglect of the structural correlation of the image. Using filtering algorithms to fuse images (as filtering algorithms can ensure that images have relatively smooth edges during fusion, but it is also because of the use of filtering algorithms that the description of fusion boundaries is not accurate enough, such as distortion and Gibbs effect) [16].

In summary, compared to the other three algorithms, the algorithm proposed by the author can provide a more complete description of the features in the image, accurately judge the focus area, maintain the structural correlation of the image, and strengthen the fusion of edge description and detail information in the fused image by fusing images [17,18,19,20].

4. Conclusion. Through the above simulation experiments and analysis, it can be concluded that the proposed hybrid image fusion and denoising algorithm based on MLT and SRS has good applicability. Under appropriate threshold conditions, it can obtain high-quality fused images and achieve the effect of removing noise from the source image, reducing the contrast and spectral information distortion of the fused image. The comparison of this algorithm with MGA, NST, and CTSR algorithms shows that the algorithm can display

high-quality visual fusion effects and maintain high PSNR values under different noise levels. However, there are also cases where there is more noise when dealing with different efficient transformation domains and less appropriate thresholds. Therefore, in the next step of research, the focus should be on improving the model transformation algorithm to address these situations, so as to better grasp the geometric shape changes in the image and graphics fusion process.

REFERENCES

- [1] Hu, Z. W. H. (2021). An efficient fusion algorithm based on hybrid multiscale decomposition for infrared-visible and multi-type images. *Infrared physics and technology*, 112(1),158.
- [2] Li, G., Lin, Y., & Qu, X. (2021). An infrared and visible image fusion method based on multi-scale transformation and norm optimization. *Information Fusion*, 71(2),145-148.
- [3] Tu, P., Huang, C., & Zhu, J. (2022). A hand gesture recognition algorithm based on multi-scale hybrid features. *Journal of Physics: Conference Series*, 2218(1), 012038.
- [4] Wei, B., Feng, X., & Wang, W. (2021). 3m: a multi-scale and multi-directional method for multi-focus image fusion. *IEEE Access*, PP(99), 1-1.
- [5] Fallah, M., & Azadbakht, M. (2021). Fusion of thermal infrared and visible images based on multi-scale transform and sparse representation. *Journal of Geospatial Information Technology*, 8(3), 39-59.
- [6] Qu, S., Liu, X., & Liang, S. (2021). Multi-scale superpixels dimension reduction hyperspectral image classification algorithm based on low rank sparse representation joint hierarchical recursive filtering. *Sensors (Basel, Switzerland)*, 21(11),63-65.
- [7] Wang, X., Su, Y., Zhang, H., & Zou, C. (2021). A new hybrid image encryption algorithm based on gray code transformation and snake-like diffusion. *The Visual Computer*, 85(7),1-22.
- [8] Wang, J., & Gao, Y. (2021). Suspect multifocus image fusion based on sparse denoising autoencoder neural network for police multimodal big data analysis. *Scientific Programming*,54(7),12-15.
- [9] Gao, X., Mou, J., Li, B., Banerjee, S., Sun, B., & Taylor, T. (2023). Multi-image hybrid encryption algorithm based on pixel substitution and gene theory. *Fractals*,69(3),52-56.
- [10] Chen, J., Zhu, Z., Hu, H., Qiu, L., Zheng, Z., & Dong, L. (2023). A novel adaptive group sparse representation model based on infrared image denoising for remote sensing application. *Applied Sciences*,74(1),63-65.
- [11] LIU Yong-sheng, CAI Shi-yang, CHEN Yi-xin, XU Zhi-bo. (2023). Point cloud denoising algorithm based on hybrid filtering and improved bilateral filtering. *Journal of Northeastern University(Natural Science)*, 44(5), 682-688.
- [12] Li, Y. Q. X. (2021). An infrared and visible image fusion method based on multi-scale transformation and norm optimization. *Information Fusion*, 71(1),65.
- [13] Wang, C., Wu, Y., Yu, Y., & Zhao, J. Q. (2022). Joint patch clustering-based adaptive dictionary and sparse representation for multi-modality image fusion. *Machine Vision and Applications*, 33(5), 1-16.
- [14] Wang, H., Li, Y., Ding, S., Pan, X., Gao, Z., & Wan, S., et al. (2022). Adaptive denoising for magnetic resonance image based on nonlocal structural similarity and low-rank sparse representation. *Cluster Computing*, 26(5), 2933-2946.
- [15] Prateek, G. V., Ju, Y. E., & Nehorai, A. (2021). Sparsity-assisted signal denoising and pattern recognition in time-series data. *Circuits Systems and Signal Processing*,124(9), 1-50.
- [16] Rebollo-Neira, L., & Inacio, A. (2023). Enhancing sparse representation of color images by cross channel transformation. *PLoS one*, 18(1), e0279917.
- [17] Miao, Y., Zakharov, Y. V., Sun, H., Li, J., & Wang, J. (2021). Underwater acoustic signal classification based on sparse time-frequency representation and deep learning. *IEEE Journal of Oceanic Engineering*,36(7),52-56.
- [18] Zhang, J., Chen, J., Yu, H., Yang, D., & Xing, M. (2021). Learning an sar image despeckling model via weighted sparse representation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14,(7) 7148-7158.
- [19] An, F. P., Ma, X. M., & Bai, L. (2022). Image fusion algorithm based on unsupervised deep learning-optimized sparse representation. *Biomedical signal processing and control*(Jan. Pt.B), 85(3),71.
- [20] Liu, Z., Wang, L., Feng, Y., Qian, Z., & Chen, X. (2021). A recognition method for time-frequency overlapped waveform-agile radar signals based on matrix transformation and multi-scale center point detection. *Applied Acoustics*, 175(4), 107855.

Edited by: Zhigao Zheng

Special issue on: Graph Powered Big Aerospace Data Processing

Received: Dec 16, 2023

Accepted: Jan 9, 2024



DIGITAL MEDIA INTERNET MODELING SYSTEM UNDER COMPUTER ARTIFICIAL INTELLIGENCE TECHNOLOGY

MIAOJUN LI* QI LI† CHANGRONG PENG ‡ AND XIAODONG ZHANG§

Abstract. With the continuous expansion of network scale, the hierarchical and modular characteristics of network structure are becoming increasingly prominent. The traditional single-layer network research paradigm has certain limitations in characterizing the complex relationships between various network systems. Analyzing and constructing models for multi-layer networks has gradually become an important direction in complex network research. The author proposes a multi-layer network model for the Internet and the significance of constructing multi-layer network models in the field of the Internet, based on the characteristics of the multi-layer network structure presented in the application process of the Internet. By analyzing the characteristics of internet data, a multi-layer network model covering three types of networks, namely the internet infrastructure layer, business application layer, and user account layer, was designed. The experimental results show that the average shortest path lengths of the three networks, namely routing relationship network, web hyperchain network, and email address network, are 3.8, 2.9, and 1.7, respectively. Due to the existence of inter layer edges in multi-layer networks, nodes can reach each other across layers in addition to intra layer edges. Experimental results show that the average shortest path length of the three networks has been shortened to some extent, with values of 3.1, 2.7, and 1.6, respectively. On the basis of the single-layer network generation model, a layer correlation method for multi-layer networks was designed, and the model construction of multi-layer networks was achieved.

Key words: Internet, Digital media, Kernel distribution, Multi layer network, model building

1. Introduction. With the continuous development and application of computer artificial intelligence technology, the field of digital media internet has shown enormous development potential. Digital media internet refers to a form of media that utilizes computer networks and digital technology for information dissemination, content exchange, and user interaction [1]. It has become an important way for people to obtain information, entertainment, and socialize in today's society. However, the rapid development of digital media internet has also brought a series of challenges and problems, such as information overload, uneven content quality, and poor user experience. Firstly, with the popularization of the Internet and the rapid development of digital media, people are facing a huge problem of information overload. In the era of digital media and the internet, people can easily access a large amount of information, but at the same time, they also face difficulties in information filtering and screening. Excessive information may lead people to be unable to quickly and accurately find the content they need, and even fall into the dilemma of information overload. In order to solve this problem, digital media internet platforms need to strengthen the research and development of information classification, recommendation algorithms, and other aspects, provide personalized information services, and help users better obtain the required information. Secondly, the uneven content quality of digital media internet is also an urgent issue that needs to be addressed. With the development of the Internet, anyone can easily publish content on digital media platforms, which leads to uneven quality of content. Some information may be misleading, false, or even illegal, causing confusion and distress to users. In order to improve content quality, digital media internet platforms need to strengthen content review and supervision, establish effective information dissemination mechanisms and content management systems [2]. At the same time, users also need to improve their ability to distinguish information and avoid blindly believing and disseminating unreliable information. In addition, the user experience of digital media internet platforms also needs to be further improved. Although the digital media internet provides users with abundant information resources and communication platforms, sometimes

*School of Architecture and Art, Ningbo Vocational and Technical College, Ningbo, 315800, China.

†College of Arts, Cheongju University, Cheongju, 28503, Korea. (Corresponding author, liqi20231113@163.com)

‡College of Art, Hebei University of Economics and Business, Shijiazhuang, 050061, China.

§College of Art, Hebei University of Economics and Business, Shijiazhuang, 050061, China.

users may encounter some problems during the process of using the digital media internet, such as slow page loading speed, excessive advertising, etc. These issues may affect the user experience and even make them lose interest. Digital media internet platforms need to continuously optimize user interface design and improve system performance to provide a smoother, more convenient, and personalized user experience.

In addition, the rapid development of digital media internet has also brought privacy and security issues. In the era of digital media and the internet, the leakage of personal information and network security threats are becoming increasingly serious. In order to protect the personal privacy and information security of users, digital media internet platforms need to strengthen the protection measures of user data, encrypt user data, and establish sound security mechanisms. At the same time, users should also raise their security awareness, strengthen the protection of personal information, and make reasonable use of digital media internet platforms [3].

In short, with the continuous development and application of computer artificial intelligence technology, the field of digital media internet has shown enormous development potential. However, the rapid development of digital media internet has also brought a series of problems and challenges [4,5]. In order to overcome these issues, digital media internet platforms need to strengthen information filtering and filtering, improve content quality, improve user experience, and protect user privacy and information security. Only in this way can digital media and the internet better provide people with high-quality information services, promote social progress and development.

2. Methods.

2.1. Current status of multi-layer network modeling research.

(1) *Multi layer network generation model.* At present, multi-layer network modeling is still in its early stages, and multi-layer network generation models can be divided into two categories. One type is the growing multi-layer network model, in which the number of nodes gradually increases according to a generalized priority connection rule. These models explain the evolution process of multi-layer networks from simple and basic dynamic laws. The process of constructing this type of model is as follows: 1) Growth, adding a new node in each layer of a multi-layer network at each time step, and connecting the new nodes in each layer to other nodes in the same layer through m links; 2) Generalized priority connection, where new nodes select existing nodes in the network to connect based on the probability of intra layer and inter layer connections [6]. Another type is multi-layer network composite models, which consider multi-layer networks as a collection of single-layer networks that satisfy certain structural constraints between layers. These sets can generate multi-layer networks with degree correlation and controllable overlap. The idea of this type of model is to first consider the generation of each network layer and then consider the connections between layers, where each layer is generated according to the static model of traditional single-layer networks: Given the node degree sequence of each layer, a configuration model is used to obtain a specific network implementation for the given connection set [7]. There are two ways to connect layers: One is to add any inter layer connections, and the other is to specify inter layer edges by giving a joint degree distribution. Although the terms "joint degree sequence", "joint degree matrix", and "joint degree distribution" are commonly used in literature to represent the number of degrees related to the end of link nodes in a simple network, they explicitly refer to the connections between layers. In the process of constructing multi-layer networks, the selection of any network model will impose certain constraints on existing research, and each network model corresponds to a set of networks that satisfy the implicit constraints imposed by the model. The computer internet is a typical complex network system, and multi-source data often has multiple types of correlations, making computer network systems have multi-layer network characteristics. Therefore, it is very important to study the generation model of multi-layer network security that conforms to the actual network characteristics.

(2) *Typical network model metrics.* The degree k of node i in a complex network is defined as the number of edges connected to that node, and the average of the total degrees of all nodes is called the average degree. $P(k)$ represents the proportion of nodes with degree k in the network to all nodes, used to represent the distribution of node degrees in the network. If there are N nodes in the network and n_k nodes with degree k , then:

$$p(k) = \frac{n_k}{N} \quad (2.1)$$

The k -kernel of a complex network is defined as the remaining subgraph after sequentially removing nodes

with degree $k-1$ from the network. If a node is in the k -kernel of the network and not in the graph with degree $k-1$, then that node is a k -kernel. $g(k)$ represents the proportion of nodes with kernel number k to all nodes in the network, used to represent the distribution of node kernels in the network[8]. If there are N nodes in the network and n_k nodes with kernel number k , then:

$$g(k) = \frac{n_k}{N} \quad (2.2)$$

The average path length L of a complex network is defined as the average of the shortest path length d_{ij} between any two points i and j in the network. This indicator is usually used to measure the efficiency of the network. The smaller the average path length of the network, the higher the efficiency of the network. Its specific expression is:

$$L = \frac{2}{N(N-1)} \sum_{i \neq j} d_{ij} \quad (2.3)$$

The diameter D of a complex network is defined as the maximum distance between all nodes in the network.

$$D = \max_{i,j \in N} d_{ij} \quad (2.4)$$

2.2. Analysis of multi-layer network structure of the Internet . With the continuous expansion of network scale, various types of networks have formed in the Internet. In addition to the interconnection of underlying network devices, various business system related networks have formed at the application layer, and various network account related networks have formed at the user layer. The entire Internet exhibits characteristics of diversity, heterogeneity, and layering [9]. Traditional single-layer networks cannot characterize the multidimensional characteristics of internet data, so it is necessary to establish a multi-layer network model in the field of the internet. The author proposes a multi-layer network structure for the Internet from three layers: basic device layer, business application layer, and user role layer.

The infrastructure layer is a physical network established by various network devices interconnected through physical links, with nodes mainly including routers, switches, servers, etc; The business application layer is mainly a logical network formed by various application systems during the communication process of various business applications [10]. Its nodes mainly include, as there are many business systems operating in the Internet, different types of business systems may form their own networks. When these application systems are laid out on the same server device, there may be correlation relationships between different business system networks. Virtual entity composition, such as various sites, application systems, and business software, including various types of websites, email addresses, etc [11]. The nodes in the user role layer mainly include various accounts registered and used by people during the process of using the Internet. The entities and attribute elements of each layer are listed in Table 2.1.

In the application process of the entire Internet, each layer can form different types of networks based on different connections. For example, in the basic equipment layer, nodes form communication networks according to the relationship between optical cables and ground cables; In the business application layer, different business relationship networks are formed based on different types of business, such as email address networks, web hyperlink networks, etc; At the user role level, different types of social networks are formed based on different social relationships, such as friendship networks, Weibo account following networks, etc. From a vertical perspective, the lower level network nodes are the foundation for the existence of the upper level network nodes. For example, each application system operates on servers in the lower level network, and each account is registered on application software. There is a strong dependency relationship between layers [12].

This multi-layered network structure representation for the Internet plays an important supporting role in characterizing network spatial trends, analyzing network vulnerabilities, and conducting network tracing and evidence collection.

From the perspective of network situation display, by constructing a multi-layer network in the Internet field, Internet data can be presented to network users in a multi-level and multi-scale form, in order to understand the structural characteristics of the network and provide model support for subsequent network decision-making

Table 2.1: List of the three-tier network model elements

Network hierarchy	Entity	Attribute
Basic equipment layer	Server, computer, router, access device	City, IP address, autonomous domain number, port, manufacturer
Business application layer	Software, various types of data	System software: Windows, Ubuntu, Centos Application software: WeChat, email, website data: account and password, text, video
User Role Layer	Account	Registration time, registration unit, etc

Table 3.1: Statistics of the actual network characteristic parameters

number	project	Routing network
1	Number of nodes	192244
2	Number of edges	609066
3	Average path length	4.61
4	Average degree	6.34
5	Maximum degree	1071
6	Network diameter	12
7	Number of network cores	31

and deployment[13]. From the perspective of analyzing network vulnerability, the analysis of key nodes based on multi-layer network structure will have a significant impact on the analysis process, which may enable cross layer attack paths that were previously unreachable in single-layer networks to be reachable in multi-layer networks. This provides a new research approach for conducting network vulnerability analysis and evaluation. In terms of network traceability, multi-layer network models can more efficiently achieve this function. For example, if a company receives a phishing email and uses a multi-layer network to analyze and trace criminals, the process is as follows: First, match the email with logs to obtain the sender's email address, locate the email sender upwards, and locate the email sender's IP downwards [14]. If the IP is not the attack source, use the association relationship of the email address to find the address that has email exchanges with the email sender, and further locate the user upwards, downward positioning to achieve multi-path traceability of the target.

3. Experiments and Analysis.

3.1. Single layer network modeling experiment. The author used Python and its NetworkX module for network modeling and related feature parameter statistics experiments. In order to better demonstrate the practicality of the CGN model constructed by the author, actual network topology data was first obtained, and then an equally sized network was generated using the CGN model. The relevant statistical parameters of the CGN model generated network and the actual network were compared and analyzed [15].

The model validation data is sourced from the open-source dataset website CAIDA1. In complex network theory, network topology is generally described using statistical characteristics such as the number of nodes, edges, degree and degree distribution, kernel number and kernel distribution. After calculation, the statistical situation of the relevant characteristic parameters of the actual network is listed in Table 3.1.

The author intends to verify the applicability of the model by comparing the relevant parameters of the actual network and the network constructed by the author's proposed model. In order to ensure the reliability of the experimental results, 10 experiments were conducted using the CGN model to generate networks of the same scale. The average of these 10 statistical results was taken as the comparison parameter for each feature parameter.

Table 3.2: The comparison of static statistical characteristics

number	project	Routing network	CGN network
1	Number of nodes	192244	192244
2	Number of edges	609066	614526
3	Average path length	4.61	4.56
4	Average degree	6.34	6.16
5	Maximum degree	1071	1108
6	Network diameter	12	11
7	Number of network cores	31	31

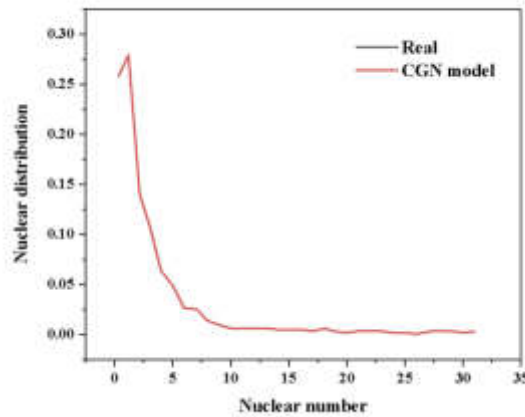


Fig. 3.1: Comparison of Kernel Distribution

From Table 3.2, it can be seen that the actual network and the number of generated network nodes are completely consistent with the number of network cores, the number of node edges is equivalent, the average path length is basically consistent, and the average degree and maximum degree are also similar, indicating that the model has good applicability [16].

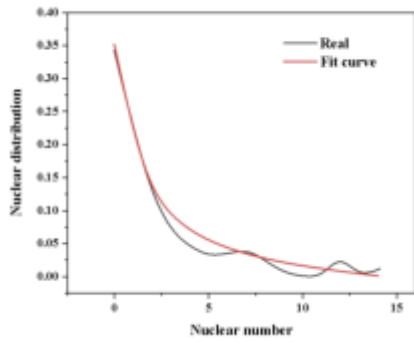
The CGN model can achieve complete controllability of the number of cores of nodes in the network, which is reflected in the fact that the network generated by the model is completely consistent with the actual network in terms of the distribution of cores, as shown in Figure 3.1.

3.2. Internet multi-layer network construction experiment . The three layers of the multi-layer network experiment are routing relationship network, web hyperchain network, and email address network. The data sources for the three layers of the network are all from the NetworkData sets network database website [17].

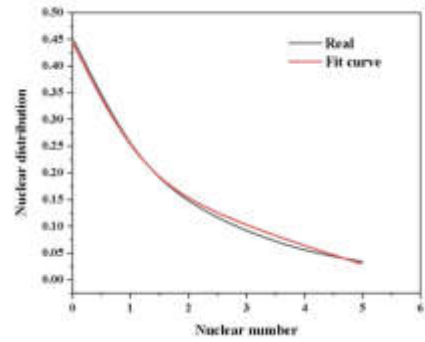
Based on the actual network data obtained, calculate the kernel distribution of the three networks, as shown in Figure 3.2(a-c). The kernel distribution of the three networks follows a power-law distribution, and the function is expressed as:

$$P(K > k) = \alpha \cdot k^\rho + \lambda(5) \quad (3.1)$$

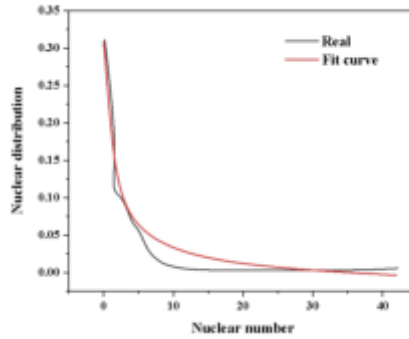
Fit the kernel distributions of three networks using the machine learning library sklearn provided by a third-party in Python, and obtain the corresponding kernel distributions for the three networks α , β , λ , as listed in Table 3.3.



(a) Route network kernel distribution



(b) Core distribution of web page hyperchain network



(c) Core distribution of the email address network

Fig. 3.2: Nuclear distribution fitting

Table 3.3: Network power-law parameters

Network type	α	β	λ
Routing network	0.40	0.79	-0.04
Web hyperlink network	0.64	0.59	-0.19
Email address network	0.32	0.93	-0.01

Based on the obtained kernel distribution functions of the three networks, the sequence of network kernels for a specified network size can be obtained. The node sizes of the three-layer network routing relationship network, webpage hyperlink network, and email address network constructed by the author are $N_1 = 200$, $N_2 = 50$, and $N_3 = 20$, respectively. Next, let's consider inter layer connectivity. In a routing network, both computers and servers are located at terminal nodes, while other nodes are router nodes. As actual business applications run on computers and servers, nodes in the business application layer will only have connectivity with terminal nodes in the basic device layer, that is, nodes with a moderate degree of 1 between the business application layer and the basic device layer will have connectivity. Therefore, the interlayer edge ratio here 1 corresponds to the proportion of connected edges between the terminal nodes in the routing network layer and the upper layer.

Randomly select interlayer node ratio $\alpha_1 = 25\%$, $\alpha_{21} = 55\%$, $\alpha_{23} = 40\%$, $\alpha_3 = 100\%$, corresponding interlayer edge connection parameters $\beta_{12} = \beta_{23} = 50\%$. Associate the three-layer network according to parameters[18].

(2) *Analysis of multi-layer network characteristics.* The average shortest path lengths for the routing relationship network, web hyperlink network, and email address network are 3.8, 2.9, and 1.7, respectively. Due to the existence of inter layer edges in multi-layer networks, nodes can reach each other across layers in addition to intra layer edges. Experimental results show that the average shortest path length of the three networks has been shortened to some extent, with values of 3.1, 2.7, and 1.6, respectively. Due to the small scale of the author's experimental data, it is not reflected clearly enough. But overall, there is a decreasing trend, and experiments have shown that compared to single-layer networks, multi-layer networks achieve multi-path reachability of nodes and reduce the average path length between nodes[19].

4. Conclusion. The author designed a three-layer network model for multi-dimensional data of the Internet based on the characteristics of the multi-layer network architecture, defining the elements, attributes, and the meanings of intra layer and inter layer edges of each single-layer network. Based on the idea of k-kernel decomposition, the author proposes a single-layer network model generation algorithm for digital media, designs a multi-layer network generation model on this basis, and finally proposes the significance of constructing multi-layer network models in the field of interconnection networks. As an emerging research direction in recent years, multi-layer networks have been applied in related fields, but there are still many scientific problems worth exploring. There is still a long way to go in establishing its theoretical framework and enriching databases in related fields. Establishing a multi-layer network model in the field of the Internet is of great significance for network security personnel to correctly assess the network security situation, master key network nodes, and make decisions and deployments. This is an important direction for future research.

REFERENCES

- [1] Qian, J. (2022). Research on artificial intelligence technology of virtual reality teaching method in digital media art creation. *Journal of Internet Technology*, 75(1), 23.
- [2] Chen, Y. (2022). Research on application of computer artificial intelligence technology in intelligent substation monitoring system. *Highlights in Science, Engineering and Technology*, 14(9), 1881-1891.
- [3] Lu, W. (2021). Research on marketing development under the background of artificial intelligence technology. *Journal of Physics Conference Series*, 1769(1), 012071.
- [4] Tavakoli, S. S., Mozaffari, A., Danaei, A., & Rashidi, E. (2023). Explaining the effect of artificial intelligence on the technology acceptance model in media: a cloud computing approach. *The Electronic Library: The International Journal for Minicomputer, Microcomputer, and Software Applications in Libraries*, 20(6), 2963-2992.
- [5] Wu, J., Wang, X., Dang, Y., & Lv, Z. (2022). Digital twins and artificial intelligence in transportation infrastructure: classification, application, and future research directions. *Computers and Electrical Engineering*, 62(5), 101.
- [6] Geelan, T. (2021). Introduction to the special issue - the internet, social media and trade union revitalization: still behind the digital curve or catching up?. *New Technology, Work and Employment*, 36(2), 85-89.
- [7] Zhuang, D., Gao, F., & Zhu, R. (2021). Construction of mechanical control system based on artificial intelligence technology. *2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture*, 56(2), 1052-1061.
- [8] Meng, Q. (2021). Research on the application of computer network technology under the background of artificial intelligence cloud technology. *Journal of Physics: Conference Series*, 1802(4), 042067.
- [9] Nazir, S., Khadim, S., Asadullah, M. A., & Syed, N. (2023). Exploring the influence of artificial intelligence technology on consumer repurchase intention: the mediation and moderation approach. *Technology in society*, 13(2), 45-56.
- [10] Korjian, S., & Gibson, C. M. (2022). Digital technologies and the democratization of clinical research: social media, wearables, and artificial intelligence. *Contemporary clinical trials*, 117(7), 106767.
- [11] Kim, Y. H., Oh, N. Y., & Park, J. W. (2021). A case study on the digital media exhibition planning and artificial intelligence and data artworks of isea 2019. *Journal of Digital Contents Society*, 22(2), 243-251.
- [12] Zhang, J. (2021). Computer assisted instruction system under artificial intelligence technology. *International Journal of Emerging Technologies in Learning (iJET)*, 11(1), 69-89.
- [13] Chen, Z. (2023). Research on the application of artificial intelligence under big data under communication network technology. *Highlights in Science, Engineering and Technology*, 11(16), 10479-10486.
- [14] Ma, W., Zhao, X., & Guo, Y. (2021). Improving the effectiveness of traditional education based on computer artificial intelligence and neural network system. *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*, 74(2), 40.
- [15] Ban, H., & Ning, J. (2021). Online english teaching based on artificial intelligence internet technology embedded system. *Hindawi Limited*, 50(35), 12316-12323.
- [16] Chen, C. H., & Zhu, X. (2021). Application research on information security of aerobics information digital system based on internet of things technology. *Journal of Intelligent and Fuzzy Systems*, 44(14), 1-8.

- [17] Tong, Y., Wu, J., & Zhang, X. (2021). Research on interdisciplinarity-teaching of digital media art under big data. *Journal of Physics: Conference Series*, 1883(1), 012145 (6pp).
- [18] Yunjie, J. (2021). Research on the practice of college english classroom teaching based on internet and artificial intelligence. *Journal of Intelligent and Fuzzy Systems*,87(1), 1-10.
- [19] Li, Y. (2021). Research on the construction of tefl resource database system based on artificial intelligence. *Journal of Intelligent and Fuzzy Systems*,147(6), 1-12.

Edited by: Zhigao Zheng

Special issue on: Graph Powered Big Aerospace Data Processing

Received: Dec 16, 2023

Accepted: Jan 9, 2024



E-COMMERCE DATA MINING ANALYSIS BASED ON USER PREFERENCES AND ASSOCIATION RULES

YUN ZHANG*

Abstract. With the development of network technology, online shopping is becoming more and more convenient. But the increasing number of products also makes it difficult for consumers to make the right decision. When there is no apparent market demand, how to recommend products with commercial potential to customers has become an urgent problem for businesses to solve. This paper proposes e-commerce product recommendation based on user preference and association rule algorithm aiming at the problems existing in e-commerce product recommendation. Firstly, this paper constructs a user interest modeling method. Through analyzing users' interests and preferences, to provide users with timely and accurate personalized services. Then, the FP_Growth algorithm is optimized and improved. A more effective CTE-MARM algorithm is designed, and an association rules database based on user benefit items is constructed and analyzed jointly. Analyze products with strong correlations. According to consumers' interest levels, TOP-N is the best product choice. Experiments show that the algorithm has higher prediction accuracy. The research results of this project can not only improve enterprises' ability to analyse data and provide data support for enterprises to carry out effective marketing management.

Key words: Data mining; Association rules; User preference; Electronic commerce; CTE-MARM algorithm

1. Introduction. In modern society, with the rapid development of science and technology, human beings are also faced with the problem of "excess" and the convenience of obtaining science and technology. This phenomenon is no exception in e-commerce. Making customers satisfied with products is a significant issue for e-commerce enterprises. This is how the product recommendation technology in e-commerce emerged [1]. It applies the method of data mining to the actual consumer behavior scenario. Possible business opportunities can be predicted through the correlation mining of historical data. This saves users time finding items they like and increases sales and customer loyalty.

In the same frame, it is of great theoretical value and practical significance to study three different types of customer evaluation: customer perceived value, customer satisfaction, and customer purchase intention. Previous studies have shown that perceived value is the pre-variable of consumer satisfaction. Risk perception, individual innovation, social impact, and perception of usefulness directly and significantly impact user intent. Perceived ease of use and social influence directly impact user availability but cannot play an indirect role through perceived usefulness. Many psychological barriers of consumers to e-commerce, especially the security and reputation problems faced by e-commerce users, will significantly impact consumers' purchase intentions and cause significant obstacles to the rapid development of e-commerce. After a questionnaire survey and analysis of e-commerce users, some researchers found that social factors have the most apparent positive effect on e-commerce intentions, while they have no noticeable effect on e-commerce expected practicability and risk perception. Mobile phone payment intention positively affects users' use, but convenience has no noticeable promoting effect [2]. Some scholars use the integrated technology acceptance model to study electronic transaction intention from six perspectives: perception of usefulness, perception of ease of use, perception of risk, trust, trust and evaluation, and attitude of use. Relevant studies mainly focus on the perception and evaluation of consumer behavior but lack systematic collection and analysis of mobile payment users' characteristics and future consumer behavior [3]. This makes it a complex problem to accurately position the market of its target users in the initial stage of the development of e-commerce.

*Zhejiang Yuexiu University, Shaoxing, 312000, China (460465456@qq.com)

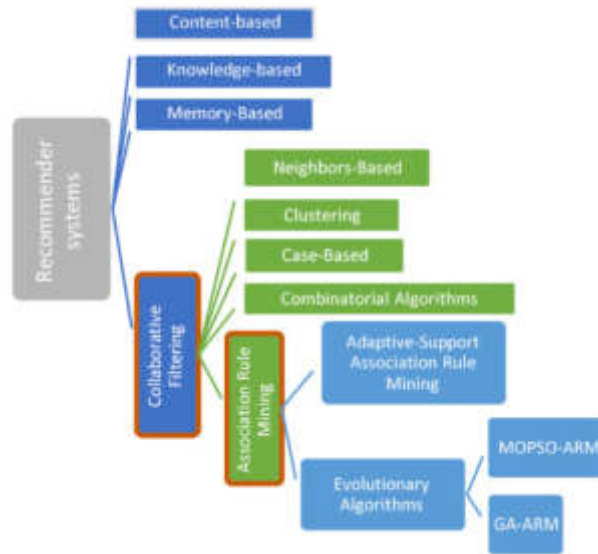


Fig. 2.1: Association rules model flow.

2. Demand analysis.

2.1. Functional Requirements. The key to an e-commerce product recommendation system is to analyze the interests and preferences of users by collecting their shopping habits and shopping records. Then, explore and predict the potential shopping opportunities [4]. The most important thing is to analyze users' personalization and real-time performance. The system implements the following essential functions.

1. Data collection: extracted relevant records and operational data.
2. Preprocess and eliminate unnecessary data to ensure data integrity.
3. The user's interest model is established and classified.
4. Establish the relevant rule database of data mining.
5. Product recommendation for products that are attractive to users.

2.2. Key Parameters.

1. Credibility is introduced to reduce the phenomenon of "rule explosion" and improve the search accuracy;
2. The prescription describes the change in consumption behavior in real life. The closer to the current consumption behavior, the more it can reflect the current demand preferences.
3. The purchase, browsing, collection, rating, and comments of e-commerce consumers can reflect consumers' interests [5]. Users have different levels of interest in different business activities. Use the triplet method to sort the required items in order IR_{ij} . The "user-benefit item chain" comprises the N items with the highest information value by the TOP-N method. The CTE-MARM method is proposed for data association [6]. When good i in the list of customer-benefit items has a strong correlation with good k , good k is added to the tripartite group. The model flow diagram is shown in Figure 2.1.

3. Association rule algorithm model design. This topic focuses on how to quickly find the user group with specific characteristics in the process of commercial marketing of e-commerce enterprises. Discover the characteristics of the target user group. This user group has more complex characteristics [7]. The value of the variable is random and difficult to transform. Each attribute is unrelated, so it is challenging to meet the needs of conventional statistical methods using standard multiple regression analysis, structural equation model, technology acceptance model, technology acceptance model, and integrated technology acceptance model. It is challenging to realize the rapid positioning of specific user groups. In the practical problems, from many incomplete, noisy, fuzzy, random, and other practical problems, we extract the information and knowledge that

people do not know in advance but have potential value [8]. Then, the information in these eigenvalues can be obtained by mining the association rules.

3.1. Principles of Association Rules. Association rules were proposed by R. Agrawal in 1993 to describe the internal relationship between various items in a database. It is one of the essential research contents in data mining [9]. Association rules can directly represent the relationship between a collection of items in a data set. These associations are not based on a specific distribution or depend on a specific pattern. It depends only on the probability of occurrence of the set of items in a pattern. Suppose $Q = \{q_1, q_2, \dots, q_m\}$ is the set of all entries. H represents the transaction database and T represents project subset ($T \subseteq Q$). Each transaction has its unique transaction identifier $TID.B$ is the set of entries. Transaction record T includes item B . Its necessary and sufficient condition is that if the entry B contains k items, it is called the set of k items. The proportion of the number of items B in the transaction database H is called the level of support for the item set. If the support level of an item set exceeds the minimum support threshold set by the user, it is called a frequent item set. The law of association is the logical implication of class $U \Rightarrow V$. Where $U \subset Q, V \subset Q$, and $U \cap V = \emptyset$. If transaction database contains $s\%$ transactions and includes UV , then the support of trading system $U \Rightarrow V$ is $s\%$. The absolute amount of support is the possible number. If $\text{support}(U)$ is used to represent the degree of support for item U , then $\text{support}(UV)/\text{support}(U)$ can be used to represent the degree of confidence of the rule, which is the conditional probability $P(V | U)$.

$$\begin{aligned}\text{support}(U \Rightarrow V) &= P(UV) \\ \text{confidence}(U \Rightarrow V) &= P(V | U)\end{aligned}$$

The related rules conforming to the minimum support and threshold are called strict rules [10]. The maximum is 0 – 100%. That is, whether the item on the right will be selected when the item on the left is purchased or whether the item on the right will be selected in any situation. In the process of selecting target customers, the size of the revenue is the most important. The greater the revenue value, the greater the customer demand for the service. This criterion is the same as the selection criteria for other data exploration modes [11]. To measure how much this criterion improves the prediction accuracy by comparing it with the "original" criterion.

3.2. Evaluate the role of the "promotion" attribute group in association rules. Assume that the mobile user base is constant. It grows at a constant rate over some time. But compared to the number of existing mobile phone users, the number of new mobile phone users is still relatively small [12]. For this reason, the number of users will grow in a particular proportion over some time, so at a certain point in time, the number of mobile phone users can be limited to:

$$\text{total users} = IE$$

I is the invariant growth factor. E is the number of mobile phone users at a given point in time. In association rule $[D, Z, S]$, D is support, Z is credibility, and S is revenue. The number of mobile phone users $\times D$ can be considered to be able to support a certain number of mobile phone users. Under the promotion effect of the previous part, the mobile user number H can promote the latter part. The number of mobile subscribers $\times D \times Z \times S$ can be understood as selling mobile phones to a specific group [13]. The number of mobile users that can make subsequent things happen. Using the number of mobile phones $\times D \times Z \times S$, the "promotion" effect of the former term on the latter term can be evaluated, and the "promotion" effect is more prominent when the effect value is more significant. Since the number of mobile phone users IE is constant at a certain point in time, a "boost" factor is introduced here:

$$\phi = D \times Z \times S$$

D stands for support, Z for credibility, and S for revenue. The "boost" factor is used to measure the strength of the "boost" connection between the first product and the later product [14]. With the increase of ϕ , this "promoting" effect gradually increases. It is targeted at this audience characteristic for marketing publicity, which will receive a good effect.

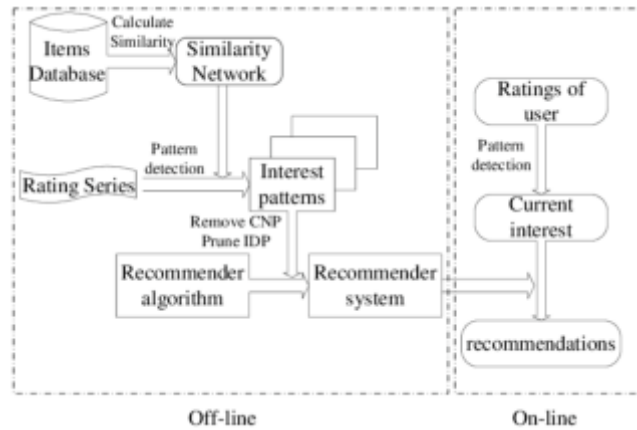


Fig. 4.1: *User interest model framework.*

4. User interest model. Building a user’s interest model can give better feedback on the user’s interest. A person’s hobbies are generally divided into two categories: one is long-term, and the other is short-term. A structural diagram of the user benefit pattern is shown in Figure 4.1.

Long-term interests reflect a person’s preference for one thing over a while. This hobby is not something that will change over some time but a constant state. Most of these are gradually accumulated over a long period of life, and naturally, it is also related to the individual’s educational experience, life background and personality [15]. Short-term interest usually refers to a specific period. People prefer a thing because of some factors and stimuli, but external stimuli will change this preference. Of course, short-term interests can turn into long-term interests. These fleeting interests will gradually dissipate over time. Because the amount of information users is interested in for a long time is considerable, it is necessary to divide it reasonably. If we use the mathematical formula $S = \{(s_1, \varepsilon_1, r_1), (s_2, \varepsilon_2, r_2), \dots, (s_n, \varepsilon_n, r_n)\}$ to describe a person’s long-term interests, then S_i is what we care about. The ε_i stands for the user’s love for the product. r_i indicates that users pay more attention to an event because they prefer it.

Because the short-term interests of users change at any time, it is challenging to transform them into long-term interests, and they are often fleeting, so this paper does not use the method of statistics on the short-term interests of users. The mathematical formula expresses the user’s short-term interest in this paper d. β is for something. ξ stands for the level at which users like a product. Then, the user’s short-term interest can be expressed by $D = (d_1, d_2, \dots, d_m)$. In a shopping system based on consumer preferences, this indicator can reflect in real time the change of consumers’ preferences for fresh items decreasing over time. We define the expression of freshness as: In a shopping system based on consumer preferences, this indicator can reflect in real time the change of consumers’ preferences for fresh items decreasing over time. We define the expression of freshness as:

$$u_t = \begin{cases} 0 & 0 \leq u_{t_0} \leq \varphi \\ \left(\frac{u_{t_0} - \varphi}{u_{t_0}} \right) & u_{t_0} > \varphi \end{cases}$$

t_0 is the past time. t means now. u_{t_0} indicates that the user has previously found the item novel. u_t represents that the user currently finds the item novel. φ stands for a constant number. If freshness is set to 0, then freshness is considered 0. Freshness means that it will gradually decrease over time until it reaches 0. He has lost interest in the item if the novelty drops to zero. According to different user behavior records, different keywords are automatically generated, and the corresponding database is built for personalized search. However, since the value of the freshness decreases over time, it is also necessary to calculate the weight of the freshness. Then, the calculation of novelty weighting for specific keywords can be expressed by expression.

$$(\beta_p, \xi_1, t_1, h_1, u_1), (\beta_p, \xi_2, t_2, h_2, u_2) \cdots (\beta_p, \xi_n, t_n, h_n, u_n) : h_1, \dots, h_n$$

The user's recent preferences on a particular tag can be weighted. u is used to represent the total weight. $\xi_p^t = \xi_1 * u'_1 + \xi_2 * u'_2 \dots + \xi_n * u'_n \cdot u'_1, u'_2, \dots, u'_n$ represents the current freshness value and takes it as the freshness value at t . After weighing each keyword, the interest set can be obtained quickly.

Keywords are weighted according to different user preferences. The detailed calculation method is as follows:

$$S = \{(s_1, \varepsilon_1, r_1), (s_2, \varepsilon_2, r_2), \dots, (s_n, \varepsilon_n, r_n)\}$$

Since each class has keyword glyph, the weight $s_i = ((\beta_1, w_1), (\beta_2, w_2), \dots, (\beta_n, w_n))$ of different classes can be obtained by statistical analysis of text feature vectors of different classes. In addition, using the relevant algorithm, the user's interest category can be accurately displayed.

Input: Important bytes that are currently of interest to the user;

Output: the user's long-term interest choice;

$$S = \{(s_1, \varepsilon_1, r_1), (s_2, \varepsilon_2, r_2), \dots, (s_n, \varepsilon_n, r_n)\}$$

1. Extract all tuples generated on A day t_0 and define them as follows: keyword β , initial weight ξ , creation time t , document containing keyword h , freshness u .
Firstly, $\xi_p^t = \xi_1 * u'_1 + \xi_2 * u'_2 \dots + \xi_n * u'_n$ weighted analysis is carried out for a particular keyword. The corresponding weights of each keyword are obtained after Δt .
2. Repeat step 1 until all tuples have been calculated. Set threshold λ and find a keyword that is larger than this value. Record it in Group ψ . Then, the corresponding text H_0 is determined according to the relationship between the keyword and the tuple.
3. The selection range of the selected associated keywords and text need not be too wide, so according to the threshold value of the initial weight of the keyword, you can find the corresponding collection of files whose original weight exceeds a particular critical value. And as a category according to their level of long-term interest in the category. Use $S_i = (s_1, S_2, \dots, S_m) s_i$ to represent a class.
4. Update keywords when they do not reach a particular range. Re-evaluate it to see if it's of long-term interest.
5. The users' long-term interest set is obtained by calculating steps 3 and 4. Calculate the number of files r_i in each category to get $((s_1, r_1), (s_2, r_2), \dots, (s_n, r_n))$. $\varepsilon_i = r_i / \sum r_i$ represents the extent to which users have different preferences for each category. Generate an end-user long-term interest statement:

$$S = \{(s_1, \varepsilon_1, r_1), (s_2, \varepsilon_2, r_2), \dots, (s_n, \varepsilon_n, r_n)\}$$

5. Design of e-commerce product recommendation system.

5.1. System Architecture. The recommendation of e-commerce products is studied using the calculation method of association rules and user preferences. Figure 5.1 shows the overall architectural design (Picture quoted from Egyptian Informatics Journal, Volume 23, Issue 1, March 2022, Pages 33-45). The system includes data collection and cleaning, user interest analysis, building association rule base, and recommending TOP-N.

- (1) Data collection and cleaning: data collection is divided into two parts: shopping and various business activity information. These two data types provide the basis for discovering and researching users' interests. In addition, many messy, fuzzy, incomplete, and dirty data must be cleaned to improve the mining effect. Through the detection of empty orders, goods in the transaction record are detected, and non-relevant fields are removed. In this way, the data is simplified.
- (2) User interest analysis module: This module is mainly used to model and update the interest model. It is automatically collected by the front end and stored in the corresponding database. A "user-interest item" chain is constructed using the above calculation method. Search for items with high correlation and apply them to the TOP-N model.
- (3) Establishment of association rule database model: the CTE-MARM method proposed above is used to discover the connections between specific levels in specific application scenarios. Set the restriction level k to 2. The rules stored in the transaction table are classified from different perspectives according to credibility.

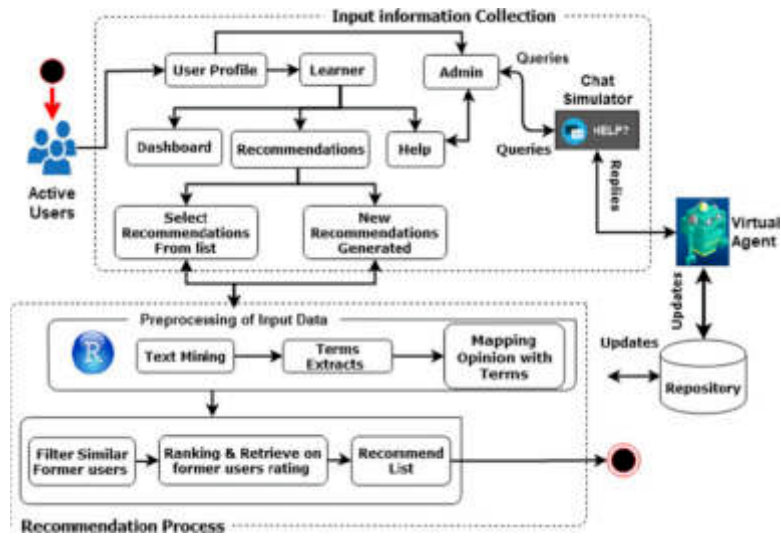


Fig. 5.1: Overall architecture diagram of the system.

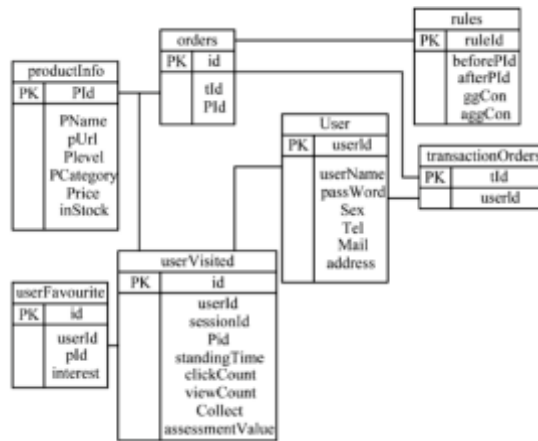


Fig. 5.2: Database E-R diagram.

(4) TOP-N network recommendation mode: The page is visually displayed through the association rules of the products that correlate more with the user’s interest mode.

5.2. Database Design. According to the characteristics of e-commerce product recommendation, a database E-R model based on user behavior and product recommendation is proposed. The database diagram for E-R is shown in Figure 5.2.

Below are the main database tables.

1. The user information form includes user ID, registration name, password, gender, contact information, address, etc.
2. The item information form includes item ID, name, link, grade, classification, price, inventory, etc.
3. The transaction table contains the transaction ID and user ID.
4. The purchase record form contains the unique identification ID, transaction processing ID and item ID.
5. The user behavior table contains unique identifiers, user identifiers, item identifiers, number of clicks,

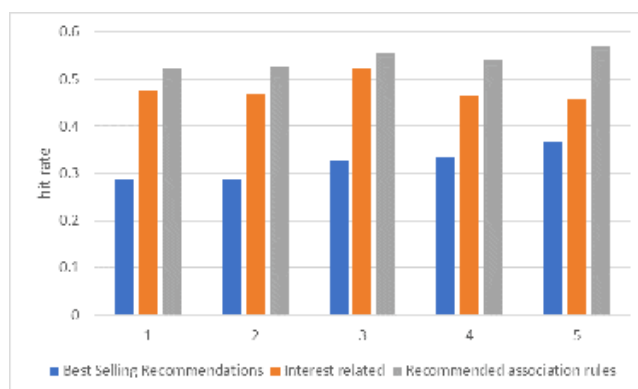


Fig. 6.1: Hit ratio analysis of manual analysis and system algorithm results.

repeat visits, browsing time, and whether to collect and evaluate scores.

6. The association rule table contains a unique identification, pre-rule component identification, post-rule component identification, inter-item confidence, classification and inter-item confidence.
7. The user interest table contains unique identifiers, user identifiers, product identifiers and interest levels.

6. System testing and verification. The shopping data of 3000 users are used as experimental data to test the accuracy of the established theory and products. The results of manual statistics are compared with user habits and system operation results of manual analysis. Cross-merge was performed after each trial. There were five trials. The result is shown in Figure 6.1. The study found that the accuracy of manual analysis of the best-selling items reached 31.8%. The interest recommendation accuracy of manual analysis reached 51.1%, while the recommendation accuracy of the e-commerce product recommendation system based on association rules reached 55.8%, which is more efficient and accurate than the method. This achieves the original design goal.

7. Conclusion. This paper uses the method based on CTE-MARM to establish an e-commerce data analysis method based on user interests and association rules. In this way, TOP-N products are recommended for users. However, the multi-level association discovery method research needs to be further explored. More influencing factors and technical means must be considered in improving the hit rate of product recommendations.

8. Acknowledgments. The work was supported by 1. Industry and University Collaborative Education Project of the Ministry of Education in 2024: Construction of University about E-commerce Big-Data Practice Base From the perspective of intelligent new media (No. 231100273283538). 2. Industry and University Collaborative Education Project of the Ministry of Education in 2022: Construction of University about E-commerce Big-Data Practice Base from the Perspective of New Media (No. 220606030204538).

REFERENCES

- [1] Dogan, O., Kem, F. C., & Oztaysi, B. (2022). Fuzzy association rule mining approach to identify e-commerce product association considering sales amount. *Complex & Intelligent Systems*, 8(2), 1551-1560.
- [2] Yang, W., & Lin, Y. (2021). Research on the interactive operations research model of e-commerce tourism resources business based on big data and circular economy concept. *Journal of Enterprise Information Management*, 35(4/5), 1348-1373.
- [3] Zong, K., Yuan, Y., Montenegro-Marin, C. E., & Kadry, S. N. (2021). Or-based intelligent decision support system for e-commerce. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(4), 1150-1164.
- [4] Chawla, N., & Kumar, B. (2022). E-commerce and consumer protection in India: the emerging trend. *Journal of Business Ethics*, 180(2), 581-604.
- [5] Abendin, S., & Duan, P. (2021). Global E-commerce talks at the WTO: Positions on selected issues of the United States, European Union, China, and Japan. *World Trade Review*, 20(5), 707-724.
- [6] Ayob, A. H. (2021). E-commerce adoption in ASEAN: who and where? *Future business journal*, 7(1), 1-11.

- [7] Hasana, Z., & Afifah, I. I. (2023). Perlindungan hukum terhadap konsumen dalam transaksi e-commerce. *Advanced In Social Humanities Research*, 1(5), 795-807.
- [8] Karn, A. L., Karna, R. K., Kondamudi, B. R., Bagale, G., Pustokhin, D. A., Pustokhina, I. V., & Sengan, S. (2023). Customer centric hybrid recommendation system for E-Commerce applications by integrating hybrid sentiment analysis. *Electronic Commerce Research*, 23(1), 279-314.
- [9] Bawack, R. E., Wamba, S. F., Carillo, K. D. A., & Akter, S. (2022). Artificial intelligence in E-Commerce: a bibliometric study and literature review. *Electronic markets*, 32(1), 297-338.
- [10] Aditantri, R., Mahliza, F., & Wibisono, A. D. (2021). Urban planning and e-commerce: Understanding the impact during pandemic covid-19 in Jakarta. *International Journal of Business, Economics, and Social Development*, 2(3), 135-142.
- [11] Belwal, R., Al Shibli, R., & Belwal, S. (2021). Consumer protection and electronic commerce in the Sultanate of Oman. *Journal of Information, Communication and Ethics in Society*, 19(1), 38-60.
- [12] Peráček, T. (2022). E-commerce and its limits in the context of the consumer protection: The case of the Slovak Republic. *Tribuna Juridică*, 12(1), 35-50.
- [13] Tran, D. T., & Huh, J. H. (2022). Building a model to exploit association rules and analyze purchasing behavior based on rough set theory. *The Journal of Supercomputing*, 78(8), 11051-11091.
- [14] Ünvan, Y. A. (2021). Market basket analysis with association rules. *Communications in Statistics-Theory and Methods*, 50(7), 1615-1628.
- [15] Cui, H., Niu, S., Li, K., Shi, C., Shao, S., & Gao, Z. (2021). A k-means++ based user classification method for social e-commerce. *Intell. Autom. Soft Comput*, 28(1), 277-291.

Edited by: Zhigao Zheng

Special issue on: Graph Powered Big Aerospace Data Processing

Received: Dec 16, 2023

Accepted: Jan 9, 2024



HIGH-RESOLUTION HOLOGRAPHIC IMAGE RECONSTRUCTION BASED ON DEEP LEARNING

FANGJU LI*

Abstract. Aiming at the shortcomings of existing holographic reconstruction algorithms, which are complex and easily affected by noise, this project proposes a semantic partitioning U-Net for high-resolution reconstruction. Firstly, a method based on the edge-neural network is proposed to obtain more image semantic information and improve the model training effect. Secondly, the effective channel attention mechanism of deep neural networks is introduced to enhance the attention to the detailed information in the holographic image. This further improves the accuracy of the neural network. The convergence rate of the neural network is accelerated by introducing a linear element with leakage correction. Experimental results show that this method can quickly reconstruct phase and brightness images with better detail, better edge texture and flat background. Holographic images of various sizes can be reproduced. The research of this project will lay a foundation for applying holographic image enhancement technology based on deep learning.

Key words: Holography; High-resolution reconstruction; Channel attention mechanism; Terminal neural network; multi-scale reconstruction

1. Introduction. The numerical reconstruction of holograms is a very crucial research work. Firstly, it needs to transform the hologram, filter it +1 times, carry out operations such as unwrapping and phase compensation, and finally realize the reconstruction of the hologram. However, conventional detection technology is more complicated and easily disturbed by noise. So, it is necessary to find a fast and effective reconstruction algorithm. In recent years, with the rapid development of neural networks and corresponding algorithms, it has played a vital role in many industries such as industry, agriculture, medicine and so on with its convenient, fast and efficient characteristics, and it also dramatically promotes the development of three-dimensional image reconstruction. It is imperative to construct an end-to-end coaxial holographic reconstruction network that can resist various optical path aberrations and has no restriction on the wavefront of the reference beam. High-resolution holographic image reconstruction with deep learning is based on deep neural networks, which can reconstruct low-resolution holographic images into high-resolution images. This method takes advantage of the powerful capabilities of convolutional neural networks (CNN) and recurrent neural networks (RNN) in deep learning to learn the features and structure of images, thereby achieving the reconstruction of high-resolution holographic images. Specifically, this approach typically includes the following steps:

1. Data preprocessing: Convert the original low-resolution holographic image into a format suitable for input into the deep learning model and perform necessary preprocessing, such as denoising, normalization, etc.
2. Feature extraction: Use the convolutional neural network (CNN) in the deep learning model to extract features from the preprocessed holographic image to extract high-level features of the image.
3. Super-resolution reconstruction: Use the deep learning model's recurrent neural network (RNN) to perform super-resolution reconstruction on the extracted high-level features to generate high-resolution holographic images.
4. Post-processing: Perform necessary post-processing on the generated high-resolution holographic image, such as denoising, sharpening, etc., to improve the quality and visibility of the image. The high-resolution holographic image reconstruction method of deep learning is a complex technology that requires a large amount of computing resources and training data. It also requires understanding and mastery of deep

*Department of Physics, School of Physics and Electrical Engineering, Weinan Normal University, Weinan, Shaanxi, 714099, China (antyl8366@163.com)

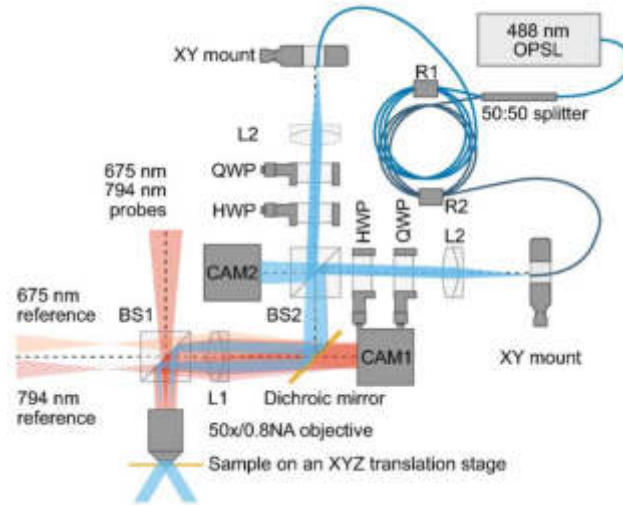


Fig. 2.1: Principle of digital holographic recording.

learning models and algorithms.

Reference [1] describes a lightweight one-to-two network, which can reproduce both light intensity and bit phase and is applied to digital image reconstruction. In literature [2], U-shaped networks were used to reconstruct three-dimensional images, and the images were not affected by DC, twins and other factors. Reference [3] introduced the attention mechanism into U-Net to achieve high-resolution cell reconstruction. In reference [4], a multi-scale complete convolutional fusion network was constructed to reconstruct the three-dimensional phase diagram of a single cell. Literature has some problems [5], such as low reconstruction rate, complex reconstructed scenes and unclear boundaries. The existing studies only conducted simulation experiments, lacking the real image test. This project will combine the end-to-end neural network with adequate channel attention (ECA) based on U-Net architecture to improve reconstruction efficiency. This project proposes an end-to-end neural network model based on U-Net to improve reconstruction efficiency further. Adequate channel attention is introduced into the feature extraction layer. The method of leakage correction was used to improve the linear cell activation function of the convolutional hidden layer. Then, an efficient and high-quality reconstruction method of multi-scale hologram images based on skeleton structure is proposed.

2. Principles and methods.

2.1. Digital holographic wavefront recording and traditional reconstruction. The reference light is coherently superimposed utilizing object-light wave transmission on the hologram. The interference fringe pattern collected by CCD and a digital hologram are obtained. The principle of digital holographic recording is shown in Figure 2.1 (the picture is quoted in the literature [6]).

Assuming that the density of the object surface is $P(u_0, v_0)$, the complex amplitude of the object wave reaching the holographic recording surface after coherent light irradiation is obtained by the method of angular spectral diffraction.

$$P(u, v) = G^{-1} \left\{ G [P(u_0, v_0)] \exp \left[k \frac{2\pi}{\eta} s \sqrt{1 - (\eta g_{x_0})^2 - (\eta g_{y_0})^2} \right] \right\}$$

G is the Fourier transform. G^{-1} inverse Fourier inversion. s is the distance from the surface of the object to the projection plane of the hologram. $P(u, v) = C(u, v) \exp[k\eta(u, v)]$. It is assumed that $C(u, v)$ represents the light wave amplitude of the object and $\eta(u, v)$ represents the phase of the light wave of the object [7]. The complex amplitude of the reference light on the hologram is called $D(u, v) = C_d(u, v) \exp[k\eta_d(u, v)]$. Where $C_d(u, v)$

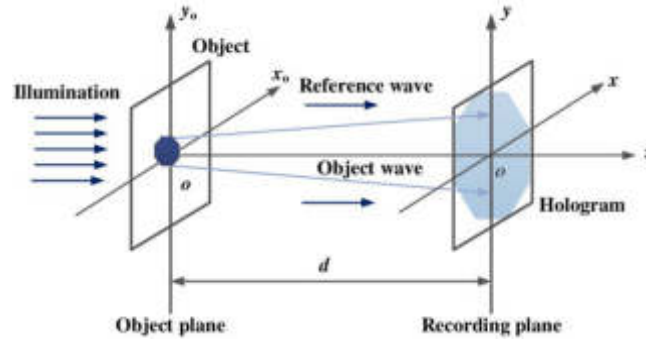


Fig. 2.2: Schematic diagram of digital holographic reconstruction.

represents the amplitude of the reference light wave and $\eta_d(u, v)$ represents the phase. Then, the distribution of light intensity on the recording surface of digital holography can be expressed as:

$$\begin{aligned} W(u, v) &= |P(u, v) + D(u, v)|^2 = \\ &|P(u, v)|^2 + |D(u, v)|^2 + \\ &P(u, v)D^*(u, v) + P^*(u, v)D(u, v) \end{aligned}$$

$|P(u, v)|^2$ represents information about the light intensity of the object and $|D(u, v)|^2$ represents information about the intensity of the reference light. The third project is to use conjugated light to regulate matter waves. The fourth project is the study of conjugated optical properties of reference light sources [8]. Therefore, the measured object's amplitude and bit equality information are recorded in the digital hologram. The reproduction principle of traditional digital holography is shown in Figure 2.2. The reconstruction of the image is realized through the steps of illumination and zero-order interference, twin image elimination and phase unwinding using regenerative light. The above processes are not included in reconstructing digital holograms using deep neural networks, so they will not be described here.

Assuming that the copied light wave is A vertically irradiated plane wave $H(u, v)$, then after irradiated digital holography, the diffraction wave generated can be expressed as:

$$\begin{aligned} W'(u, v) &= H(u, v)W(u, v) = H(u, v) [|D(u, v)|^2 + |P(u, v)|^2] + \\ &H(u, v)D^*(u, v)P(u, v) + H(u, v)D(u, v)P^*(u, v) \end{aligned}$$

From the diffraction equation of the angular spectrum, the light field of the reproducing object can be obtained:

$$P(u_i, v_i) = G^{-1} \left\{ G [W'(u, v)] \exp \left[k \frac{2\pi}{\eta} s \sqrt{1 - (\eta g_u)^2 - (\eta g_v)^2} \right] \right\}$$

2.2. Neural network structure. The experiment was divided into two parts. The network structure used in the first experiment is shown in Figure 2.3. Where $U \times Q$ is the proportion of the input image. $U/2 \times Q/2$ represents that the digital holographic image on the $U \times Q$ scale is 1/2 of the original scale after being maximized. The rectangular frame is the characteristic graph obtained after processing each operation module [9]. The number of channels on the current feature map is displayed above the rectangular box. Here, the input and output represented by 1 are two gray images. $Z1$ means that the number of channels in the current characteristic mapping is not fixed and is given according to specific test requirements. $2Z1, 4Z1$, etc., represent the current number of characteristic graph channels is 2 times, 4 times, and $Z1$ so on. The network architecture comprises five operation modules: hierarchical segmentation block, down sampling block, up sampling block, jump link block, output convolutional layer and so on. See Figure 2.3 for a detailed description of each module.

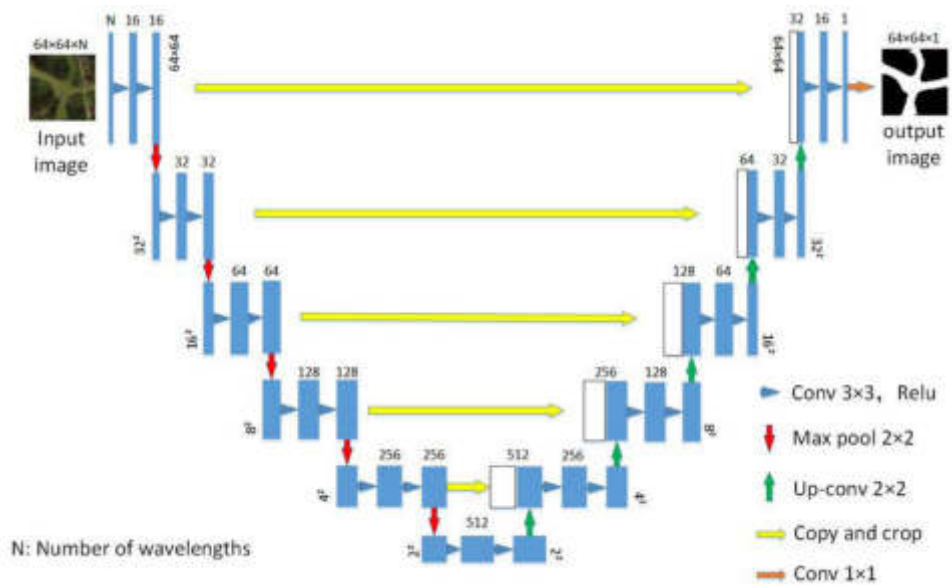


Fig. 2.3: Improved U-Net structure.

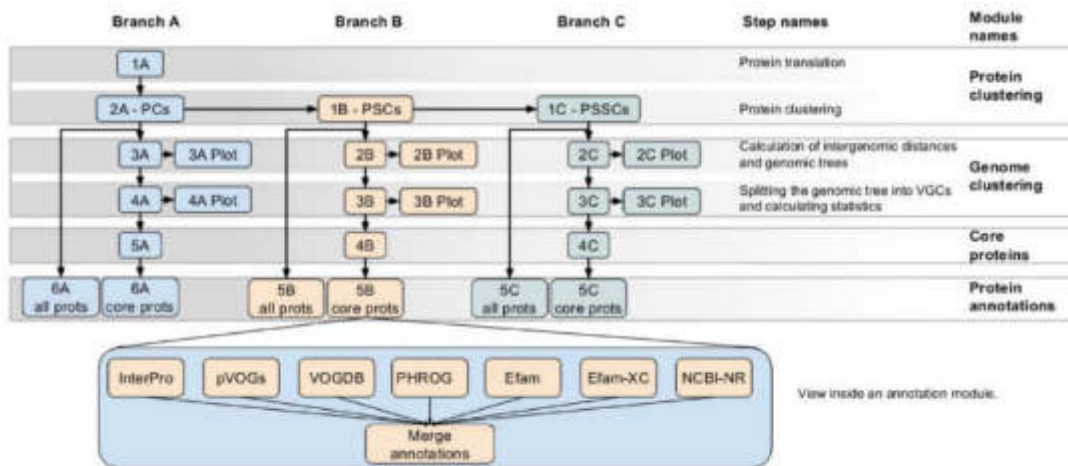


Fig. 2.4: HS-Block structure.

2.2.1. Dividing blocks by layers. This construct is shown in Figure 2.4. The core idea is to use a 3x3 convolution kernel to extract the feature map, perform Batch Norm and ReLU operations, and divide it into five channels. The output of the first set is fed directly to the next level. The second set of images is also extracted by a 3x3 convolutional network with a convolution kernel, and these images are evenly divided into two sub-sequences [10]. The first sub-sequence is connected with the second sub-sequence, and finally, the depth features are further extracted by convolution and the average segmentation is carried out until the fifth set of images is completed. Finally, the output characteristic diagram is integrated and used as the input of the next layer. The advantage of this method is that it reduces the redundant features based on reducing the network structure and improves the operational efficiency and speed of the neural network.

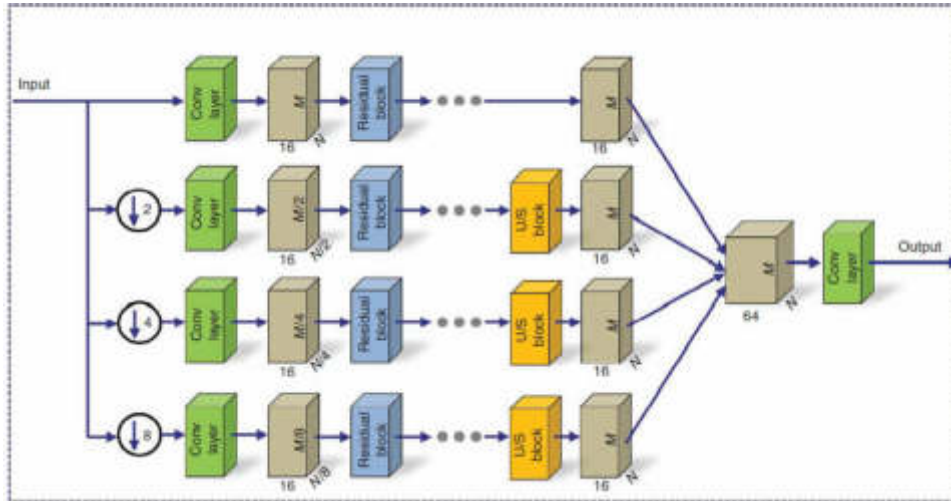


Fig. 2.5: A network of reconfigurable multi-scale digital holograms.

2.2.2. Pooled core subsampling. The lower sampling used the largest pool size, a 2x2 pool. Its primary purpose is to reduce the scale of feature mapping, maintain helpful information, increase the perception field, and, to some extent, prevent overfitting and reduce the difficulty of solving the network structure.

2.2.3. Using Nearest Neighbor Padding. The goal is to decode step by step to restore the original image size. After up-sampling the image, the size of the image will be doubled.

2.2.4. Long Jump Connection. This operation connects the feature graph channels of the output of the shallow network and the deep network together, and the result is used as the input of the next layer so that the deeper network can retain useful shallow information while avoiding gradient disappearance and improving network performance [11].

2.2.5. Output convolution layer. This is simply a combination of convolution and activation layers used to produce a single-channel image.

The network structure used in the second experiment is represented in Figure 2.5. The U-Net in this architecture is represented by the dotted boxes in the Figure 2.3 architecture and otherwise has the same functionality as the corresponding modules in the Figure 2.3 architecture [12]. The numbers above the matrix box represent the channels in the current feature graph. One represents the input and the output as grayscale images, and 40, 20, and 20 represent the number of channels. They also indicate that C1 is 40, 20, and 20 in the modified U-Net architecture. There are three levels from top to bottom [13]. The first level is a 256x256 three-dimensional image. Fix the current model parameters when done. A digital holographic image with a size of 512x512 is passed through the two front and back network structures. Only the second network is learned, and then the parameters of the second network are fixed. The three-level network architecture is added so that every network can pass 640x480 digital holographic, while the three-layer network only needs to learn the network parameters [14]. Since the following layer network can use the local characteristic map extracted from the previous layer network, the number of all network parameters can be reduced. This network structure can effectively solve the problem of multiple extractions of the same image, thus reducing the complexity of modeling and realizing the reconstruction of images of different sizes.

2.3. Generation of data sets and network training. Digital holograms are generated when a computer simulates digital holographic imaging. This project intends to scale up to 18,000 handwritten drawings in the MNIST standard library to obtain amplitude information [15]. The 18,000 images of Chinese characters are normalized and multiplied by 1.9π according to A particular proportion, that is, the wavelength of the illuminated object is $\eta = 632.8$ nm, the distance between the illuminated object and the recording surface of

Table 3.1: Simulation Environment Settings.

Name	Settings
CPU	Intel265GHz4 core
RAM	16G
SDD	500
OS	Win10PersonalEdition

the hologram is $s = 0.3086m$, and the size of the recording surface of the hologram is 5 mm. The reference light is parallel. The Angle of reference light with the x-axis and the Y-axis is set as $\pi/2.02$ and $\pi/2$, and the sampling quantity is 256×256 , 512×512 , 640×480 . The complex amplitude distribution of object wave on the digital holographic plane is obtained by formula (1), and then it coherently overlaps with the reference wave to obtain 6000 digital holograms of different sizes. Five thousand cases were selected as training and 1000 as test samples [16]. The training loss function uses the mean square error loss function, which represents the sum of squares of errors at the corresponding points of the predicted and original data. The expression is

$$\text{lose}_{MSE} = \frac{1}{n} \sum_{i=1}^n (v_i - \bar{v}_i)^2$$

n represents the size of the photo. v_i represents the amplitude information or the i value of the phase information after reconstruction; \bar{v}_i represents i values associated with the initial data [17]. The weights are optimized by the Adam optimization method. Measures were taken to end the training session early when mistakes were made to avoid over-learning. A deep neural network learning framework based on NvidiaRTX2080Ti is constructed by using Pytorch technology.

3. Case analysis.

3.1. Environment Settings. This project intends to use a convolutional neural network and BP neural network for detection [18]. The above algorithms are compared under the experimental conditions given in Table 3.1.

3.2. Experimental object of image reconstruction in laser holography. Five typical three-dimensional images are selected for experiments to verify the universality of the algorithm [19]. Fifty images were selected as simulation tests for different types of laser holograms. The representative image is shown in Figure 3.1.

3.3. Study on image reconstruction in laser holography. The reconstruction results of various stereo images obtained by the three measurement methods are compared, and the results are shown in Figure 3.2. The results show that the accuracy of laser hologram image reconstruction based on deep learning neural networks is better than 95%. The advantages of laser holographic image reconstruction based on deep learning neural network theory are proven.

3.4. Comparison of image reconstruction effects of laser holography. The reconstruction speed in existing large three-dimensional images is also a significant problem. According to the analysis of Figure 3.3, it can be seen that the reconstruction of the 3D stereo image reconstructed by the neural network method based on deep learning takes about 20 milliseconds, the 3D reconstruction based on the BP network takes about 30 milliseconds, and the reconstruction of the convolutional neural network takes about 35 milliseconds.

4. Conclusion. A new idea of three-dimensional image reconstruction using a deep neural network is studied. Its main objective is to improve image reconstruction accuracy in laser holography. The deep neural network model is applied to the reconstruction of laser holographic images, improving the reconstruction effect of laser holographic images. This will help the processing of laser holograms in the future. This project will open up a new way to improve the image quality of laser holographic image reconstruction.



Fig. 3.1: *Experimental subjects of image reconstruction in laser holographic imaging.*

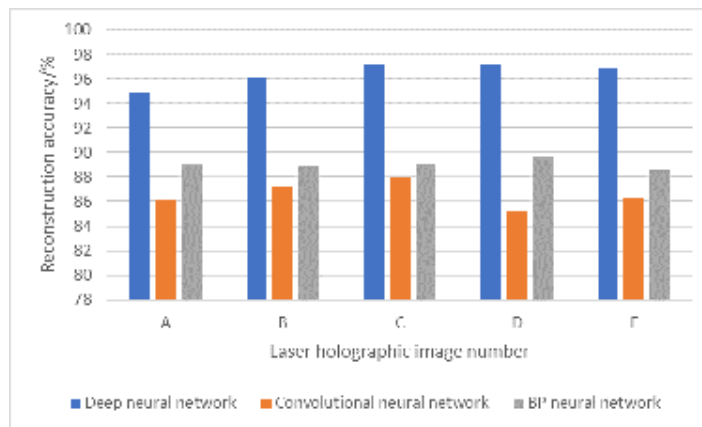


Fig. 3.2: *Image reconstruction accuracy of laser holographic imaging by different methods.*

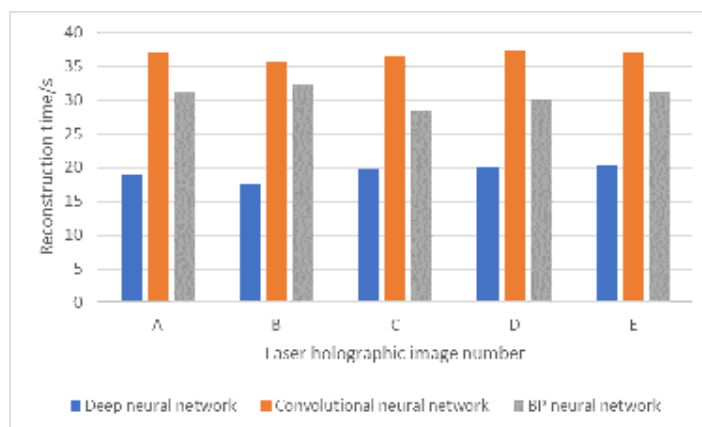


Fig. 3.3: *Image reconstruction time in different ways in laser holography.*

Acknowledgements. The work was supported by Weinan Science and Technology Bureau (WXQY002-011).

REFERENCES

- [1] Huang, L., Liu, T., Yang, X., Luo, Y., Rivenson, Y., & Ozcan, A. (2021). Holographic image reconstruction with phase recovery and autofocusing using recurrent neural networks. *ACS Photonics*, 8(6), 1763-1774.
- [2] Melanthota, S. K., Gopal, D., Chakrabarti, S., Kashyap, A. A., Radhakrishnan, R., & Mazumder, N. (2022). Deep learning-based image processing in optical microscopy. *Biophysical Reviews*, 14(2), 463-481.
- [3] Shi, L., Li, B., Kim, C., Kellnhofer, P., & Matusik, W. (2021). Towards real-time photorealistic 3D holography with deep neural networks. *Nature*, 591(7849), 234-239.
- [4] Pirone, D., Sirico, D., Miccio, L., Bianco, V., Mugnano, M., Ferraro, P., & Memmolo, P. (2022). Speeding up reconstruction of 3D tomograms in holographic flow cytometry via deep learning. *Lab on a Chip*, 22(4), 793-804.
- [5] Zeng, T., Zhu, Y., & Lam, E. Y. (2021). Deep learning for digital holography: a review. *Optics Express*, 29(24), 40572-40593.
- [6] Rekola H, Berdin A, Fedele C, et al. (2020). Digital holographic microscopy for real-time observation of surface-relief grating formation on azobenzene-containing films. *Scientific Reports*, 10(1), 19642.
- [7] Lee, M. H., Lew, H. M., Youn, S., Kim, T., & Hwang, J. Y. (2022). Deep learning-based framework for fast and accurate acoustic hologram generation. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 69(12), 3353-3366.
- [8] Rahmani, B., Oguz, I., Tegin, U., Hsieh, J. L., Psaltis, D., & Moser, C. (2022). Learning to image and compute with multimode optical fibers. *Nanophotonics*, 11(6), 1071-1082.
- [9] Kim, W., Park, B. S., Kim, J. K., Oh, K. J., Kim, J. W., Kim, D. W., & Seo, Y. H. (2020). Deep learning-based super resolution for phase-only holograms. *Journal of Broadcast Engineering*, 25(6), 935-943.
- [10] Gao, P., & Yuan, C. (2022). Resolution enhancement of digital holographic microscopy via synthetic aperture: a review. *Light: Advanced Manufacturing*, 3(1), 105-120.
- [11] MacNeil, L., Missan, S., Luo, J., Trappenberg, T., & LaRoche, J. (2021). Plankton classification with high-throughput submersible holographic microscopy and transfer learning. *BMC Ecology and Evolution*, 21(1), 1-11.
- [12] Potter, C. J., Hu, Y., Xiong, Z., Wang, J., & McLeod, E. (2022). Point-of-care SARS-CoV-2 sensing using lens-free imaging and a deep learning-assisted quantitative agglutination assay. *Lab on a Chip*, 22(19), 3744-3754.
- [13] Sun, J., Tárnok, A., & Su, X. (2020). Deep learning-based single-cell optical image studies. *Cytometry Part A*, 97(3), 226-240.
- [14] Huang, L., Chen, H., Liu, T., & Ozcan, A. (2023). Self-supervised learning of hologram reconstruction using physics consistency. *Nature Machine Intelligence*, 5(8), 895-907.
- [15] Sakib Rahman, M. S., & Ozcan, A. (2021). Computer-free, all-optical reconstruction of holograms using diffractive networks. *ACS Photonics*, 8(11), 3375-3384.
- [16] Xu, G., Jin, B., Yang, S., & Liu, P. (2023). Field recovery from digital inline holographic images of composite propellant combustion base on denoising diffusion model. *Optics Express*, 31(23), 38216-38227.
- [17] ONUR, T. Ö., & KAYA, G. U. (2022). Application of Binary Genetic Algorithm for Holographic Vascular Mimicking Phantom Reconstruction. *Balkan Journal of Electrical and Computer Engineering*, 10(1), 16-22.
- [18] Kang, I., De Cea, M., Xue, J., Li, Z., Barbastathis, G., & Ram, R. J. (2022). Simultaneous spectral recovery and CMOS micro-LED holography with an untrained deep neural network. *Optica*, 9(10), 1149-1155.
- [19] Meng, Z., Pedrini, G., Lv, X., Ma, J., Nie, S., & Yuan, C. (2021). DL-SI-DHM: A deep network generating the high-resolution phase and amplitude images from wide-field images. *Optics Express*, 29(13), 19247-19261.

Edited by: Zhigao Zheng

Special issue on: Graph Powered Big Aerospace Data Processing

Received: Dec 25, 2023

Accepted: Jan 15, 2024



OPTIMIZATION OF E-COMMERCE PRODUCT RECOMMENDATION ALGORITHM BASED ON USER BEHAVIOR

YIFAN JI*, LAN CHEN† AND RUI XIONG‡

Abstract. In order to implement personalized recommendation algorithms for e-commerce, the author proposes a genetic fuzzy algorithm based on user behavior to improve the sales, personalized recommendation, user satisfaction, and purchase matching performance of e-commerce. Collect data based on e-commerce personalized preference recommendation information, extract the associated feature quantities of personalized data for clustering processing, and then combine fuzzy B-means clustering method to achieve e-commerce personalized recommendation. According to the individual preferences of e-commerce, the collected data samples are fitted with differences and restructured, and a genetic evolution method is adopted for global optimization. The experimental results show that the optimized genetic fuzzy algorithm used in this method has improved stability and accuracy compared to the PSO method, with an accuracy increase of 4%. This proves that the algorithm can provide the services needed by users more quickly and is an effective means.

Key words: Genetic algorithm, Personalized recommendations, Fuzzy clustering, User behavior

1. Introduction. The internet generates a massive amount of information every day, some generated during purchases and some generated during page browsing. Behind this complex information lies the unique behavioral characteristics of each internet user. If this information is well utilized, better services can be provided more accurately for specific users [1]. With the skyrocketing amount of information on the Internet, methods for analyzing this information face severe challenges. Therefore, it is of great commercial value to analyze and predict user behavior through data mining and other related technologies, and directly recommend the predicted results to users, providing personalized information recommendation services. In recent years, after successful e-commerce models such as B2B and B2C, the O2O model has for the first time connected online virtual operations with offline physical stores, indicating that internet technology has further spread to people's daily lives. Multiple internet giants are actively participating in the O2O business. In July 2015, Baidu announced an investment of 20 billion yuan in the O2O field within three years, indicating its level of importance. In this context, the daily transaction volume of the Internet will increase, and e-commerce will be the most important market in the future [2].

However, the increasingly serious problem of information overload still troubles internet users. Selecting information that users are interested in from the ocean of data is like finding a needle in a haystack. At the same time, it is difficult to distinguish the accuracy and authenticity of information. Due to limitations in one's own abilities and concerns about expenses, a large amount of information actually brings confusion and confusion to users. In fact, what consumers want is not the amount of information, but the information tailored to their personal interests and hobbies. In recent years, internet companies have shifted their approach from passively providing information to actively guessing what information users need. Nowadays, whether using apps to listen to music or online shopping, there is a "guess what you like" section for personalized recommendations. The rapid development of e-commerce has prompted e-commerce websites to provide higher quality services and more professional information. Personalized services are an important research direction in the field of e-commerce. Currently, e-commerce platforms that only passively provide information services are no longer able to establish themselves. Only by actively attracting users and providing the goods and services they need can

*Sichuan Vocational College of Finance and Economics, Chengdu, Sichuan, 610101, China

†Sichuan Vocational College of Finance and Economics, Chengdu, Sichuan, 610101, China (Corresponding author's e-mail: chenlannncncnc@163.com)

‡Sichuan Vocational College of Finance and Economics, Chengdu, Sichuan, 610101, China

they attract users. Driven by this emerging business model with billions of users, new science and technology have been continuously developed, and recommendation systems have played a crucial role in it [3].

From a research perspective, the massive amount of data brought by the Internet has led to the emergence of a new research model, which has shifted from studying actual experimental objects to analyzing virtual data. In this model, the dependence on theory is not very sensitive, and there are some problems that cannot be accurately explained in theory, but have been proven effective in practical applications. This research method is a breakthrough attempt, it not only prompts the academic community to re-examine research methodology, but more importantly, it itself promotes the progress of current science and technology [4].

From a business perspective, accurate recommendation algorithms can better personalize marketing for users. When placing advertisements on a page, if recommendations are made based on the user's recent browsing data, it can improve the accuracy of advertising placement. On shopping websites, corresponding recommended products can also be provided based on users' recent purchasing behavior to increase sales. It can be seen that high-performance recommendation algorithms can generate significant economic benefits for both advertising and e-commerce businesses.

From the perspective of internet users, with the emergence of recommendation systems, we don't have to waste a lot of time finding things we are interested in from search engines or merchant lists. Instead, data providers can directly push recommendation results to us through recommendation systems on the page. This is not only an efficient way for users to obtain the required information, but also a pleasant user experience.

2. Methods.

2.1. E-commerce personalized recommendation information model and feature extraction.

(1) *Personalized information transmission.* Extract user personalized consumption data based on e-commerce correlation features, obtain fuzzy decision functions, construct non-linear mapping $\varnothing : n \in R^n \rightarrow Q$ to represent user personalized guidance space, combine data information with decision functions, and use intelligent algorithms to map to feature space F [5]. Assuming the e-commerce recommendation sample set is $\{(a_1, m_1), (a_2, m_2), \dots, (a_n, m_n)\}$, the personalized feature quantity $a_i \in R^n$ represents the model input vector, $m_i \in R^n$ represents the target test value, and n represents the quantity, the personalized recommendation objective function is calculated:

$$\begin{aligned} \min \text{imize } & \frac{1}{2} \|w\|^2 + \sum_{i=1}^n B (J_i + J_i^*) \\ \text{subject to } & m_1 - (w' \varnothing(a_i) + b) \leq \varepsilon - J_i \\ & (w' \varnothing(a_i) + b) - m_i \leq \varepsilon - J_i \\ & J_i J_i^* \geq 0, i = 1, 2, \dots, n; B > 0 \end{aligned} \quad (2.1)$$

In Equation 2.1, J_i and J_i^* are ontology attributes and association rule variables, and the cost factor is represented by B . The difference function is obtained by using control optimization method as shown in Equation 2.2.

$$Q(a) = \sum_{i=(e_i - e_i^*)}^n F(a_i, a_j) + b \quad (2.2)$$

In Equation 2.2, e_i and e_i^* represent personalized attribute values and the number of Template categories, while the symmetric kernel function $F(a_i, a_j)$ represents the recommendation threshold. Based on the preference information in the database Web cloud, adaptive optimization is performed to extract the information multipath gradient map and obtain the conduction model a_1, a_2, a_3, a_4 , which is represented as:

$$\begin{cases} a_1 = L_1 - u \\ a_2 = L_2 - u \\ a_3 = L_3 - u \\ a_4 = u \end{cases} \quad (2.3)$$

In Equation 2.3, u represents the genetic fuzzy clustering correlation attribute, and the domain of information subspace dimension is Ω . Extract associated data and feature quantities based on e-commerce information, establish labels, and analyze group interaction relationships.

(2) *Extraction of associated feature quantities.* The personalized preference data is fitted with sample differences and restructured to extract associated feature quantities, $x(s)$, $s = 0, 1, \dots, n-1$, representing the time series of feature quantities in the user area. Under the constraints of relevant rules, the structural distribution function obtained is:

$$E^{bu}(b_1, b_2) = T \cdot \text{Length}(b) + m \cdot \text{Area}(\text{inside}(b)) + P_1 \int \text{inside}(b) |i - b_1|^2 + P_2 \int \text{outside}(b) |i - b_2|^2 \text{dadm} \quad (2.4)$$

In Equation 2.4, b_1 and b_2 are the adaptive feature coefficients, $\text{Length}(b)$ is the length coefficient, and $\text{Area}(\text{inside}(b))$ is the size of the feature region; Construct a feature vector set and control decision function, with a mixed kernel function as shown in Equation 2.5.

$$F_{min} = \mu F_{poly} + (1 - \mu) F_{rbf}, \mu \in (0, 1) \quad (2.5)$$

In Equation 2.5 $F_{poly} = [(a \cdot a_i + 1)]^2$, the preference trust kernel function $F_{RBQ} = \exp(-a||a - a_i||^2)$ is RBQ, and the adaptive clustering process is recommended under the rule, resulting in:

$$Q_{lg-c}(o) = (Q_{lg}(o), Q_{lg-a}(o), Q_{lg-m}(o)) = (Q_{lg}(o), h_a^*(o)) \quad (2.6)$$

In Equation 2.6, $Q_{lg}(o)$ is the user's project rating value, and a quadruple is obtained based on the associated features:

$$\max \left\{ \begin{array}{l} |bh(z) - bh(z) \cap bh(z_2)| + |bh(z) \cap bh(z_2)| \\ |bh(z_2) - bh(z) \cap bh(z_2)| + |bh(z) \cap bh(z_2)| \end{array} \right\} = \max \left\{ \begin{array}{l} bh(z) \\ bh(z_2) \end{array} \right\} \leq \Delta \quad (2.7)$$

In Equation 2.7, $bh(z)$ is a regular coefficient with correlation, which facilitates the extraction of data correlation features.

2.2. Optimization of personalized recommendation algorithms for e-commerce.

(1) *Genetic evolution optimization method.* Personalized data association feature extraction is based on sample difference fitting and structural restructuring. The author uses genetic fuzzy clustering method to extract association features based on kernel function construction. Utilizing the advantages of genetic algorithms, global optimization control is performed on potential user preference variables. The following Equation 2.8 expresses the meaning of the genetic evolutionary control function:

$$m_i = T m_i (1 - m_i) \quad (2.8)$$

In Equation 2.8, T is a personalized recommendation control parameter, and the construction of $m_i \in [0, 1]$ random numbers is completed. Assuming that in a fuzzy clustering space W , u is the mutated individual, $t = \{L_1, L_2, \dots, L_w\}$ is the population, $L_i^d(s) (i = 1, 2, \dots, w)$ represents the user's search for individual i using latent features in the dimension space W , $V_i^w(s) (i = 1, 2, \dots, t)$ is the optimization speed of individual i , $L_{best}^w(s)$ represents the best position of individual i , and $G_{best}^w(s)$ represents the optimal solution, in the process of genetic evolution, the expression for optimizing individual extremum and global extremum at each iteration is:

$$\begin{cases} N_i^w(s+1) = A \cdot N_i^w(s) + B_1 \cdot R_1(L_{best}^w(s)) \\ -L_i^w(s) + B_2 \cdot R_2 \cdot (G_{best}^w(s) - L_i^w(s)) \\ L_i^w(s+1) = L_i^w(s) + N_i^w(s+1) \end{cases} \quad (2.9)$$

In Equation 2.9, the conduction coefficient and correlation eigenvectors of particle i at the current and next time points are $N_i^w(s)$, $N_i^w(s+1)$, and $L_i^w(s)$, $L_i^w(s+1)$; In the formula, B_1 and B_2 represent learning factors,

with values between -25 and 25; The search radius and threshold of genetic fuzzy clustering are represented by R_1 and R_2 , respectively, and are randomly selected[6]. Combining genetic fuzzy clustering optimization with A, a recommended adjustment formula is obtained in interval $[A_{min}, A_{max}]$:

$$A(s + 1) = 4.0w(s)(1 - A(s)) \tag{2.10}$$

$$A(s) = A_{min} + (A_{max} - A_{min})A(s) \tag{2.11}$$

The personal universality analysis of e-commerce users is based on the value of the inertia factor in Equation 2.11 and the convergence control of the pattern neural network recommendation process.

(2) *Data information mining and clustering processing.* Cluster the obtained personalized preference data and combine it with B-means to extract quadruples represented by $\{E_1, E_2, \dots, E_p\}$. Extract the conduction control model under the premise of controlling constraint variables:

$$\gamma'_{desira} = \gamma_1 \cdot \frac{Density_i}{\sum_i Density_i} + \gamma_2 \frac{YL_i}{YL_{init}} \tag{2.12}$$

$$\begin{cases} \gamma_1 + \gamma_2 = 1, \gamma_1, \gamma_2 \in [0, 1] \\ \gamma_2 = \frac{max_i(YL_i) - min_i(YL_i)}{YL_{init}} \end{cases} \tag{2.13}$$

Introducing radii R_1 and R_2 into cluster learning, the recommendation process update Equation 2.14 is obtained, with $R_i(s) \in (0, 1), i = 1, 2$ in Equation 2.14. The phenomenon of using fuzzy clustering method to jump out of the optimal value and merge with user rating measurement is described by Equation 2.15:

$$R_i(s + 1) = 4.0R_i(s)(1 - R_i(s)) \tag{2.14}$$

$$N_i^w(s + 1) = 4.0N_i^w(s)(1 - N_i^w(s)) \tag{2.15}$$

$$N_i^w(s) = N_{min} + (N_{max} - N_{min})N_i^w(s) \tag{2.16}$$

$$\rho_i(f, l) = \frac{\gamma(f)\sigma_{fl}d_l(o_{s+1})\mu_{s+1}(l)}{\sum_{f+1}^V \sum_{l+1}^V \gamma_s(f)\sigma_{fl}d_l(o_{s+1})\mu_{s+1}(l)} \tag{2.17}$$

The range of pheromone values for mutated individuals in Equation 2.16 is represented by $[N_{min} + N_{max}]$. Under the constraint of association rules, Equation 2.17 represents the clustering center of the association feature quantity, where $t + 1(j)$ is the number of nodes, $d_l(o_{s+1})$ is the coefficient point set, and y_{fl} is the score measurement information. The genetic evolution algorithm is used to calculate the variance δ^2 and determine whether $\delta^2 < H$ is valid. In order to achieve personalized recommendation in e-commerce, genetic evolution and optimization processing are combined with clustering algorithms to determine whether the convergence criteria are met [7].

2.3. Introduction to User Group Behavior Analysis Platform. The process of the product recommendation function platform, as shown in Figure 2.1, is mainly to provide users with personalized product recommendation services [8,9]. Users log in to the platform to browse products (the products they browse are basically of interest to the user), and obtain the parameters and prices of the products, at the same time, understanding the type of product, regardless of whether the end user has successfully purchased, the user transaction database will record information such as product number, product type, and matching. The platform is divided into the following three modules to present the user behavior data analysis process:

- 1) User behavior tracking module: By tracking the user's clicking, browsing, bookmarking, storing and other behaviors on the website, the product data and browsing customer data are transformed into user purchasing behavior operation data, and users are grouped for the first time through log data;

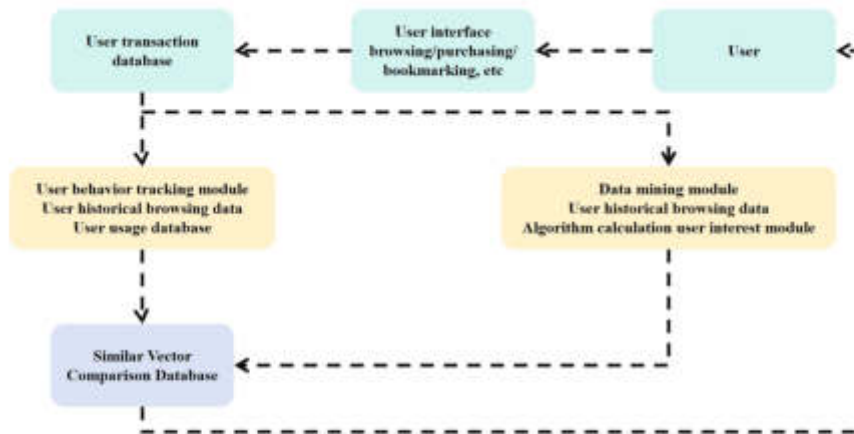


Fig. 2.1: Process of Product Recommendation Function Platform

- 2) Data mining module: Analyze user log data using traditional algorithms and improved fuzzy algorithms based on genetic algorithms, and obtain association rules for frequent items of users respectively. This can uncover the rule habits of users in their recent behavior;
- 3) Result display module: By using the user behavior tracking module to group user data, as well as the association rules of products analyzed by the data mining module, based on similarity vector comparison of user similarity, gather similar rule users, and make Top-N recommendations based on the similarity comparison results [10,11].

3. Experiments and Results. Using Matlab7 to design matrix simulation experiments, the data used in the experiments were all sourced from publicly available online data from JD.com and Alibaba, mainly to verify the accuracy and convergence of personalized e-commerce recommendations[12]. According to the simulation experiment parameters, $Q=41$ is set as the clustering center, the fuzzy control parameters are set to $b_1 = 124, b_2 = 364$, the e-commerce simulation time is 2 minutes (120 seconds), and 3120 iterations are run. In the W individual dimension space, the number is set to 1040, and the population size is 10. The experimental parameters are set to start the simulation analysis.

Test the sample information collected from e-commerce users and compare the convergence of traditional methods with the experimental process to obtain a comparison curve, as shown in Figure 3.1.

The comparison results of Figure 3.1 show that the author's use of e-commerce personalized recommendation has a very good response in terms of user recommendation satisfaction and information accuracy, which fully reflects the good performance of e-commerce personalized recommendation[13]. Analyzing and comparing personalized e-commerce recommendation methods with traditional methods, and comparing PCA algorithm with PSO, the results are shown in Figure 3.2.

The above experimental results indicate that in the personalized dimension space W of e-commerce personalized recommendation, the collection and setting of all experimental data, and the experimental process are compared with traditional methods. E-commerce personalized recommendation provides accurate user information recommendation, promotes transaction completion, and greatly improves efficiency[14]. It is confirmed that genetic evolution method is combined with B-means in global optimization, and preference data is clustered to achieve personalized recommendations for e-commerce. User satisfaction and accuracy are also reflected in Figures 3.1 and 3.2.

By collecting and analyzing a large amount of test data, three constructed methods were used for product recommendation, and the experimental data results shown in Figures 3.3 and 3.4 were obtained.

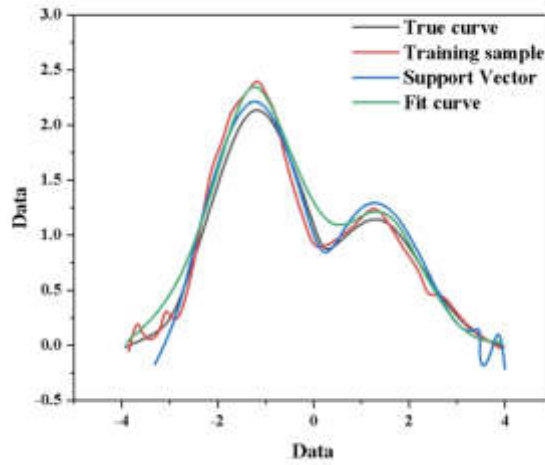


Fig. 3.1: Comparison of personalized recommendation curves for e-commerce

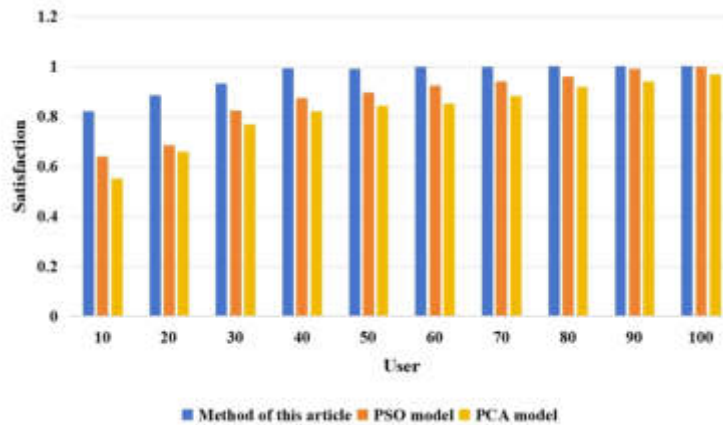


Fig. 3.2: Comparison of satisfaction with personalized e-commerce recommendations

Figure 3.3 shows the accuracy distribution of some users. It can be seen that in a small number of users, the accuracy difference between the PSO method and this method is not significant. However, as the number of users increases, this method demonstrates its advantages of high efficiency and accuracy [15].

Figure 3.4 shows the distribution of recall rates for some users. Similarly, it can be seen from the figure that when the number of users is relatively low, both accuracy and recall rates do not show the advantages of the algorithm. As the number of users increases, the improvement in system performance after algorithm optimization can gradually be seen. This also fully demonstrates that the results of big data analysis are more accurate and effective.

Finally, the author performed simple average calculations on a large amount of accuracy and recall data, and obtained experimental results as shown in Table 3.1 by calculating the time complexity of the three methods during operation.

Due to the fact that in the PCA model, the entire system does not display the function of product recommendation, but only conducts data analysis on the use and purchase of products, and presents the analysis

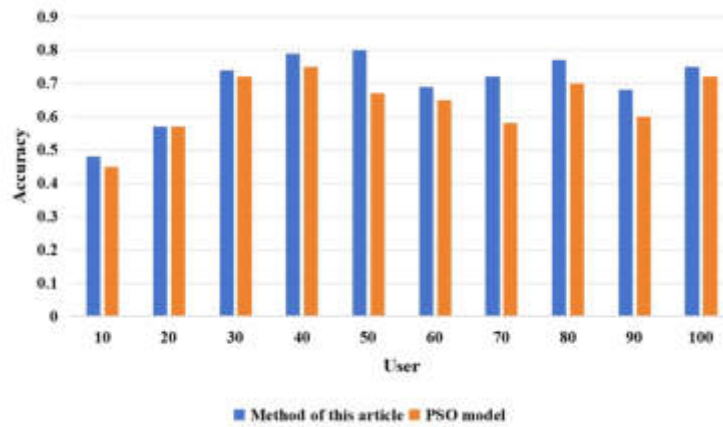


Fig. 3.3: Partial user usage accuracy chart

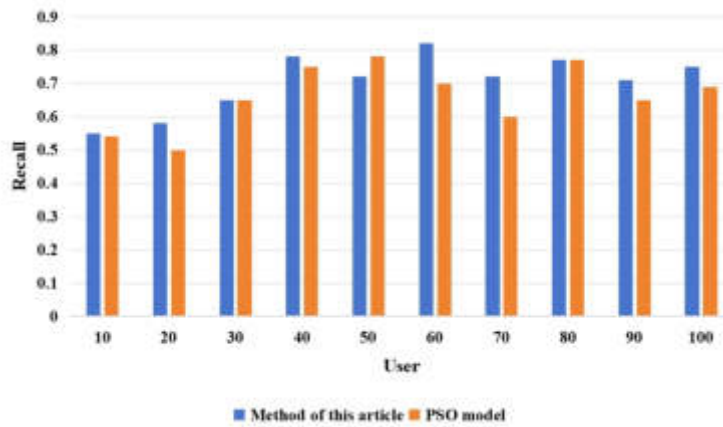


Fig. 3.4: Partial User Function Recall Rate Chart

results to users in the form of bar charts or pie charts, therefore, in the PCA model, users need to rely on these analysis data to determine the products that are suitable for their own configuration, thus unable to provide time complexity [16].

As can be seen, for the PSO method, the author uses traditional incremental mining algorithms to optimize the entire system, and the accuracy and recall of the product recommendation function have greatly increased. However, for the data platform Storm used in the system, traditional algorithms need to be compatible with its segmented data processing method and massive scanning times, which greatly increases the time and space required, the operating cost of the system has also generated tremendous pressure, therefore, the minimum time complexity of the matching algorithm is $O(n)$. For this method, a half search is used, so the time complexity of the matching algorithm is set to $O(1gn)$. Considering the cost, both types of time complexity are acceptable, but the system pursues higher efficiency. Moreover, Table 3.1 shows that the accuracy and recall of this method have also been improved to a certain extent compared to the PSO method. Therefore, considering all indicators comprehensively, this algorithm has indeed played an important role in improving product recommendation effectiveness [17,18,19,20].

From the above data analysis, it can be seen that the PSO method, using traditional incremental mining

Table 3.1: Comparison of evaluation indicators for three experimental methods

	PCA model	The method of this paper	PSO model
Accuracy	0.14	0.68	0.72
Recall	0.19	0.70	0.76
Time complexity	Not have	O(n)	O(lgn)

algorithms, has already achieved very good results in mining user preferences, and can more accurately and efficiently recommend products to users compared to before. And this method utilizes the optimized genetic fuzzy algorithm, which has improved stability and accuracy compared to the PSO method. More importantly, it greatly compresses the system's running time in terms of time cost, and can provide the services needed by users more quickly. It is an effective means.

4. Conclusion. The collection of personalized e-commerce data is sourced from publicly available data from JD.com and Alibaba. The author restructured the data and fitted the sample difference, then optimized and extracted the associated feature quantity, which was processed using the B-means clustering method to achieve personalized e-commerce recommendations. The author confirmed the effectiveness and stability of this recommendation method through experimental results. The effect of using weight increment mining is better than that of general mining, and it is also more efficient and fast. The accuracy of this experiment is as high as 72%, and the recall rate is as high as 76%, which is more accurate than other recommendation methods. Based on the above experimental methods, it can be proven that the algorithm optimized product recommendation method used by the author is an effective product recommendation strategy and has basically achieved the expected results.

Acknowledgement. This research was supported by the China West Normal University—Sichuan Provincial Department of Education Humanities and Social Sciences Research Base: Chengdu Chongqing Region Education and Economic and Social Collaborative Development Research Center (Grant No. CYJXF24061).

REFERENCES

- [1] Li, Y., Li, Y., Sun, B., & Chen, Y. (2023). Zinc ore supplier evaluation and recommendation method based on nonlinear adaptive online transfer learning. *Journal of Industrial and Management Optimization*, 19(1), 472-490.
- [2] El-Ashmawi, Walaa H.Ali, Ahmed F.Slowik, Adam. (2021). Hybrid crow search and uniform crossover algorithm-based clustering for top-n recommendation system. *Neural computing & applications*, 33(12),74-78.
- [3] Zhang, Z. (2021). An optimization model for logistics distribution network of cross-border e-commerce based on personalized recommendation algorithm. *Security and Communication Networks*, 2021(4), 1-11.
- [4] Ni, J. (2021). Predictive analysis of user behavior of e-commerce platform based on machine learning image algorithm in internet of things environment. *Journal of Intelligent and Fuzzy Systems*, 58(7),1-8.
- [5] Hussien, F. T. A., Rahma, A. M. S., & Abdulwahab, H. B. (2021). An e-commerce recommendation system based on dynamic analysis of customer behavior. *Sustainability*, 13(19), 10786.
- [6] Raja, D. R. K., Kumar, G. H., Basha, S. M., & Ahmed, S. T. (2022). Recommendations based on integrated matrix time decomposition and clustering optimization. *International Journal of Performability Engineering*,74(4), 18.
- [7] Zou, F., Chen, D., Xu, Q., Jiang, Z., & Kang, J. (2021). A two-stage personalized recommendation based on multi-objective teaching-learning-based optimization with decomposition. *Neurocomputing*, 452(6),96-102.
- [8] Peng, Z., Wan, D., Wang, A., Lu, X., & Pardalos, P. M. (2023). Deep learning-based recommendation method for top-k tasks in software crowdsourcing systems. *Journal of Industrial and Management Optimization*, 19(9), 6478-6499.
- [9] Yang, F., Wang, H., & Fu, J. (2021). Improvement of recommendation algorithm based on collaborative deep learning and its parallelization on spark. *Journal of Parallel and Distributed Computing*, 148(2), 58-68.
- [10] I. Glük, & Ozsoydan, F. B. (2021). Q-learning and hyper-heuristic based algorithm recommendation for changing environments. *Eng. Appl. Artif. Intell.*, 102(3), 104284.
- [11] Fang, Z., & Chiao, C. (2021). Research on prediction and recommendation of financial stocks based on k-means clustering algorithm optimization. *Journal of Computational Methods in Sciences and Engineering*,147(2), 1-9.
- [12] Zhong, D., Yang, G., Fan, J., Tian, B., & Zhang, Y. (2022). A service recommendation system based on rough multidimensional matrix in cloud-based environment. *Computer Standards and Interfaces*,(Aug), 82(34),65-69.
- [13] Yan, H. C., Wang, Z. R., Niu, J. Y., & Xue, T. (2022). Application of covering rough granular computing model in collaborative filtering recommendation algorithm optimization. *Advanced Engineering Informatics*, 51(54), 101485.

- [14] Ilker Glük, & Ozsoydan, F. B. (2021). Q-learning and hyper-heuristic based algorithm recommendation for changing environments. *Engineering Applications of Artificial Intelligence*, 102(104284),96.
- [15] Liu, S., Yang, Y., & Wang, Y. (2021). Integration of museum user behavior information based on wireless network. *Mobile Information Systems*, 2021(5), 1-8.
- [16] Zhang, H., Li, X., Kan, Z., Zhang, X., & Li, Z. (2021). Research on optimization of assembly line based on product scheduling and just-in-time feeding of parts. *Assembly Automation*,71(5), 41.
- [17] Sun, P., & Gu, L. (2021). Optimization of cross-border e-commerce logistics supervision system based on internet of things technology. *Complexity*,147(12),1-6.
- [18] Zhong, H., Wang, Y., & Yue, W. (2022). An e-commerce product recommendation method based on visual search and customer satisfaction. *International journal of knowledge and systems science*,96(5),52-53.
- [19] Sharma, S. N., & Sadagopan, P. (2021). Influence of conditional holoentropy-based feature selection on automatic recommendation system in e-commerce sector. *Journal of King Saud University - Computer and Information Sciences*,85(7),45-48.
- [20] Gupta, S., & Dave, M. (2021). Product recommendation system using tunicate swarm magnetic optimization algorithm-based black hole renyi entropy fuzzy clustering and k-nearest neighbour. *Journal of Information & Knowledge Management*,65(2),35-38.

Edited by: Zhigao Zheng

Special issue on: Graph Powered Big Aerospace Data Processing

Received: Dec 25, 2023

Accepted: Jan 23, 2024



APPLICATION OF MULTI-OBJECTIVE OPTIMIZATION ALGORITHM BASED ON ARTIFICIAL FISH SCHOOL ALGORITHM IN FINANCIAL INVESTMENT PORTFOLIO PROBLEMS

HONGXING ZHANG*

Abstract. In order to comprehensively measure these two indicators and make reasonable portfolio investment decisions, the author proposes using swarm intelligence optimization algorithm - artificial fish swarm algorithm to solve multi-objective investment portfolio problems, and has achieved good results. In order to verify the effectiveness and superiority of the artificial fish school algorithm, the author used MATLAB programming to conduct simulation experiments using AFSA algorithm and genetic algorithm (GA), and compared the results obtained. The results show that compared to the GA algorithm, the artificial fish school algorithm can obtain better investment portfolio decision-making solutions for investing in five types of assets, making investment returns as large as possible while minimizing risks, indicating the efficiency and superiority of the algorithm in solving multi-objective investment portfolio problems.

Key words: Artificial fish school algorithm, Multi objective optimization algorithm, Financial investment, Combinatorial problem

1. Introduction. With the continuous development of the market economy in today's society and the increasingly fierce social competition, the competitive field of asset investment has gradually penetrated into various industries and even people's daily financial life. However, when investing, there will inevitably be complex multi-objective investment portfolio problems, which cannot avoid complex calculations, therefore, in the face of the difficulty that existing mathematical programming methods cannot directly solve multi-objective investment portfolio problems, the author considers solving this problem from the perspective of highly efficient swarm intelligence optimization algorithms that have emerged and developed in recent years, which involves handing over a large amount of complex computational work to computers to complete. The so-called optimization algorithm basically describes search or rule based strategy or specific process, and gets the solution which can achieve user expectation by some search rules and channels. The theoretical research and practical application of optimization algorithms have made great progress, and many scholars have attempted to apply them to solve various engineering optimization problems. Many studies have shown that their problem-solving ability is higher than other traditional algorithms, and they have achieved good results, especially in combinatorial optimization problems. Therefore, people insist on innovative research on the theory and application of optimization algorithms. The author chooses a new type of swarm intelligence optimization algorithm - artificial fish swarm algorithm for research and applies it to solve practical problems [1]. The artificial swarm algorithm adopts a new method which is different from other optimization algorithms. There are some differences between the specific implementation of algorithms and the overall design idea, compared to the design method and the solution. However, it can also be combined with other traditional methods. AFSA does not require the initial values, measurement values, or properties of the target function. It has the potential to overcome local weaknesses and globalization. It is precisely because of the excellent characteristics that artificial fish swarm algorithm has attracted great interest and extensive attention from researchers from all walks of life at home and abroad since it has been applied. At present, the research and application of artificial fish swarm algorithms have entered many disciplines and been used to solve practical problems. This conclusion has become a research topic with high research value. So far, although the artificial fish school algorithm has been studied and applied by many scholars in related fields, its application scope and improvement work on the algorithm itself still need further improvement and expansion. Therefore, the author's research on using

*College of Finance, Henan Finance University, Zhengzhou, Henan, 451464, China (18538776139@163.com)

artificial fish school algorithms to solve multi-objective investment portfolio problems has important theoretical significance and practical application value for both the field of economics and computer science [2]. At the same time, it is also a preliminary exploration of the application of artificial fish swarm algorithm in the field of multi-objective investment portfolio.

Artificial Fish Swarm Algorithm (AFSA) is an intelligent swarm algorithm based on the behavior of fish game, which simulates the basic behavior of animals. Since its proposal, AFSA has received widespread attention from domestic and foreign researchers, as well as in-depth research and practical applications. So far, there have been hundreds of relevant references on the research and application of AFSA, indicating the significant influence of this algorithm both domestically and internationally. Related studies have shown that although the artificial fish swarm algorithm has many superior characteristics, it still has some shortcomings, such as low solving accuracy, being trapped in local extremum and low efficiency, and slow convergence speed in the later stage. In response to these shortcomings, scholars have adopted different design methods for exploratory research and improvement, making the algorithm more effective in solving practical problems.

In the process of capital market, rational investors will decide how to allocate some risk factors which they hold in an appropriate way, in order to achieve the goal of increasing risk as much as possible and obtaining the maximum funds. How to comprehensively measure the two main factors that constrain capital investment, namely investment risk and investment return, and make reasonable portfolio investment decisions, this also leads to a multi-objective investment portfolio problem. And this problem belongs to a complex nonlinear programming problem, which is difficult to be effectively solved by traditional algorithms. Therefore, the author proposes to use swarm intelligence optimization algorithm - artificial fish swarm algorithm to solve multi-objective investment portfolio problems, and has achieved good results [3].

2. Methods.

2.1. Multi-objective investment portfolio problem and its model. Assuming there are n assets S_i ($i=1, \dots, n$) available for investors to choose from in the market, with an existing amount of M of funds used for a period of investment. The average return on purchasing S_i during this investment period is r_i , and the expected risk loss rate is q_i . There is a certain level of risk in the investment process. Consider using the largest risk among n assets to measure overall risk. Assuming that a certain transaction fee needs to be paid when purchasing an asset, let the transaction rate paid for purchasing S_i be p_i . When the purchase amount does not exceed the given value u_i , the transaction fee will be calculated based on u_i ; Furthermore, assuming that the deposit interest rate deposited into the bank is fixed [4]. Establish the following multi-objective investment portfolio model, assuming that the proportion of funds for purchasing asset S_i to the total M is x_i , the required transaction cost is:

$$T(x_i) = \begin{cases} 0, & x_i = 0 \\ u_i p_i, & 0 < Mx_i \leq u_i \\ (Mx_i) p_i, & Mx_i > u_i \end{cases} \tag{2.1}$$

The total net income during an investment period is recorded as V :

$$V = \sum_{i=1}^n (Mx_i r_i - T(x_i)) \tag{2.2}$$

For an investment period, the overall risk (taking the maximum risk) is recorded as R :

$$R = \max_{1 \leq i \leq n} \{Mx_i q_i\} \tag{2.3}$$

In order to maximize returns and minimize risks, the mathematical model for the multi-objective investment portfolio problem is:

$$\begin{aligned} \max \quad & v = \max \left\{ \sum_{i=1}^N (Mx_i r_i - T(x_i)) \right\} \\ \min \quad & R = \min \{ 1 \leq i \leq n \{ Mx_i q_i \} \} \end{aligned} \tag{2.4}$$

2.2. The principle of artificial fish school algorithm . The main idea of the Artificial Fish Swarm (AFSA) algorithm is that in the body of water, the area where fish are most abundant is usually the area with the most nutrients. This decision is based on these characteristics to simulate the food, crowd, end-to-end collisions, and behavior of fish schools, in order to achieve global development.

(1) *Individual definitions of artificial fish.* The individual state of artificial fish can be represented as a vector $X = (x_1, x_2, \dots, x_n)$, where x_i ($i=1, 2, n$) is the variable to be optimized; The food concentration at the current location of the artificial fish is expressed as $Y=f(X)$; The distance between artificial fish individuals is expressed as $d_{ij} = ||X_i - X_j||$; The maximum number of attempts per foraging for artificial fish is $trynumber$, the perceived distance for artificial fish is $visual$, the maximum movement step for artificial fish is $step$, and the crowding factor is δ [5,6].

(2) *Description of artificial fish behavior.*

Foraging behavior: Set the current state of the artificial fish to X_i , and randomly select a state X_j within its perceptual range ($visual$), if the food concentration in this state is better than its current state, the artificial fish will move one step forward in that direction; On the contrary, randomly select state X_j again to determine whether the forward condition is met. After repeated attempts at $trynumber$, if the condition is still not met, randomly move one step.

Crowding behavior: Set the current state of artificial fish to X_i , explore the number of ($d_{ij} < visual$) partners nf and their central position X_c in the current field, if $Y_{cnf} < \delta Y_i$, it indicates that the food concentration at the center position is better than the current state and not too crowded, and the artificial fish moves forward towards the X_c position. Otherwise, they perform foraging behavior [7].

Tail chasing behavior: Set the current state of the artificial fish to X_i , and explore the optimal partner X_{max} in the current field, if $Y_{maxmf} < \delta Y_i$, it indicates that the food concentration at the partner's location is better than the current state and the surrounding area is not too crowded. If not, the artificial fish will move forward towards the X_{max} position. Otherwise, they will engage in behavioral activities.

Random behavior: This behavior is easy to execute and is the original behavior for the behavior.

Report board: used to record the status of the best fishermen. During the refining process, each artificial fish compares itself with the status of the reports. If the status of the news report itself is improved, it will be adjusted to its own status so that the news report can record the best events in history.

2.3. Multi-objective investment portfolio problem based on AFSA algorithm.

(1) *Model optimization.* For the solution of the multi-objective investment portfolio problem described by the author, the following method can be used to first transform it into a single objective problem, and then solve it: Set weights ρ for the weight of the total net investment return, then $1 - \rho$ weighting the overall investment risk ($0 \leq \rho \leq 1$), so the above multi-objective investment portfolio model can be transformed into a single objective model as follows:

$$\begin{aligned}
 & \max F(x) = \rho \times V + (1 - \rho)(-R) \\
 & \max F(x) = \rho \times \sum_{i=1}^N (rMx_i - T(x)) + (1 - \rho) \times (\max_{1 \leq i \leq n} \{Mx_i q_i\}) \\
 & \text{s.t.} \begin{cases} \sum_{i=1}^N x_j = 1, 0 \leq x_i \leq 1 (i = 1, 2, \dots, n) \\ 0 \leq \rho \leq 1 \end{cases}
 \end{aligned} \tag{2.5}$$

For the model ρ investors can make appropriate adjustments based on their own situation to obtain appropriate investment portfolio decisions [8].

(2) *Steps for solving the model.* The steps to solve multi-objective investment problems using AFSA algorithm are as follows:

Step 1: Start the size of artificial fish stocks, automatically create N artificial fish stocks, set the requirement for the algorithm, including the quantity of simulation and simulation of artificial fish stocks movement, the detection by artificial fish stocks, the large-scale step artificial fish stocks movement, and the crowds are equal δ the maximum number of iterations to achieve the algorithm.

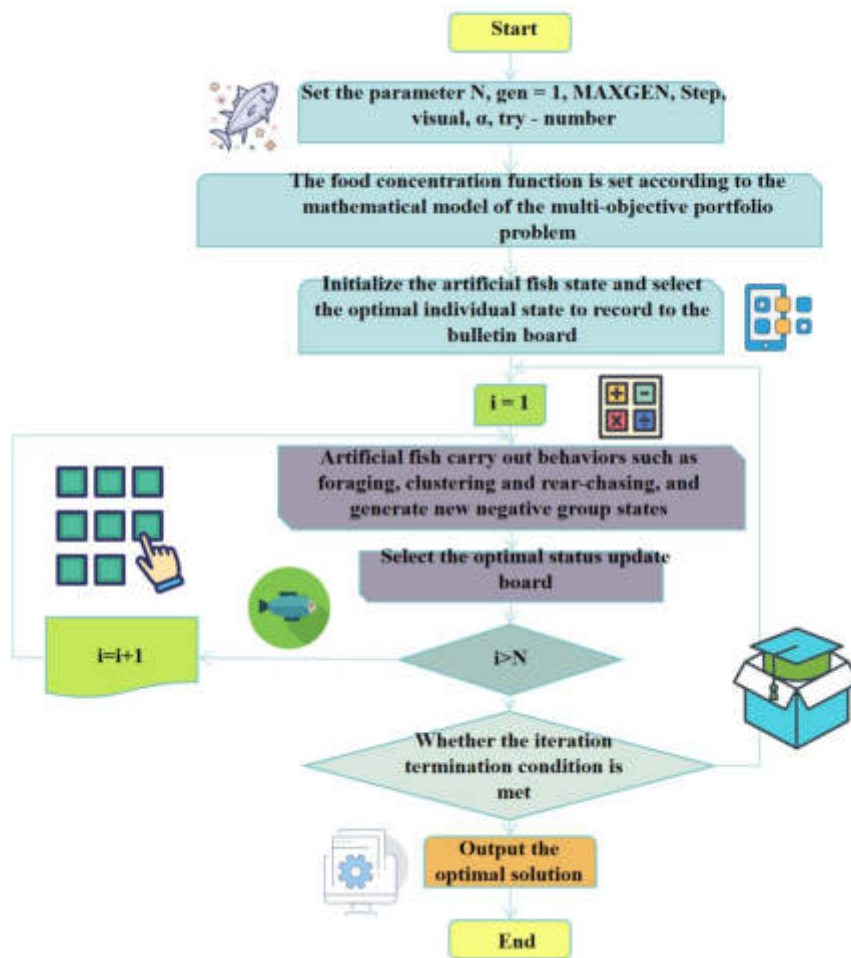


Fig. 2.1: Steps for solving multi-objective investment portfolio problems using AFSA algorithm

- Step 2: According to the mathematical model of multi- objective investment problem, set up the food concentration function of the algorithm.
- Step 3: Sample the current menu of all the fake fishermen and list the status of the best person on the poster.
- Step 4: Each artificial fish acts as a laborer, crowded, and tailor -made, chooses the best practices, and modifies its own state.
- Step 5: After the update, each fake fish will compare its status with the reports. If the effect is greater than the sign, the status of the sign will be adjusted.
- Step 6: Determine the decision for the algorithm. If iteration termination condition is met, output the calculated result; output otherwise, continue with step 4.

Figure 2.1 shows the steps to solve a multi-objective investment portfolio problem using the AFSA algorithm.

3. Simulation experiments. In order to verify the effectiveness and efficiency of the artificial fish school algorithm, the author made a simulation of AFSA algorithm and genetic algorithm (GA) using MATLAB programming, and compared the results. There are five investment assets available for selection, with a total investment amount of M of 1000 yuan. The data on the return rate, risk loss rate, and other factors of each asset are shown in Table 3.1 [9].

The parameters of artificial fish school algorithm were set fish school size of 50%, the quantity of iterations

Table 3.1: Relevant data of various investment assets

S_i	r_i	q_i	p_i	u_i
S_1	29	2.6	1.1	104
S_2	24	5.6	4.6	53
S_3	22	1.6	2.1	199
S_4	6	0	0	0
S_5	26	2.7	6.6	41

Table 3.2: Different ρ optimal investment portfolio corresponding to the value

ρ	x_1	x_2	x_3	x_4	x_5	V(%)	R(%)	F(x)
0.0	0.000	0.000	0.000	1.000	0.000	5.00	0.00	0
0.1	0.2882	0.0825	0.3876	0.0029	0.2393	21.105	0.721	14.63
0.3	0.6786	0.3125	0.000	0.0092	0.000	24.146	1.697	60.57
0.5	0.8668	0.0094	0.0147	0.0136	0.0959	25.135	2.168	114.85
1.0	1.000	0.000	0.000	0.000	0.000	27.01	2.51	270.01

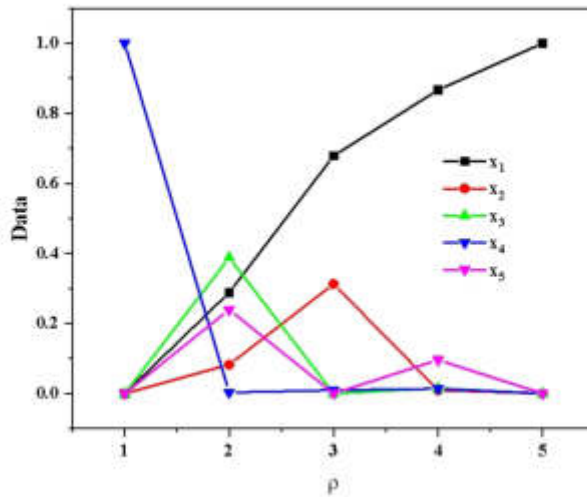


Fig. 3.1: Different ρ optimal investment portfolio corresponding to the value

of 100%, observed = 2.5%, moving step size = 0.4%, and population coefficient. respectively. $\delta=0.618$, and detection calculation=100; Using MATLAB programming to adopt different risk ρ Resource allocation method is shown in table 3.2 and Figure 3.1.

From Table 3.2, it can be seen that for multi-objective investment portfolio problems, under different risk preferences, the artificial fish school algorithm can obtain better fund investment portfolio schemes. The simulation experiment results show the effectiveness and feasibility of AFSA algorithm to solve such problems. The range of iterations for genetic algorithm used for comparison is $N=100$, with a crossover probability of 0.8, a mutation rate of 0.01%, and a population size of $P=100$; The comparison of investment portfolio schemes based on the calculation results is shown in Table 3.3 and Figure 3.2 (assuming $\rho= 0.5$).

From Table 3.3, it is evident that for investing in five types of assets, compared to the GA algorithm, the

Table 3.3: Comparison of Investment Portfolio Plans

	x_1	x_2	x_3	x_4	x_5	V(%)	R(%)
GA	0.9073	0.0022	0.0008	0.0893	0.0009	25.008	2.269
AFSA	0.8668	0.0094	0.0147	0.0136	0.099	25.135	2.168

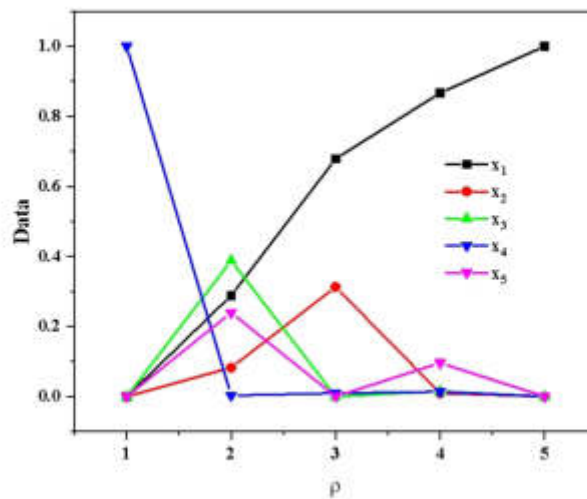


Fig. 3.2: Comparison of Investment Portfolio Plans

artificial fish school algorithm can obtain better investment portfolio decision-making solutions, maximizing investment returns while minimizing risks, indicating the efficiency and superiority of the algorithm in solving multi-objective investment portfolio problems [10].

4. Conclusion. The author proposed to study and apply the artificial fish school algorithm to multi-objective investment portfolio problems, and achieved good results through MATLAB programming. Experimental results show that this algorithm performs very well in solving such problems, but as a new type of optimization algorithm, there are still some shortcomings. Therefore, further research and improvement are needed to improve the performance of solving such problems when there are a large number of assets and investment funds.

REFERENCES

- [1] Bo, X., Wenlong, F., & Zhang, J. (2022). Whale optimization algorithm based on artificial fish swarm algorithm. Springer, Cham.3(79),54
- [2] HU Zhiyuan, WANG Zheng, YANG Yang, YIN Yang. (2022). Three-dimensional global path planning for uuv based on artificial fish swarm and ant colony algorithm. Acta Armamentarii, 43(7), 1676-1684.
- [3] Zhang, L., Fu, M., Li, H., & Liu, T. (2021). Improved artificial bee colony algorithm based on damping motion and artificial fish swarm algorithm. Journal of Physics: Conference Series, 1903(1), 012038 (9pp).
- [4] Zhang, X., Lian, L., & Zhu, F. (2021). Parameter fitting of variogram based on hybrid algorithm of particle swarm and artificial fish swarm. Future Generation Computer Systems, 116(1), 265-274.
- [5] Lin, M., Yuan, X., Lei, H., & Ji, Z. (2021). Kinematic analysis of tensegrity mechanisms based on improved artificial fish swarm algorithm with variable step size. Journal of Physics: Conference Series, 1903(1), 012071 (6pp).
- [6] Wang, X., & Li, Y. (2021). Chaotic image encryption algorithm based on hybrid multi-objective particle swarm optimization and dna sequence. Optics and Lasers in Engineering, 137(11), 106393.

- [7] Ibrahim, R., Abualigah, L., Ewees, A., Al-Qaness, M., Yousri, D., & Alshathri, S., et al. (2021). An electric fish-based arithmetic optimization algorithm for feature selection. *Entropy (Basel, Switzerland)*, 23(9),12.
- [8] Yi, H., Wang, J., Hu, Y., & Yang, P. (2021). Mechanism isomorphism identification based on artificial fish swarm algorithm. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 235(21), 5421-5433.
- [9] Qingyao, T., Ran, W. U., Jiajia, J. I., Junyuan, L., & Zhangyong, L. I. (2021). Research on illumination optimization of phototherapy led based on multi-objective artificial fish swarm algorithm. *Journal of Applied Optics*, 42(2), 352-359.
- [10] Huang, J., Zeng, J., Bai, Y., Cheng, Z., & Liang, D. (2021). Layout optimization of fiber bragg grating strain sensor network based on modified artificial fish swarm algorithm. *Optical Fiber Technology*, 65(1), 102583.

Edited by: Zhigao Zheng

Special issue on: Graph Powered Big Aerospace Data Processing

Received: Jan 2, 2024

Accepted: Feb 18, 2024



OPTIMIZATION ALGORITHM FOR GREEN ENVIRONMENT DESIGN BASED ON ARTIFICIAL INTELLIGENCE

KAI QIAN*

Abstract. In order to build energy-efficient commercial buildings in bustling urban centers and utilize passive means such as natural ventilation and natural lighting as much as possible to improve indoor environmental quality, the author proposes a green environment design optimization algorithm based on artificial intelligence. Taking Building A as an example, simulation technology was used to optimize the performance of the enclosure structure, anti condensation in the glass atrium, and natural ventilation during the transition season. The sunlight and shadow simulation software BSAT was used to simulate the mutual occlusion and self occlusion of the building. It was found that external rolling shutters were not required for shading the facade between axes L2-L5 and 3-9. Using DeST-C, the energy consumption and investment of south facing, east facing, and west facing envelope structures were compared when using different types of glass. Combining economic efficiency and energy-saving effects, the optimal southbound enclosure structure scheme for this building is to use Low-e membrane coated hollow double glass with ten horizontal louvers for external shading; It is recommended to adopt the scheme of hollow double glass (Low-e membrane) or hollow double glass (Lowe membrane)+external roller shutter in the east-west direction. It is worth noting that for east-west oriented glass, the Lowe film should be low permeability, mainly to improve the thermal performance in winter and reduce the radiation heat gain in summer. Using DeST-C for simulation 2 calculation, it was found that when the thermal performance of windows increased from $3.0W/(m^2 \cdot K)$, shading coefficient 0.7 to $2.0W/(m^2 \cdot K)$, and shading coefficient 0.5, the maximum cold and heat load of the building and the cumulative consumption of cold and heat throughout the year were significantly reduced, providing an effective reference for solving problems.

Key words: Artificial intelligence, Green environment, Design optimization algorithms, Energy saving, Enclosure structure

1. Introduction. Due to the high-density agglomeration of urban population and pollution caused by production and daily life, the ecological environment of urban residential areas is increasingly valued by people [1]. Green environment design is an important means of ecological environment design. Green environment design for residential areas can fully improve the climate characteristics of residential areas, and achieve unity and harmony with nature in the living environment of residential areas. Generally speaking, the form of green environment design in residential areas includes grassland greening, tree (tree, shrub) greening, balcony and terrace greening, interior and exterior corridor greening, exterior wall greening, roof platform greening, and indoor greening. It is a three-dimensional and all-round green environment. In residential areas, grasslands not only provide recreation, but also provide a broad view and become the center of public activities, serving as the "breathing space" of residential areas. But in some current residential areas, developers have made the central green space area too large in pursuit of grandeur, and there is a lack of coordination between shrubs and trees. Due to the limitation of the overall plot ratio, residential areas have reduced the green coordination of the adjacent green spaces, on the other hand, there is a lack of trees to shade the sun, resulting in a lack of humanized space in the central green space, which lacks a lot of fun and even becomes purely a decoration. Trees can absorb dust and noise in residential areas, reduce wind speed, preserve soil and water, and block the scorching sun in summer. On the landscape, it can enclose the space and block the view. Meanwhile, the oxygen production of trees is significantly higher than that of grass. When selecting tree species, attention should be paid to selecting trees that are suitable for the local climate, soil, and hydrological conditions, emphasizing the combination of common and precious tree species, the combination of deciduous and evergreen trees, the arrangement of tree clusters and clusters, and the selection of tree size, height, and morphology. The ecological environment of balconies varies greatly due to differences in their enclosed form, orientation, height, and shape. The main way of balcony greening is to place potted flowers inside the balcony and on the concrete handrails

*Shandong University Of Arts, Jinan, Shandong, China, 250300 (18615545701@163.com)

of the railing, and to build various types of planting slots in sync with the civil engineering project. It can be installed around and along the balcony railing, and can also be combined with the solid balcony railing to form a flower hopper groove shape. At the same time, various planting pots can be hung on the hanging board of the railing, which not only enriches the shape of the balcony railing, but also increases the function of planting flowers. At the same time, by setting up greenery on balconies and terraces, the building facade can be enriched, adding a moving and bright color to the building. In architectural design, flower beds and racks can be set up on balconies and terraces, with potted plants as the main greenery, accompanied by climbing plants, which should have strong ornamental value and can generally be floral plants.

Currently, energy conservation and improving indoor environmental quality are receiving increasing attention in architectural design. How to build energy-efficient commercial buildings in bustling urban centers and make the most of passive means (such as natural ventilation, natural lighting, etc.) to improve indoor environmental quality is of great significance for the sustainable development of China's construction industry. In the current development process of large-scale commercial buildings, due to excessive pursuit of aesthetic effects and consideration of eye-catching benefits in the early stage of investment attraction, glass curtain walls are extensively used as facade elements in building design, but the thermal effects and reasonable selection of glass curtain walls are not fully considered, often leading to the deterioration of building thermal performance, increased initial equipment investment, and high energy consumption of air conditioning system operation. At the same time, the summer solar radiation near the glass curtain wall is strong, and the average indoor radiation temperature is high; In winter, when the temperature is low and the radiation is strong, the indoor thermal comfort will significantly decrease. In addition, for the deep interior area, it is also necessary to consider how to effectively organize ventilation and cooling during the transitional season and improve indoor thermal comfort.

In order to avoid the problems that glass curtain wall buildings are prone to in building A, the author fully implemented the concept of energy-saving design in the early design stage, adopted advanced computer simulation technology, optimized the enclosure structure and passive natural ventilation design, and achieved certain results [2,3].

2. Methods. Project A is one of the renovation and construction projects for the commercial area on the west side of Xidan North Street in City A. It is located in the bustling center of the city, with a land area of about $16700m^2$ and a building area of about $200000 m^2$. The design requirements meet various commercial activities such as large-scale commercial, high-end office, catering, entertainment, and hotels. The entire building is lightweight and transparent, using a large number of glass components to enhance the integration between humans and the natural environment; In addition, there is a transparent inverted cone atrium with a unique and prominent design, which can also serve as a natural ventilation channel. For such a super large commercial building, in order to avoid the problems that glass curtain wall buildings are prone to, the concept of energy-saving design was fully implemented in the early design stage. Advanced computer simulation technology was used to optimize the enclosure structure and natural ventilation passive design, and certain results were achieved. Here is a brief introduction.

2.1. Ventilation optimization design. Combining the characteristics of this building, we focus on using natural ventilation or mechanical assisted natural ventilation methods to solve the transitional season ventilation problems in most areas. The layout of various cinema buildings in Building A is relatively dense, making it difficult to organize natural ventilation. Therefore, it is recommended to use a full air system during the transition season to operate under full fresh air and full exhaust conditions [4]. Thermal natural ventilation (or mechanical assisted thermal natural ventilation) mainly relies on the tall inverted cone atrium in the middle of the building and the south side atrium space of the cross. Figure 2.1 shows the indoor natural ventilation node diagram. In order to ensure the effectiveness of thermal ventilation, a certain open external window area is required to construct a calculation program for the thermal fluid network, using meteorological conditions of typical days in the transitional season for calculation [5]. The calculation method is as follows:

- ① Select representative days for the transitional season, calculate the solar heat gain in the atrium and the internal disturbances in various functional areas;
- ② Estimate the appropriate natural ventilation rate to determine the temperature distribution (average air exchange rate) in the atrium; The temperature distribution can be determined according to a linear distribution relationship based on existing research results;

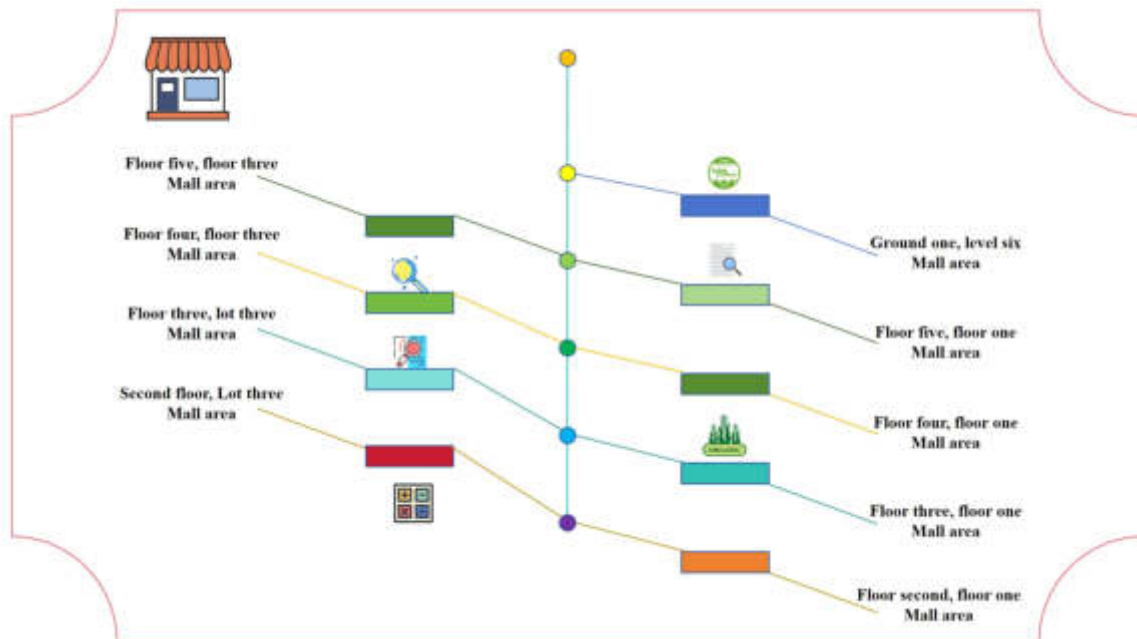


Fig. 2.1: Indoor Natural Ventilation Node Diagram

- ③ Using regional network method to calculate the natural ventilation flow rate of each branch, and guiding and optimizing building design; If there is a discrepancy between the calculated ventilation volume and the estimated value in , the estimated value in should be adjusted, and the above steps should be repeated to solve the calculation until the calculated value and estimated value match.

In this way, the openable window areas for different parts were calculated separately. The result is that the opening area of the north exterior window should not be less than $20m^2$ /floor; The south and west sides of Building D must ensure that the opening area of the external windows on each floor is not less than $30m^2$; The first floor of building B should ensure that the opening area of the external windows is not less than $15m^2$. In addition, on the top of the south facing part of the cross courtyard, an outer window that can be mechanically controlled to open should be installed, and mechanical ventilation equipment should be installed to timely discharge accumulated heat under adverse working conditions. The exhaust equipment can be used in conjunction with the atrium smoke exhaust equipment. Due to the presence of mechanical equipment, the area of the external window that can be opened can be slightly smaller, but should not be less than $20m^2$. For the inverted cone atrium, mechanical control side windows are installed at the top of the twelfth and thirteenth floors, with a total opening area of not less than $50m^2$.

2.2. Energy saving design of exterior facade enclosure structure . Due to the fact that almost all of the peripheral protective structures in the original plan of this building were glass components (K value of about $3.0W/(m^2 \text{ } ^\circ K)$), in the deepening design, according to the original building drawings and the functional and usage requirements of the space, glass curtain walls were minimized as much as possible, and walls (or designed as inner solid walls and outer glass) were added to improve the thermal performance of windows, reducing the drawbacks of operating load and high operating costs caused by the relatively poor thermal performance of too many glass curtain walls. Under this principle, for functional spaces such as shopping malls, elevator rooms, air conditioning rooms, fire centers, garbage stations, storage rooms, bathrooms, etc., the outer enclosure structure will be as solid as possible with some external windows added (some of which are mainly used for ventilation and lighting); Or design according to the combination of inner solid wall and outer glass. In addition, for areas with LCD screens and billboards on the facade, the corresponding exterior walls are also adjusted to solid

Table 3.1: Comparison of energy-saving effects and initial investment increase of different enclosure structure schemes in the south direction

	Hollow double glass +single glass+horizon- -tal shading (double- layer curtain wall)	Hollow double glass (Low-e film)+horizon- -tal shading	Hollow double glass (coated with ten inert gases) +horizontal shading
Annual cumulative energy savings /(kWh/ $m^{2\circ}$ (year))	56.3	75.6	90.7
Annual operating cost savings (yuan/($m^{2\circ}$ year))	39.4	53	63.5
New initial investment/ (yuan/ m^2)	500	300	800

Table 3.2: Comparison of energy-saving effects and initial investment increase of different enclosure structure schemes in the east and west directions

	Hollow double glass +external rolling shutter	Hollow double glass+lou- -ver shading +single glass	Hollow double glass (Low-e membrane)	Hollow double glass (Low-e film)+exter- -nal roller shutter	Hollow double glass (coated with Low-e film)+lou- -ver shading+single glass
Annual cumulative energy savings /(kWh/($m^{2\circ}$ year))	84	92	77	119	119
Annual operating cost savings (yuan/($m^{2\circ}$ year))	59.1	64.7	54.2	83.6	83.6
New initial investment /(yuan/ m^2)	150	500	300	450	800

walls. Using the sunlight and shadow simulation software BSAT for building mutual occlusion and self occlusion simulation, it was found that the facade between L2-L5 and 3-9 axes does not need to be designed with external rolling shutters for shading. However, for the 6th floor and above of Building C (including facades outside the 3-9 axis), it is recommended to design external roller blinds or other sunscreen measures. For Building D (office building), the facade below the seventh floor in the west direction and between the 1/2 axis of the 12-15 axis can be blocked by its adjacent buildings in the west direction between 15:00 and 17:00, and external rolling shutters can be omitted for shading.

3. Results and Analysis.

3.1. Energy saving design results of exterior facade enclosure structure. Using DeST-C, the energy consumption and investment of south facing, east facing, and west facing envelope structures were compared when using different types of glass. The calculation results for the south, east, and west directions are shown in Tables 3.1 and 3.2, respectively. The thermal performance parameters of various glasses are shown in Table 3.3 [6].

Combining economic efficiency and energy-saving effects, the optimal southbound enclosure structure scheme for this building is the combination of Low-e membrane coated hollow double glass and horizontal louver external shading; It is recommended to adopt the scheme of hollow double glass (Low-e membrane) or hollow double glass (Low-e membrane)+external roller shutter in the east-west direction. It is worth noting that for east-west oriented glass, the Low-e film should be low permeability, mainly to improve the thermal performance in winter and reduce the radiation heat gain in summer.

Using DeST-C for simulation 2 calculation, it was found that when the thermal performance of windows

Table 3.3: Thermal performance settings for different glass curtain walls

Glass type	Heat transfer coefficient K value/(W/(m ² °K))	Sunshade coefficient SC
Ordinary hollow	3.1	0.71
Ordinary hollow+single glass (double-layer ventilation curtain wall)	2.1	0.54
Hollow coated with Low-e film	1.9	0.52
Inert gas filling+hollow coating	1.6	0.52
Low e hollow+single glass	1.5	0.41
Inert gas filling +coating hollow+single glass	1.3	0.41

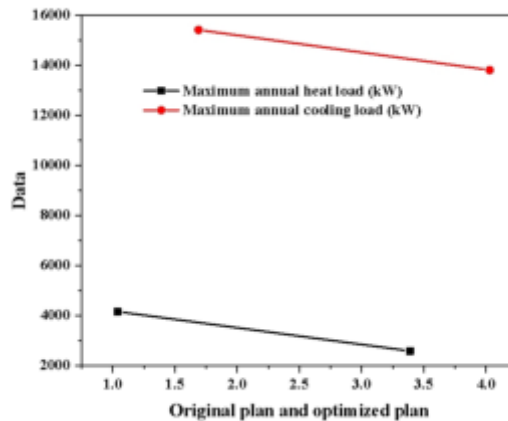


Fig. 3.1: Energy saving effect (load) of glass thermal performance optimization

increased from 3.0W/(m².K), shading coefficient 0.7 to 2.0W/(m² °K), and shading coefficient 0.5, the maximum cold and heat load of the building and the cumulative consumption of cold and heat throughout the year all decreased significantly, as shown in Figures 3.1 and 3.2 [7].

3.2. Analysis of Glass Inverted Cone Atrium Envelope Structure Scheme. The glass inverted cone atrium is located in the middle of this building complex. Due to the use of a full glass curtain wall structure, in winter, the outdoor temperature is very low, and the strong sky radiation at night at the top of the cone will cause a lower surface temperature inside the glass cone, which may lead to large-scale condensation [8-10]. Therefore, it is necessary to analyze the inner surface temperature of the glass cone during the most unfavorable working conditions in winter and make improvements. Using CFD simulation software PHONEICS to simulate and calculate the air temperature distribution inside the cone. The upper diameter of the cone is 28m, the lower diameter is 12.6m, and the height is 25m. The lower part is an opening, and the air entering the cone has a dry bulb temperature of 20 °C, a relative humidity of 40%, a dew point temperature of 6 °C, and an outdoor temperature of -10 °C. The heat transfer coefficients of the inner and outer surfaces of the glass are 8.7W/(m² °K) and 18.6W/(m² °K), respectively. Considering the strong radiation heat transfer between the top of the cone and the sky at night, the heat transfer coefficient of the outer surface of the top glass is set at 30W/(m² °K). For a single glass, the comprehensive heat transfer coefficient is set at 6.0W/(m² °K), while for ordinary hollow double glass, it is set at 3.0W/(m² °K). It can be seen that indoor air at 20 °C rises from below, cools at the top, and then descends along the side and flows out of the cone. The air temperature inside the cone is around 12-15 °C, and in some areas such as the area around the top of the

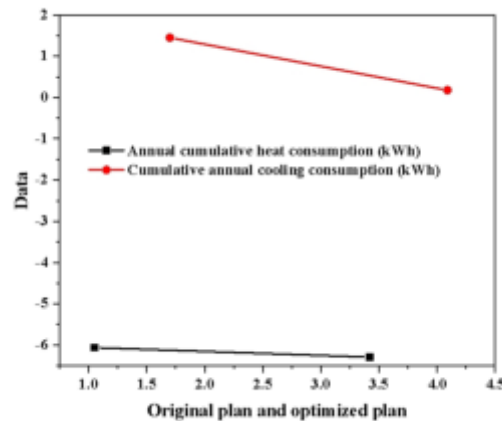


Fig. 3.2: Energy saving effect (heat) of glass thermal performance optimization

cone, the air temperature can drop to $8\text{ }^{\circ}\text{C}$. When the outdoor air temperature is $-10\text{ }^{\circ}\text{C}$, according to the method of heat transfer resistance difference, the temperature on the inner surface of the glass cone is about $3\text{ }^{\circ}\text{C}$, and the temperature on the inner surface of the glass at the top of the cone is also very low, about $1\text{--}6\text{ }^{\circ}\text{C}$. The temperature on the inner surface of the edge area of the cone is relatively low, about $1\text{ }^{\circ}\text{C}$, and a higher value of $6\text{ }^{\circ}\text{C}$ can be reached at the center. Therefore, if a single-layer glass scheme is adopted, a large area of condensation will inevitably occur on the inner surface of the top of the glass cone. If double-layer glass is used, most of the air temperature can be at $14\text{ }^{\circ}\text{C}$ or above, and even the edge area of the most unfavorable cone top can reach an air temperature of $13\text{ }^{\circ}\text{C}$. On the other hand, the temperature on the inner surface of the conical glass has also increased to a certain extent. According to the difference calculation, the temperature on the inner surface of the conical side can be around $6.3\text{ }^{\circ}\text{C}$, slightly higher than the dew point temperature of the air. The inner surface temperature at the top of the cone is $4.4\text{--}6.3\text{ }^{\circ}\text{C}$, and the surface area below the dew point temperature of $6\text{ }^{\circ}\text{C}$ accounts for about 60% of the total top area. This means that a considerable portion of the inner surface at the top will still experience condensation. According to the above analysis method, it is recommended that the heat transfer coefficient of the inverted atrium glass should not exceed $2.5\text{W}/(\text{m}^2\text{ }^{\circ}\text{K})$.

4. Conclusion. In response to the current pursuit of aesthetic and eye-catching benefits in the development process of commercial buildings, the large-scale adoption of transparent enclosure structures has led to difficulties in meeting indoor thermal comfort and high operating costs. For a large commercial building, analysis and calculation were conducted from several aspects, including optimizing the performance of glass curtain wall enclosure structures, designing anti condensation measures for glass atriums, and designing natural ventilation during the transition season, thus, the energy-saving design of the enclosure structure of the large commercial building and the improvement of indoor environmental quality were achieved. The author has made beneficial explorations on how to build energy-efficient commercial buildings in bustling urban centers, and how to use passive means (such as natural ventilation, natural lighting, etc.) as much as possible to improve indoor environmental quality, which has reference significance for other similar buildings.

REFERENCES

- [1] Xu, X. (2021). Link optimization of the new generation instant messaging network based on artificial intelligence technology. *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*3(4), 40.
- [2] Dong, X., & Zeng, L. (2021). Research on query optimization of classic art database based on artificial intelligence and edge computing. *Wirel. Commun. Mob. Comput.*, 2021, 6118113:1-6118113:11.

- [3] Li, Y., Shi, X., Diao, H., Zhang, M., & Wu, Y. (2021). Optimization of warehouse management based on artificial intelligence technology. *Journal of Intelligent and Fuzzy Systems*56(6), 1-8.
- [4] Zhang, X., Lian, L., & Zhu, F. (2021). Parameter fitting of variogram based on hybrid algorithm of particle swarm and artificial fish swarm. *Future Generation Computer Systems*, 116(1), 265-274.
- [5] Wang, Y. (2023). Research on design methodology for railway freight service combination plans to meet diverse demands. *Railway Sciences*, 2(4), 525-538.
- [6] Li, Q., & Li, S. (2021). Optimization of artificial cnn based on swarm intelligence algorithm. *Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology*88(4), 40.
- [7] Liu, D. (2021). Research on modern urban environment design based on gradient descent method in machine algorithm. 2021 2nd International Conference on Intelligent Design 23(ICID), 106-110.
- [8] Yin, Y. (2021). Research on ideological and political evaluation model of university students based on data mining artificial intelligence technology. *J. Intell. Fuzzy Syst.*, 40(42), 3689-3698.
- [9] Yu, H., & Liu, Q. (2022). Research on quality prediction of iron ore green pellets based on optimized neural network. 2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering66 (ICBAIE), 377-382.
- [10] Guo, A., & Yuan, C. (2021). Network intelligent control and traffic optimization based on sdn and artificial intelligence. *Electronics*, 10(6), 700.

Edited by: Zhigao Zheng

Special issue on: Graph Powered Big Aerospace Data Processing

Received: Jan 2, 2024

Accepted: Feb 21, 2024



IMAGE RECOGNITION TECHNOLOGY BASED ON DEEP LEARNING IN AUTOMATION CONTROL SYSTEMS

JINGJING WANG*

Abstract. In order to solve the problem of recognizing multiple product images, the author proposes the research of image recognition technology based on deep learning in automated control systems. Firstly, the FasterRCNN method is improved by proposing a non class specific FasterRCNN, which can be used for pre annotation of product images by training only on publicly available datasets. Due to the use of position correction networks, the pre annotation effect is more accurate than that of candidate region networks. Then, combining Grabcut with non class specific FasterRCNN, a sample enhancement method was proposed to synthesize a large number of training images containing multiple products and use them for model training. In addition, based on non class specific FasterRCNN, a re identification layer was proposed to improve detection accuracy. In the end, the recognition and positioning of multiple products achieved a recall rate of 93.8% and an accuracy of 96.3%.

Key words: Supermarket product image recognition, Deep learning, Data annotation algorithm, Non class specific FasterRCNN

1. Introduction. In the design process of a product, in order to stimulate consumer purchasing desire while ensuring product recognition, the product image features are very rich. Common image features can be divided into two categories: One is the low-level visual features, which are the global and local features of the product, another type is the intermediate semantic features of the product, which are mainly applied in the process of product recognition based on the underlying visual features. That is to say, the image features of products can be specifically divided into the following types, including color features, shape features, texture features, point features, semantic features, and other aspects [1]. Among these features, color feature is the most direct way. By using color histogram, the color distribution of the product image can be obtained, while shape regions and contour boundaries can be distinguished. Through specific detection algorithms, the feature points of the product can be effectively extracted. Deep learning technology is a perception technique built on neural networks, commonly including neural models, perceptrons, BP algorithms, convolutional neural networks, etc. Taking convolutional neural networks as an example, they are currently the most widely used type of neural network and have been widely used in facial recognition, speech recognition, license plate recognition, object detection, and other fields [2, 3]. Convolutional neural networks and BP neural networks are similar in that they both consist of an input layer, an intermediate layer, and an output layer. However, compared to the former, the intermediate layer is more complex. In practical applications, this neural network has enormous advantages, especially in network depth and massive image data processing. The most important thing is that using this neural network can better complete learning and training. After determining the specific neural network, deep learning can be carried out, utilizing massive data and network models for learning. Feature learning in the data is a very important content, which can ensure the accuracy of model prediction is improved.

In recent years, due to the rapid development of mobile internet, with the support of cloud computing and big data, online shopping like e-commerce has sparked a new wave of people's lifestyles [4]. However, due to the gradual saturation of online users, transaction profits have also gradually decreased. More and more companies are turning their attention to the offline trading platform. The layout of offline transactions by companies such as Alibaba, Meituan, and Ele.me is evident. The retail industry has also become a focus of offline transaction competition. Enterprises are increasingly focusing on how to use artificial intelligence technology to transform product production, circulation, and sales, reshape the industry ecosystem, and integrate online and offline

*Department of Information Engineering, Yangzhou Polytechnic College, Yangzhou, Jiangsu, 225000, China (yzwj2023@163.com)

experiences. That is to say, future innovative retail may be entirely achieved through artificial intelligence technology. In the current era of booming artificial intelligence algorithms, researching retail product recognition technology to improve productivity has become one of the hotspots in the field of artificial intelligence. In the past two years, some automated retail stores have emerged both domestically and internationally, such as AmazonGo, Taobao Coffee, etc. The use of artificial intelligence technology to achieve automation and unmanned retail scenarios has become a trend and is gaining momentum. Nowadays, artificial and intelligent technologies are using algorithms to process information in order to reach the level of humanity. As is well known, humans are best at using vision and hearing to receive, process, and transmit information. In the process of purchasing goods in supermarkets, people have a clear understanding of the products they want to purchase and can visually locate and identify them. In the settlement process, it is still common to use the cashier to scan the code one by one for settlement. With the development of artificial intelligence algorithms in the field of computer vision in recent years, machines can recognize universal objects like humans, such as roads, pedestrians, vehicles, etc. However, in specific tasks, such as the settlement process of supermarket products, how to use image vision to identify items, in order to liberate productivity, improve automation and intelligence, has become a research hotspot in the field of artificial intelligence. For the recognition and localization of multiple products, the author improved the regression layer of FasterRCNN and fully utilized publicly available datasets to learn object bounding box regression knowledge for pre labeling of product images, which is more accurate than the labeling algorithm of RPN. Then, a combination of pre labeling algorithm and Grabcut algorithm, an image synthesis algorithm, is proposed to solve the recognition and localization problem of multiple products, and its working principle is described in detail [5].

2. Methods.

2.1. Problem Description. In recent years, deep learning has been used in many fields. Such as facial recognition, object detection, etc. Deep learning can capture useful information from a large amount of data, and due to the advent of the big data era and the improvement of computing device performance, the application of deep learning has become a reality. FasterRCNN and its extended versions have become one of the most effective methods in recent years. However, if the above methods are transferred to a new task, a large amount of calibration data needs to be used to readjust the model on the new task. However, in practical scenarios, data calibration is a very difficult task that requires significant financial, material, and human resources.

The author proposes a method to solve the data bottleneck problem, which can detect and locate products without the need for border calibration. The dataset constructed by the author contains only a single product in the training images without border calibration, while the test images contain multiple products [6]. The author first improves FasterRCNN by proposing a non class specific FasterRCNN, and combines transfer learning to pre calibrate the training data; Then, combined with the unsupervised Grabcut2l method, the training data is sample enhanced to generate realistic training images of multiple objects; Then train the non class specific FasterRCNN to detect multiple objects; Finally, the author proposes a re identification method based on FasterRCNN, adding a re identification layer to FasterRCNN to improve the accuracy of multiple object detection.

2.2. System Framework Design. The product recognition and localization method proposed by the author can be divided into three modules: The first part is the bounding box calibration module of non class specific FasterRCNN. The second part is a sample enhancement method that combines non-specific FasterRCNN with Grabcut [7]. The third part is a region based re identification method. The first part, the proposal of non-specific FasterRCNN, is aimed at solving data barriers and fully utilizing prior knowledge in public datasets to solve the calibration work of borderless calibration training samples. On the one hand, it removes the impact of complex redundant backgrounds. On the other hand, it provides border interaction input for Grabcut, and at the same time, it can complete border prediction of product areas without training, effectively solving the problem of product positioning. In the second part, by combining Grabcut and non-specific FasterRCNN methods, accurate region extraction is achieved by effectively utilizing the pre calibration information of borders. Based on the extracted regions, fuse the training images of the products to generate a large number of images of multiple products. By using the generated images for overall model training, it is possible to distinguish between cases of product occlusion or overlapping borders, effectively solving the problem of multiple product

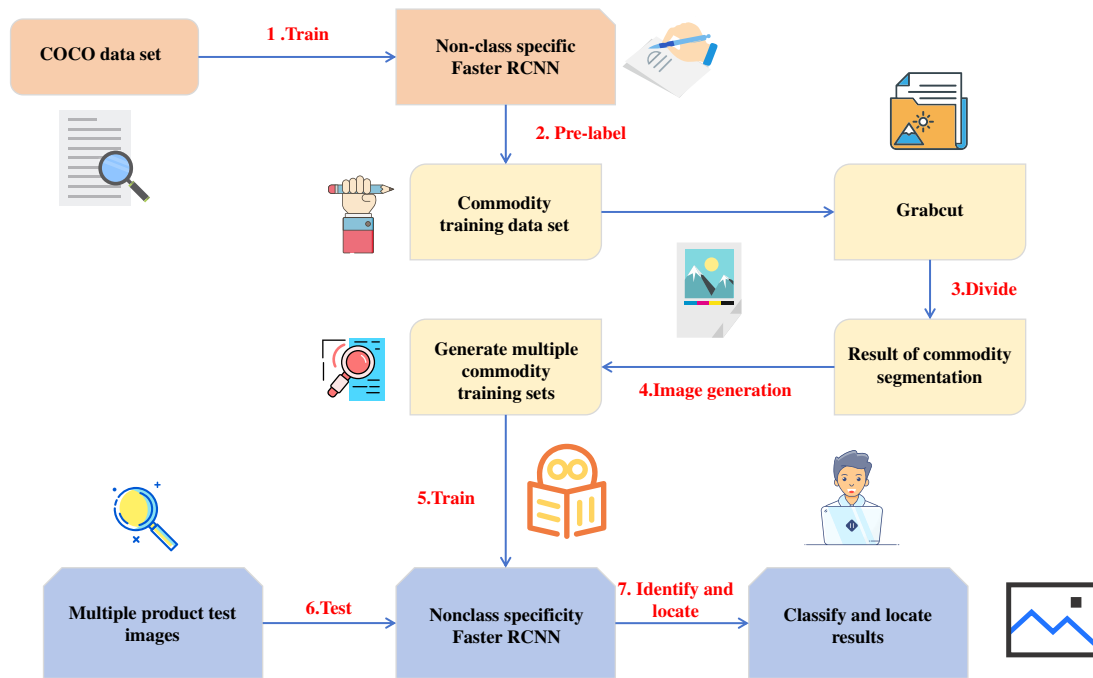


Fig. 2.1: Block diagram of multiple product identification and positioning systems

neighbors or occlusion. Realized the recognition and localization of multiple product images based solely on a single product training image calibrated without borders. In the third part, in order to improve the accuracy of system recognition, a region based re identification method is proposed to solve the problem of misidentifying objects and improve the prediction accuracy of the system. As shown in Figure 2.1.

2.3. Specific System Implementation. Combining Grabcut's sample enhancement method: The non class specific FasterRCNN solves the problem of bounding box regression, making the pre calibration of product positions in training images more accurate. When there is only a single product in the training data, and there are multiple products in the test images, the classification problem still has significant difficulty. Even if a single product can be used to train a classification model, multiple products in the test image may have overlapping or even occluded borders. Due to the lack of product overlap in the training images, training the model will make its classification ability less robust and insufficient to recognize products with overlapping or even occluded borders, increasing the difficulty of classification. Therefore, the author proposes a sample enhancement method to generate training images with multiple products by processing a single product training image. Because using category non-specific FasterRCNN, the borders of individual product images can be obtained. A direct idea is to extract the border part of the product in the image, rotate or translate it, and combine it with other products. However, this method will result in the background area of the product border covering other product areas, and the edge area of the border will differ significantly from the actual image. Therefore, using only the borders of product images is not enough to achieve realistic sample generation. If precise area information of the product can be obtained, such as the product object mask, the background area can be separated to solve the problem of the product being obscured by the background in the generated sample. Therefore, the author uses the Gabcut method to segment the products in the training images.

Both Grabcut and Graphcutsl methods are interactive image segmentation methods. Graphcut needs to provide precise foreground and background pixel seed regions during interaction, calculate the similarity between other pixels and foreground and background, and use graph theory algorithms to calculate the optimal segmentation. The Grabcut algorithm has less user interaction and only needs to provide a rectangular border containing the foreground. Then, the pixels inside the border are used as the foreground, and the pixels outside

the border are used as the background. Gaussian Mixture Model (GMM) is used to model the background, and graph theory algorithm is used for segmentation. The GMM background modeling and graph theory algorithm are repeatedly used until the iteration converges, completing the segmentation.

After utilizing the non class specific FasterRCNN proposed by the author, the rectangular border of a single product in the training image can be obtained. Therefore, it is only necessary to combine the Grabcut algorithm to segment the precise region of the product. Then, the individual product areas in the training set are randomly rotated and translated, and randomly combined to generate training images for multiple products. It is worth noting that, considering the accuracy of the data, products cannot be completely covered between each other, because if the products are excessively covered, it will lead to almost all the real products in the area being covered, and the products occupying a large area will not match the actual labels, which will mislead the training of the recognition model. Therefore, when combining randomly, it is necessary to constrain the overlapping area of the products, assuming that the upper limit of the overlapping area is Sup . Consider three fusion strategies: (1) Perform random rotation and translation, only constraining the upper limit of the overlapping area, that is $sc \leq sup$, which means that during fusion, the product may be far away, at which point $sc=0$. (2) While limiting the upper limit of overlapping area, constrain the lower limit of overlapping area, that is $sc>0$. This approach requires overlap between products, ensuring a close distance between them without extensive coverage. (3) Increase the constraint on the lower limit of overlap area, that is $sc \geq sm$, in order to increase the possibility of overlap between products and differentiate the overlapping situation through model training.

The re identification layer found that the model trained by the above method will have a small number of adjacent bounding boxes predicting two different results, one of which is correct and the other is incorrect. Because FasterRCNN is a two-level (two-stage) method, in the first stage, the RPN (candidate region network) first filters out candidate regions and filters out a portion of the background region; The second level uses the head network to finely classify candidate regions, while correcting the borders of each candidate region, known as border regression. Obviously, the candidate regions extracted by RPN are imprecise, which can affect the accuracy of head network recognition. Because some candidate regions only cover local areas of the object, although bounding box regression can correct them to the correct position, directly using candidate regions for prediction may inevitably lead to misidentification. Therefore, the author proposes a re identification layer to improve the accuracy of FasterRCNN recognition. Because the bounding box position is more accurate after passing through the bounding box regression layer of the head network, the bounding box regression layer here is a non class specific regression method proposed by the author. Moreover, the classification layer of the head network filters out a large portion of the background area. Therefore, the results of bounding box regression can be classified first, and then more accurate bounding box regression results can be used as candidate regions, followed by some re identification. The author will use the precise regions obtained from the head network regression, combined with the ROIAlign method, as inputs to pass through the classification layer of the head network again. Traditional FasterRCNN can be defined as:

$$\begin{cases} c = f_{cls}(roi) \\ reg = f_{reg}(roi) \end{cases} \quad (2.1)$$

Among them, roi represents the candidate regions generated by RPN, f_{cls} represents the classification layer, f_{reg} represents the regression layer, c represents the classification result corresponding to the candidate region, and reg represents the bounding box regression result of the candidate region. The added re identification layer selects the regions classified as non background in the candidate regions, with the background category represented by 0. Then, its regression border is used as the new candidate region for classification and regression, represented as:

$$\begin{cases} c' = f_{cls}(reg_{c \geq 0}) \\ reg' = f_{reg}(reg_{c \geq 0}) \end{cases} \quad (2.2)$$

Among them, c' represents the final classification result after re identification, and reg' represents the final bounding box regression result.

Table 2.1: Basic parameter settings for non class specific FasterRCNN

base_lr	0.001
momentum	0.9
weight_decay	0.0001
ROI number	2000
ROI positive and negative sample ratio	1:3
mini_batch	1
Step size	1000
max_iter	300

Network Model and Training. The author proposes a non-specific regression layer to improve the class specific regression layer of the original FasterRCNN, forming a non-specific FasterRCNN model. I hope to learn border regression knowledge from public datasets and directly apply it to pre annotation of individual product training images.

Firstly, the original FasterRCNN is trained on COCO, and the Resnet model in its backbone network is pre trained using ImageNet. Then, the classification layer and regression layer are trained, and finally, the overall network model is jointly trained [8]. This is done to enable the model to learn effective feature generalization ability from the COCO dataset. Change the bounding box regression layer in the trained FasterRCNN model to the non class specific regression layer proposed by the author, while keeping the parameters of other parts unchanged, and only train the non class specific regression layer on the COCO dataset. For the new non class specific FasterRCNN, tune the entire network using the COCO dataset. This is to enable the features obtained by ROIAlign to balance the ability of classification and bounding box regression.

The model trained through the above steps can be directly used for annotating product training images. RoiPooling/ROIAlign performs feature extraction in the corresponding area of FeatureMap, and normalized to 7×7 in size, with 256 channels. Then followed by an 7×7 convolutional layer, where an 7×7 convolutional kernel of the same size as the input is used, with an output dimension of 1×1 and a channel count of 1024, the convolutional layer here can be considered as a fully connected layer. Then use the convolutional kernel of 1×1 , with an output dimension of 1×1 and a channel count of 1024. The convolutional layer of this layer is also an alternative to the fully connected layer. Moreover, the output dimension is the same as the output dimension of the previous layer. This is generally done to enhance the non-linear ability of the network model, as each convolutional layer is followed by an activation function, where Relu is used as the activation function. The last layer is the softmax layer, which outputs 4 channels, which is the position parameter for bounding box regression. In addition, the hyperparameters in the model are shown in Table 2.1.

By combining non class specific FasterRCNN with Grabcut, a large number of product image samples can be generated. And used for training the overall model. The training steps are as follows:

- (1) Firstly, based on the parameters of the non class specific FasterRCNN used for pre calibration of training samples, training is carried out while keeping the backbone network and non class specific regression layer parameters unchanged, and only training the classification layer model.
- (2) Then, while keeping the parameters of the non class specific regression layer unchanged, train the classification and regression layers of the RPN network, as well as the classification layer in step (1).
- (3) Train the entire network, including the Resnet parameters in the backbone network, with only fixed non class specific regression layers. This is because the features in the backbone network are trained by COCO, and in order to better extract features from product data, it is necessary to train the backbone network parameters.

Through the above steps, the model trained using the generated samples can be used for the detection task of real product images. Moreover, the non class specific regression layer does not need to be retrained on the target dataset, which further confirms its knowledge transfer ability.

In the experiment, the MaskRCNN method from the FasterRCNN algorithm set was used, which, in addition to the FasterRCNN method, utilized the Feature Pyramid Network (FPN) and Region of Interest Alignment (ROIAlign) methods. During the training process, there was no need for segmentation prediction,

Table 3.1: Comparison of Properties of Different Datasets

Data set	Number of categories	Average training samples for each class	Does the training sample contain multiple targets	Is there a location label
COCO	80	25000	Yes	Yes
VOC	20	1350	Yes	Yes
data set	40	8	No	No

so the segmentation branch in MaskRCNN was removed and only its classification and regression branches were used. The experiment was conducted on two NVIDIA TITAN X GPUs. The initial learning rate is 0.001 and is manually adjusted during training. The momentum parameter Momentum is 0.9. Among them, conv1, conv2, conv3, conv4, and conv5 are components of the Resnet network structure.

3. Experiments and Analysis.

3.1. Dataset Introduction. The author validated the proposed method on the constructed product dataset. The training set consists of training images for a single product. Using the author's method, there is no need to train bounding box regression on the product dataset, so the constructed product dataset training images only contain category information [9]. The training images were captured from four different perspectives using two cameras, with each image containing only one product object. There are a total of 40 product categories, each of which includes 8 training images. The expanded single product image training set contains a total of 3200 training images. The test set consists of multiple product images captured using another camera, and the product positions and angles in the images are diverse. The test set consists of 400 test images, totaling 40 product categories. The proposed non class specific FasterRCNN was trained on the COCO dataset and directly applied to pre labeling of product training image data. The COCO dataset consists of 80 categories, including a large number of images, borders, and category annotations. The main difference between the product dataset constructed by the author and the COCO dataset is that the objects in the product dataset are rotated, and the training data is much less than the COCO dataset. In addition, the dataset constructed by the author only contains a single product in the training images and does not require border calibration.

As shown in Table 3.1, the differences between this dataset and the publicly available dataset were compared. Among them, there are as many as 40 categories of the dataset, and the objects can be rotated, which adds some difficulty to the recognition task. More importantly, the dataset has very few category samples and is a single object sample, without position border labels, greatly increasing the difficulty of localization and recognition tasks.

3.2. Experimental Results.

(1) *Sample Enhancement Strategy Analysis.* After extracting the product border from the training image, the Grabcut algorithm is combined to segment the product area. Because the training image contains a large area of background, if the Grabcut algorithm is directly used to segment the original training image, the segmentation effect is very unsatisfactory. Because there is no border to calibrate the background area of the image, the outermost pixel of the image is generally taken as the background. However, its area is very small, making it difficult to model the entire background. Combining the author's proposed non class specific FasterRCNN pre labeling algorithm with Grabcut algorithm for product image segmentation in the training set. Then use simple image processing methods to generate training images of multiple products for the FasterRNN model training.

When using the Grabcut algorithm for training image data generation, the upper limit of object overlap area sup is set to 10000. The author compared the performance of the trained model on the testing machine for training data generated by different overlapping areas sc. As shown in Figure 3.1, when generating training images, when the overlap area is 0, that is, when the product distance is far, the effect is not good. Because the product distance is far, it is difficult to see folding in the training data, making it difficult for the network to receive training on folding, so the testing effect is relatively poor. When the overlapping area is 6000, the recall and precision of the model reach 93.8% and 96.3%, respectively, and the testing effect is the best. When

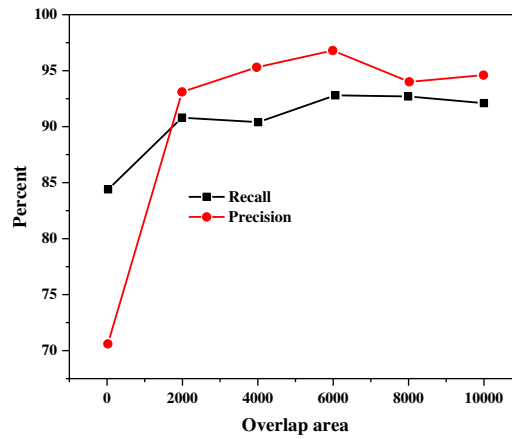


Fig. 3.1: Detection results of different fusion strategies

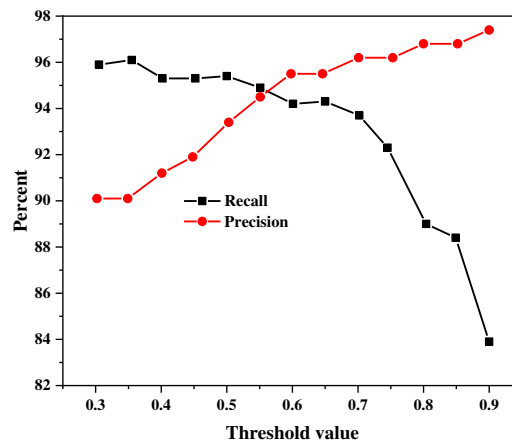


Fig. 3.2: Detection performance of different probability thresholds

the overlap area is too large, the large coverage between products in the training image tends to mislead the network into misidentification and reduce the testing effect.

(2) *Analysis of the effects of each part of the model.* When the model identifies and locates, it will output the probability of its corresponding category for each region. When arranging the model, it is usually necessary to normalize the probability, filter out predictions with low probability, and retain results with high probability. Therefore, the impact of different thresholds on the recall and accuracy of the model was analyzed, as shown in Figure 3.2. In general, the higher the probability threshold, the higher the accuracy, and the lower the recall. The lower the probability threshold, the lower the accuracy, and the higher the recall rate. In Figure 3.2, when the probability threshold is 0.3, the model can achieve high accuracy and recall at the same time. This is because the model has a high probability of predicting categories, and a low threshold has little effect on it. The model has strong predictive ability. In order to balance accuracy and recall, the author determined a probability threshold of 0.7, which is a recall rate of 93.8% and an accuracy of 96.3%.

As shown in Table 3.2, by analyzing various parts of the model, the proposed sample enhancement method combined with Grabcut improved the detection recall by over 40% and accuracy by 30%. In order to improve the accuracy of multiple product detection, the author proposes a re identification layer, which corrects the candidate regions after classification and regression through the bounding box regression layer and inputs them

Table 3.2: Analysis of the effectiveness of the methods proposed by the authors

Method	Recall (%)	Precision (%)
Non class specificity FasterRCNN	58.14	50.96
Non class specificity FasterRCNN +Grabcut	90.43	92.76
Non class specificity FasterRCNN++Grabcut + Re identification	93.81	96.3

Table 3.3: Performance Comparison of Multiple Product Identification Methods

Method	Recall (%)	Precision (%)
SIFT	33.47	20.30
VGG16	41.21	29.38
VGG19	36.50	26.25
Xception	58.50	42.50
Resnet	58.92	43.75
Author's method	93.80	96.29

into the classification layer again. After correction by non-specific bounding box regression layers, classification errors caused by imprecise candidate regions can be effectively avoided. When using the re identification layer, the recall rate is increased by 3% and the precision rate is increased by 4% compared to when not using the re identification layer.

(3) *Comparison of detection models.* Because the proposed non class specific FasterRCNN can detect a single product border, the main problem when applied to the detection of multiple products is that it can interfere with recognition when there are areas of other products within the product border. The bounding box regression of FasterRCNN, which is not specific to category, is not affected by multiple products. Therefore, when using training data from multiple products generated by the author for training, only the parameters of the classification layer are trained while keeping the parameters of the non class specific regression layer unchanged.

Through the proposed image enhancement technology, the detection of multiple products has been achieved, and the non class specific regression layer is only trained on public datasets, and regression knowledge has been learned. Moreover, there is no need for further training when transferring to product image detection. Multiple products can also be well positioned in situations with local occlusion, different perspectives, or across scenes. However, there is no perfect system in the world, and when there is severe occlusion, missed detections can also occur. Among them, the product "Hanshan Pepper and Pickled Vegetable" was not detected due to being partially occluded. The author quantitatively validated the proposed method in the constructed product dataset. Due to the author's intention to address data bottlenecks. The constructed training dataset only has category labels and no border calibration. In this case, traditional image detection methods generally use unsupervised features to calculate local features of the retrieved image and perform similarity matching with the features of the image in the training set. The currently most effective deep learning methods, such as VGG16, VGG9, Xception, and Resnet, are generally regarded as multi label classification tasks for recognition.

The author compared these methods and as shown in Table 3.3, the performance of SIFT and other currently optimal deep learning methods is significantly lower than the method proposed by the author [10]. On the one hand, SIFT does not distinguish background features, which leads to background features affecting the matching effect; On the other hand, it is an unsupervised artificial feature that is not as effective as supervised methods in recognition, and the product packaging will have severe reflection, which also results in lower feature performance. Other deep learning methods can be extended from a single product training image to multiple product training images. Not learning the distinguishing information when multiple products are similar, and also not distinguishing background features, resulting in low recognition rate. Some deep learning models, such as VGG16 and VGG19, have similar performance to SIFT, this is because the cross task recognition task of training images from a single product to identifying and locating multiple products results in low performance

of deep learning models. And this method proposes a sample annotation and sample enhancement method that does not require training on the target dataset. It can use the training image of a single product to learn the distinguishing information of multiple products, serving as a bridge across tasks and greatly improving performance.

4. Conclusion. The author designed and implemented multiple product recognition and localization methods based on improved FasterRCNN and GrabCut. Regarding the above issues. Firstly, the author cleverly improves FasterRCNN by proposing a non class specific bounding box regression layer that can learn prior knowledge of object positions using only public datasets for training, without the need for retraining on the target dataset, and uses it for pre calibration of object positions in commodity training data. Secondly, based on the pre labeled position information of non class specific FasterRCNN products, combined with the GrabCut interactive segmentation method, accurate segmentation can be performed on the training images of individual products. By fusing individual product images from multiple categories, a large number of training images containing multiple products and labels for product positions are generated for model training. At the same time, it solves the problems of limited data volume, no object position annotation, and the expansion of single product image recognition tasks to multiple product recognition tasks.

REFERENCES

- [1] B. Wu and C. Zheng, Pattern recognition of holographic image library based on deep learning, *Journal of Healthcare Engineering*, 2022 (2022), p. 1–9.
- [2] J. Gao, An image recognition method for speed limit plate based on deep learning algorithm, *International Journal of Information and Communication Technology*, 20 (2022), p. 216.
- [3] T. Takayama, T. Yashiro, S. Sanada, T. Katsuragi, and R. Sugiura, Evaluation of detection accuracy of image recognition for automatic counting of rice planthoppers captured on sticky boards, *Agricultural Information Research*, 30 (2022), p. 174–184.
- [4] J. Xu, S. Zhou, F. Xia, A. Xu, and J. Ye, Research on the lying pattern of grouped pigs using unsupervised clustering and deep learning, *Livestock Science*, 260 (2022), p. 104946.
- [5] S. Liu, R. Hu, J. Wu, X. Zhang, J. He, H. Zhao, H. Wang, and X. Li, Research on data classification and feature fusion method of cancer nuclei image based on deep learning, *International Journal of Imaging Systems and Technology*, 32 (2021), p. 969–981.
- [6] L. Wang, Q. Kou, Q. Zeng, Z. Ji, L. Zhou, and S. Zhou, Substation switching device identification method based on deep learning, in *2022 4th International Conference on Data-driven Optimization of Complex Systems (DOCS)*, 2022, pp. 1–6.
- [7] W.-x. WANG, J. JIA, K. GUI, and J. GAO, Research on automatic compliance checking method based on bim and ontology technology for architectural drawing, *Journal of Northeastern University (Natural Science)*, 43 (2022), pp. 1346–1353.
- [8] P. Rani, S. Kotwal, J. Manhas, V. Sharma, and S. Sharma, Machine learning and deep learning based computational approaches in automatic microorganisms image recognition: Methodologies, challenges, and developments, *Archives of Computational Methods in Engineering*, 29 (2021), p. 1801–1837.
- [9] Z. Wang, H. Chen, Q. Zhong, S. Lin, J. Wu, M. Xu, and Q. Zhang, Recognition of penetration state in gtaw based on vision transformer using weld pool image, *The International Journal of Advanced Manufacturing Technology*, 119 (2022), p. 5439–5452.
- [10] J. Sun and X. Yuan, Application of artificial intelligence nuclear medicine automated images based on deep learning in tumor diagnosis, *Journal of Healthcare Engineering*, 2022 (2022), p. 1–10.

Edited by: Zhigao Zheng

Special issue on: Graph Powered Big Aerospace Data Processing

Received: Jan 4, 2024

Accepted: Feb 26, 2024



THE CONSTRUCTION AND APPLICATION OF RESIDENTIAL BUILDING INFORMATION MODEL BASED ON DEEP LEARNING ALGORITHMS

SHUANG ZHAO* AND YU YANG†

Abstract. In order to explore the construction and application of building BIM models, the construction industry is actively exploring a method that can quickly reshape the 3D information model of existing buildings in the wave of digital twins and smart cities. Starting from the perspective of deep learning 3D object detection algorithms, the author starts with the generation of large-scale building datasets and the theory of point cloud deep learning, analyzes the input data types required for point cloud deep learning frameworks, and focuses on the creation process of 3D bounding boxes and 3D point clouds for various building components. The author compares different point cloud datasets with the same data structure and implements an object detection algorithm based on the ScanNet dataset, furthermore, a feasible technology route for automatic generation of BIM models from 3D point clouds based on deep learning is integrated. Through this technology route, the trained neural network can input unknown building 3D point clouds and output BIM model parameters.

Key words: Building dataset, 3D point cloud, Deep learning, BIM

1. Introduction. BIM (Building Information Model), also known as Building Information Model, contains the geometric information, performance, and functionality of all components in the model. It encompasses all information throughout the entire lifecycle of a building project into a single model [1]. In recent years, BIM has gradually been applied to the design and construction of modern buildings. At the same time, due to the development of new technologies, people have begun to study the parameterized information model of ancient buildings and apply it to the design and construction of some antique buildings.

With the transition of urban renewal from the "incremental era" to the "stock era", the relationship between data on built environments and corresponding human behavior data has become increasingly close. Big data demonstrates a people-oriented perspective, timely and real-time information, and fine resolution spatial dynamics. In the face of data research on built environments such as remote sensing images and street view images, after several years of image analysis research in the field of land planning, the semantic labels of high-resolution (VHR) remote sensing images assign a category task to each pixel in the image, including land use planning, infrastructure management, and urban expansion detection. The use of deep learning intervention has been widely adopted. With the deepening of deep learning technology research, street view images have gradually become an important data source for quantitative research on built environments due to their numerical characteristics and the accompanying geographical location information [2]. Spatial quantitative evaluation based on a humanistic perspective has become an important research direction, including the detection of street style features, environmental features, building materials and functions, semantic segmentation of building facade components, and the relationship between street scenery environment.

Illegal construction of structures is one of the important works in the law enforcement and supervision of natural resources. Illegal buildings not only affect urban planning and urban beauty, but also bring trouble to the management of state-owned land. All departments need to monitor and control the illegal construction information, and need a large number of people to obtain first-hand information with human and material resources. The traditional illegal construction information investigation is mostly conducted by relevant departments organizing field investigators to conduct field investigation and field comparison. If newly built or expanded buildings are found, relevant data should be transferred in the land management system in time to

*School of Architecture and Art Design, Hebei Academy of Fine Arts, Shijiazhuang, Hebei, 050800, China

†School of Architecture and Art Design, Hebei Academy of Fine Arts, Shijiazhuang, Hebei, 050800, China (Corresponding author, duoduoduo0620@163.com)

verify whether there is illegal construction. Although this method can obtain illegal construction information, it is time-consuming and laborious, and has low timeliness. With the continuous development of remote sensing technology and application, the use of remote sensing images for land resources and national conditions dynamic remote sensing monitoring has been a part of the annual law enforcement inspection in large and medium-sized cities. Law enforcement inspection based on remote sensing technology has good objectivity, but the manual identification of illegal land use and illegal construction is still the main method. In recent years, with the increasing maturity of artificial intelligence deep learning algorithms, such as using video images to carry out statistical analysis of road traffic flow, target recognition algorithm to carry out cargo ship operation, and semantic segmentation to carry out natural resources survey, etc.

Village building environment with the development of economic construction, the traditional residential buildings are increasingly suffer gradually eating, the style of the village is gradually alienation, fortunately, this problem has gradually get attention, but for the protection of traditional village development, how to evaluate the village residential architectural features, classified statistical management, is indeed a very necessary and difficult work. For the protection of traditional villages, an urgent need to the number of village residential buildings, style, building quality, building height, and other information for quantitative evaluation and analysis, and improve the village planning development management, currently in the practical village planning work, which also put forward specific requirements, but the present for village building information statistics mainly through the way of artificial field investigation, after multi-directional residential photos, artificial interpretation of its architectural characteristics. Such a way, on the one hand, is easy to be limited by the traffic, climate and terrain factors of the local villages, which brings inconvenience to data collection, but also increase a lot of research costs; On the other hand, due to the way of manual interpretation, it is bound to bring some uncertain changes due to the subjective factors such as the subject background, life experience, mood and emotion, and bring certain disturbance to the result of the definition of architectural style. The continuous emergence of new technologies and new data provides a rich data basis for more detailed spatial quality research. At the same time, applying intelligent technologies such as machine learning and edge computing to various industries is a concave solution that conforms to the development of The Times. Such a study first requires a data collection process to collect the required data that collects critical architectural imaging data, often relying on field investigations. Such a high level of labor-intensive and time-consuming work makes a large-scale assessment of architectural features extremely difficult. In this regard, the collection and integration of architectural landscape data in an effective way remains a challenge for current academic research.

Digital twin cities can make forward-looking predictions about people and things in the city to a certain extent, with the premise of reshaping high-precision and multi coupled Building Information Modeling (BIM). However, in the initial digital modeling process of BIM, very few existing buildings have complete and accurate information data. In the field of BIM model generation for existing buildings, current methods focus on preprocessing the 3D point cloud data of existing buildings, extracting point cloud features through prior knowledge and corresponding algorithms (such as Hough transform, RANSAC (RandomSampleConsensus), and then generating BIM models through certain human-computer interaction processing [3,4]. This involves a large amount of manual work, which is time-consuming and subjective and prone to errors, in addition, the generalization performance of 3D reconstruction methods based on prior knowledge is poor, and the point cloud features extracted by traditional algorithms cannot be optimized end-to-end to obtain the global optimal solution. In the process of generating BIM models from 3D point clouds, the segmentation and localization of point cloud data are the most critical steps, and automatic implementation of this step is often difficult. Going deep into the basics, the difficulty lies in the fact that BIM models are composed of various types of components, and point clouds are a whole. In the process of automatically creating BIM models, component identification, localization, and modeling need to be synchronized. With the development of deep learning theory, deep neural networks have been widely adopted in different industries and have achieved good performance in computer vision tasks. The author explores the combination of deep learning and the generation of existing building BIM models. Combining the input data required for training point cloud neural networks, a set of point cloud datasets (SYNBIM) and corresponding creation methods are proposed for conventional building components that can be applied to 3D object detection tasks, by comparing the data storage formats of the SYNBIM dataset and the Scan Net point cloud dataset proposed by the author, the input and output data of the point cloud

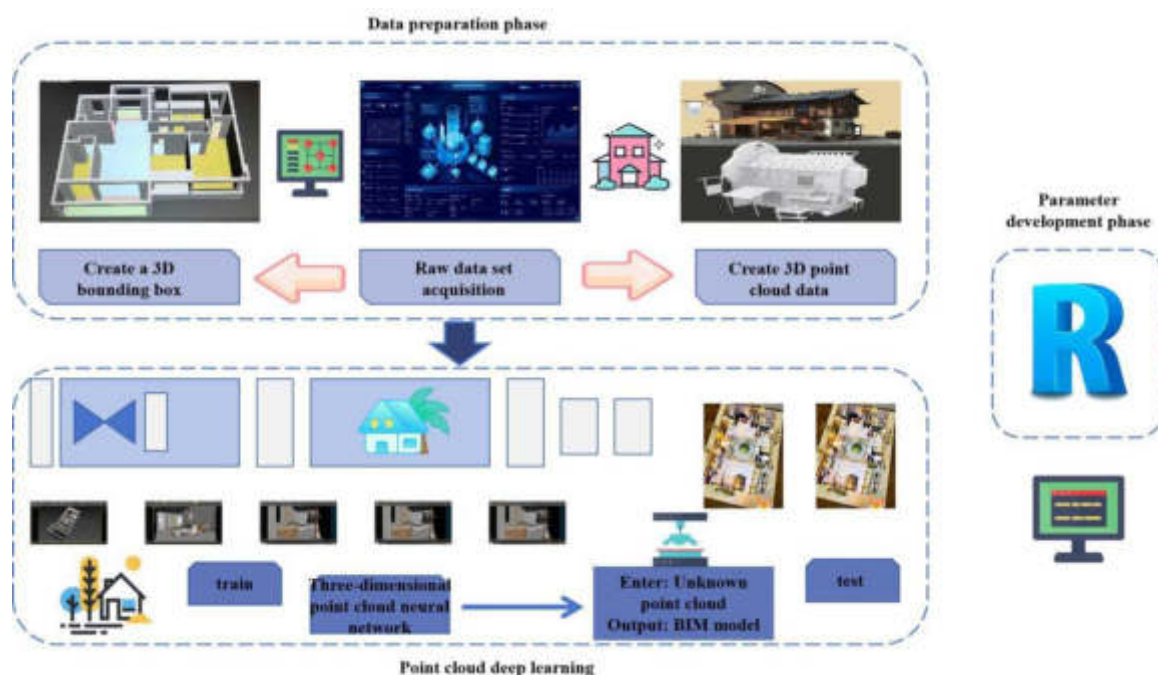


Fig. 2.1: Technical route of BIM model for automatic generation of 3 D point cloud based on deep learning

neural network framework during the testing phase were analyzed, and it was proved that the trained neural network can achieve "input unknown point clouds and output BIM model parameters" [5].

2. Methods.

2.1. Technical roadmap. The artificial intelligence algorithm is the most ideal method for reshaping BIM models in 3D and point clouds, but currently, the theory of point cloud processing and training data resources are not yet mature. The technical roadmap proposed by the author based on these two points is shown in Figure 2.1. In the data preparation stage, the SUNCG dataset 9, released by the Computer Vision and Robotics Research Group at Princeton University in 2017, was selected as the original dataset. This is a manually synthesized large-scale 3D scene virtual dataset with dense volume annotations. However, the original dataset is not mature enough for 3D bounding box annotations of conventional components in the field of civil engineering, such as walls, doors, windows, etc. In order to achieve the task of deep learning 3D object detection, the author selected five types of building components, namely "walls", "doors", "windows", "floors", and "ceilings", as the research objects. With the help of the Open3D 3D semantic library, each type of component was processed separately and corresponding hypotheses were proposed during the processing process. Finally, instance level 3D bounding boxes were generated as the labeled information for supervised learning. Then, combined with the camera pose estimation method, multiple indoor observation points are generated and stereo cameras are placed to capture multiple depth maps and two-dimensional images. These images with depth information are then synthesized into point cloud data, which can be used as input to the neural network. In terms of point cloud deep learning, by combining the input and output data of the point cloud neural network framework Votenet, and analogizing the data storage structures of the SYNBIM dataset and ScanNet dataset, a trained Votenet model is used to test the ScanNet point cloud scene, and the 3D object detection result is the BIM model [6]. For the five types of conventional building components, the data output by neural networks is well-established, and the parameter information of the components can be obtained through relevant geometric operations. In theory, corresponding BIM models can be created through Revit secondary development programming.

2.2. Obtaining the original dataset. The original dataset provides 45622 sets of 3D virtual scenes containing multi story buildings. Each 3D scene is saved in a house.json file. Based on the content of the JSON file, select a hierarchical processing method to generate the target 3D bounding boxes for each scene, and use separate processing for five different objects. With the help of Open3D, an open-source library that supports 3D data processing, and its advantage of customizable operation functions, the original dataset can be read and operated on.

2.3. Deep learning principles.

(1) *The fully convolutional neural network.* Full convolutional neural network (fully convolutional networks, FCN) uses convolutional neural network to realize the transformation from image pixel to pixel category. It can realize the classification of pixel level and solve the problem of semantic level image segmentation. Therefore, it is more suitable for the extraction of building map spots. The traditional convolutional neural network (convolutional neural networks, CNN) usually squashes the original two-dimensional matrix into one-dimensional at the last fully connected layer, losing spatial information, and finally trains to output a label. The building change extraction task is not only to know the types of objects contained in the image, but also to divide the location of different objects. Compared with other CNN, the difference is to remove the original CNN last fully connected layer, using deconvolution layer of the last convolution layer feature map sampling, make it back to the input image of the same size, the last pixel calculation classification loss, so can produce a prediction for each pixel, retain the spatial information in the original input image. Since high-resolution remote sensing images have a lot of detailed information, the underlying low-resolution semantic features directly classify the images, which will produce obvious errors. Therefore, FCN upsamples the results of different pooling layers, forms the optimization output, and then combines them with the downsampling to obtain the final classification results.

(2) *Add deep learning features.* In the sample training, the deep learning algorithm usually carries out a lot of iterative calculation based on the texture and spectral information of the sample, but there will be false identification for the structures with different shapes and different uses. On the basis of the identification and extraction of building information by deep learning algorithm, plus the features of NDVI and SAVI, it can effectively improve the differentiation degree of woodland, bare land, vegetation and building, and then improve the extraction accuracy of building changes. Both indices are calculated as follows:

$$N = \frac{B_{NIR} - B_R}{B_{NIR} + B_R} \quad (2.1)$$

$$S = \frac{B_{NIR} - B_R}{B_{NIR} + B_R + L} (1 + L) \quad (2.2)$$

where N indicates the vegetation index NDVI; BNIR represents the near-infrared band; BR red light band; S is the regulated soil brightness index SAVI; L is the soil regulation coefficient, the value range is 0~1, L=0 means the vegetation coverage is 0, L=1 indicates the soil background effect is 0. The texture features, spectral features and the above two exponential features of the image are combined into multi-source index, and the optimal training accuracy is obtained by the continuous iterative calculation of the image model within the sample range.

CNN can be applied in scene classification and image classification. LeNet is one of the earliest CNN structures, which is mainly used in the character classification problem. Since the convolution operation is used in the program, not only the features of the picture can be extracted, but also the convolution operation maintains the spatial relationship between the pixels. In the CNN, a filter is used as a feature extractor, and the matrix obtained by convolution is called a "feature graph". When selecting a specific CNN, the image characteristics of the target object, such as the difference between rural and urban buildings, and the situation of coarse-grained buildings. Because the real world classification problems are non-linear, and convolution operation is linear operation, so when using CNN to solve, must use a nonlinear function such as ReLU (or other nonlinear functions, such as Tanh and Sigmoid,) to add the results of the nonlinear properties, then adopt the form of downsampling, extract the features after the ReLU processing, or extract the element average or extract the maximum value, so as to reduce the dimension of the feature graph while maintaining the important information

Table 3.1: Part of Wall Ground Truth Data Styles

xc	yc	zc	x_size	y_size	z_size	yaw
7.46×10^{-2}	2.69×10^0	1.4	5.26	7.86×10^{-2}	2.73	$1.58 / \times 10^0$
9.27×10^0	1.38×10^1	1.39	5.04	1.03×10^{-1}	2.73	$1.58 / \times 10^0$
1.51×10^1	1.54×10^1	1.38	2.15	1.03×10^{-1}	2.78	$1.58 / \times 10^0$
4.68×10^0	1.27×10^1	1.35	2.71	9.36×10^{-2}	2.73	$1.58 / \times 10^0$
1.73×10^1	1.13×10^1	1.35	4.06	1.06×10^{-1}	2.7	$1.75 / \times 10^{-3}$
9.3×10^0	2.68×10^0	1.37	4.9	1.04×10^{-1}	2.73	$1.58 / \times 10^0$
1.38×10^0	1.63×10^1	1.39	2.41	9.56×10^{-2}	2.81	$3.45 / \times 10^{-3}$
3.62×10^0	5.27×10^0	1.37	6.61	9.86×10^{-2}	2.71	$3.15 / \times 10^0$
...

of the picture. Finally, the fully connected layer (multilayer perceptron) is combined together and classified through the whole layer using a softmax activation function. To is a vector with a value of 0-1 to judge picture classification by the probability value.

In recent years, deep learning methods, especially CNN performance in various computer vision tasks have gone beyond traditional methods, such as its excellent research results in target detection, semantics and image segmentation. The method of labeling image pixels with labels is based on the semantics in the image, that is, the algorithm will exist in the image, such as cars, trees or buildings as semantics from the extraction of the whole image, and each semantics. Moreover, in the field of computer vision, there is a large amount of research on the various modules used in convolutional neural networks that utilize the concept of "per object classification". These modules, such as convolution and pyramid pooling, improve the algorithmic performance for semantic segmentation tasks. In recent years, with the significant improvement of chip processing power (such as GPU units), the significant reduction of the cost of computing hardware, and the significant progress of machine learning algorithms, deep learning has made rapid progress in the field of image recognition, thus greatly improving the processing power of computers.

3. Results and Analysis.

3.1. Creating 3D bounding box walls. The manifestation of walls in the field of multi story and multi room indoor scene research is very complex. In order to prevent the calibration requirements of model bounding boxes in 3D object detection, a set of paradigms is needed to define the representation of wall objects, therefore, the author proposes the following assumptions when defining wall objects: (1) Shortest wall assumption: When a long wall and a short wall intersect in the same floor space, the long wall is split into two walls at the intersection; (2) One wall corresponds uniquely to a 3D bounding box. The bounding boxes extracted directly from the geometric information of the walls in the original data through programming, where different walls are marked with different colored bounding boxes. However, based on the above assumptions, there are several issues: Firstly, multiple long walls are not broken when encountering short walls; Secondly, there are multiple instances where the same wall is repeatedly calibrated by multiple bounding boxes, resulting in severe box overlap; Finally, there is no height difference between different bounding boxes, which may hide walls with repeated calibration.

By calling the open3d library functions through programming and customizing relevant functions to read and operate on the raw data, a set of bounding boxes is generated. The single-layer building walls are processed to meet the above assumptions.

The data of the wall bounding box is shown in Table 3.1, where a single row of data represents information about a bounding box, and seven parameters represent the coordinates of the three center points of the box in the overall coordinate system, the length, width, height of the box, and the deflection angle along the height direction. These generate the calibration parameters for the wall, and the other four components are processed and have the same data structure as the wall bounding box [7].

3.2. Creating 3D bounding boxes - doors and windows. During the process of processing raw data, doors and windows encountered the problems shown in Table 3.2 and Figure 3.1. Through investigation, it

Table 3.2: Initial Boundary Box Problems and Solutions for Doors and Windows

Initial problem	Initial error rate/%	Solution	Final accuracy rate/%
Thickness greater than wall	23.48	Detect the wall where it is located	99.18
Detaching from the overall building	9.7	Recalibrate the detection&delete it	99.8
Repeated calibration	19.87	Merge	98.18

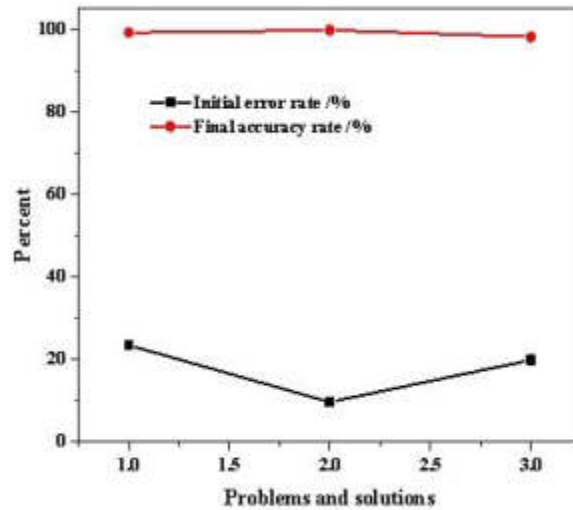


Fig. 3.1: Error rate and final accuracy rate of initial bounding boxes for doors and windows

was found that some buildings had small doors and partially open windows [8]. According to the minimum bounding box theory, the thickness of the bounding box obtained by directly extracting data from partially open doors and windows exceeded the thickness of the wall.

3.3. Creating 3D bounding boxes - Floors and ceilings. The author defines the following rules when creating board class bounding boxes in the program: (1) Board objects cannot cover two or more rooms; (2) For multi story buildings, the boundary boxes of two floors or two ceilings cannot overlap, but overlapping is allowed between the two types of boxes of floors and ceilings; (3) The edges of at least three edges of the floor and ceiling bounding boxes are in contact with the edges of the wall.

3.4. Creating 3D bounding boxes - entire building . For single story buildings, create 3D bounding boxes for each target instance using the above methods to provide marker information for deep learning 3D object detection training tasks; For multi story buildings, the author first separates the information of each layer separately during the programming process of reading and manipulating the raw data, then sequentially processes and obtains the 3D bounding boxes of each layer's target instances. Finally, the processed multi story data is combined together to complete the processed multi story 3D bounding boxes [9].

3.5. Building Datasets - Creating 3D Point Clouds. In current practical applications, the digital acquisition based on 3D laser point clouds, which uses laser scanners to scan indoor and outdoor buildings in 3D, is a common method to obtain building point cloud data [10]. However, LiDAR has problems such as high cost, mirror black holes, and low lifespan, the method of generating depth images through stereo vision cameras and synthesizing point clouds is the future development direction. Depth images, also known as distance images, refer to images that use the distance values from the image collector to each point in the scene as pixel values. he author takes a long-term perspective and chooses to place a stereo vision camera at the virtual observation

Table 3.3: Dataset generation process and partial result presentation

Classification	subclass	3D Scene	3D Scene	3D Scene
		#1 Layer	#2 Layer	#3 Layer
Input	#Number of Layers/# Number of	1/1596/	2/768/	3 /549 /
	Depth Maps/#Number of Point Clouds/	3041.407	1472.243	2685.806
Number	Number of rooms/area/m2	10/180.32	6/142.38	21 /278.52
	Boundary Box #	28/9/3	11+9	23+23+26
	Wall/# Door/#		/2+1	/4+4+4
	Window		/1+4	/4+4+3
	Boundary Box #	10/9	3+2/2+2	6+6+6
	Floor/# Ceiling			/3+3+3
	Downsampling/Original	29987/	18785/	36429/
	Point Cloud/k	16849.562	12492.263	19443.005
Running time/s	All bounding boxes	12.45	10.6	17.9
	Depth image+	654.76+	445.67+	953.16+
	color rendering	99.46	67.42	137.67
	3D point cloud	1843.13	1403.78	1930.19

points contained inside each building in the SUNCG dataset and generate corresponding depth images. The obtained depth images are 16 bit PNG format files, which can be used to generate 3D point clouds through relevant registration and registration algorithms.

Stereoscopic vision cameras need to capture objects from different observation points, and the depth map information obtained from different stations has their own independent coordinate systems [11]. After unifying the coordinate system, data fusion needs to be carried out. The basic formulas involved are:

$$\begin{cases} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta z \end{bmatrix} + R_z(\alpha)R_z(\beta)R_z(\gamma) \begin{bmatrix} x \\ y \\ z \end{bmatrix} \\ R_z(\alpha)R_z(\beta)R_z(\gamma) = \begin{bmatrix} \cos\alpha, -\sin\alpha, 0 \\ \sin\alpha, \cos\alpha, 0 \\ 1, 1, 1 \end{bmatrix} \\ \begin{bmatrix} \cos\beta, 0, -\sin\beta \\ 0, 1, 0 \\ -\sin\beta, 0, \cos\beta \end{bmatrix} \begin{bmatrix} 1, 0, 0 \\ 0, \cos\gamma, \sin\gamma \\ 0, -\sin\gamma, \cos\gamma \end{bmatrix} \end{cases} \quad (3.1)$$

In the formula: (X, Y, Z) represents the coordinates after point cloud registration; (x, y, z) are the original coordinates of the point cloud; $\Delta x, \Delta y, \Delta z$ is the translation parameter; α, β, γ is the rotation parameter[12]. By creating bounding box marker information and point cloud data, the dataset required for training the point cloud neural network was obtained, consisting of 45622 sets of large-scale 3D scene data. The hardware information and related calculation time used in the data generation process are shown in Table 3.3 [13].

3.6. Point Cloud Neural Network. At present, the application of deep learning to input 3D point cloud data into neural networks is still an open problem, and there are mainly three methods used: (1) Projecting point cloud data onto a 2D plane, which does not directly process 3D point cloud data, but first projects the point cloud to certain specific perspectives before processing, such as MUCNNL; (2) Dividing point cloud data into voxels with spatial dependencies, this approach involves segmenting three-dimensional space, introducing spatial dependencies into point cloud data, and then using methods such as 3D convolution for processing. However, the computational complexity of 3D convolution is relatively high, such as PointCNN; (3) Directly processing the original point cloud, this method generally involves a multi-step 3D object detection algorithm.

Table 3.4: Comparison of Dataset Properties

Data set	Scale	Data sources	Object scope	Data in	Output data	Using Lingcheng
ScanNet	1513	Real scan	Indoor scenes	3D point cloud	3D bounding box	Computer Vision
Hours (recalibration)	45622	synthetic	Indoor scenes	3D point cloud	3D bounding box	Architectural reshaping

Table 3.5: Geometric parameters of target components

Component category	Geometrical parameter
Wall	Wall bottom elevation, wall top elevation, wall centerline starting and ending wall thickness
Doors and windows	Placement coordinates, door bottom elevation, and door top elevation
Floors and ceilings	Outline boundary and plate thickness

In order to deal with the perspective invariance of each target in the point cloud data and obtain accurate 3D box regression, the algorithm needs to perform multiple coordinate transformations, such as Votenet [14]. The reason for the above three input methods is that: (1) Point cloud data is a collection that is not sensitive to the order of data, and the model processing point cloud data needs to maintain invariance to different arrangements of data; (2) The target represented by point cloud data should have invariance to certain rigid transformations, such as rotation and translation. This network does not rely on any RGB information and only relies on simple point cloud geometric information [15,16]. The entire network framework is divided into two parts: The first part uses a point cloud feature extraction network to process the original point cloud data and generate seed points and growth points; The second part aims to achieve classification and localization by training the extracted growth point clusters. It can directly input point cloud data and output target 3D bounding boxes in 3D object detection tasks, achieving good results in the benchmark challenge of the real scanning dataset ScanNet. The comparison of some key parameters between the ScanNet dataset and the SYN BIM dataset created by the author is shown in Table 3.4 [17].

3.7. Parametric modeling. From the perspective of data coherence, parameterized modeling of BIM components only requires extracting model parameters to achieve model creation. Therefore, in response to the author's selection of component categories, the required parameters for creating a BIM model are summarized and listed in Table 3.5 [18].

For these five types of building components, the data output by neural networks is well-established, and the parameter information of the components is obtained through corresponding algorithm operations. Then, through Revit secondary development, corresponding BIM models can be created [19,20].

4. Conclusion. The author explores the combination of deep learning and existing building BIM model generation based on 3D object detection technology, and integrates a feasible technology roadmap for automatic generation of BIM models from 3D point clouds based on deep learning. The following important issues are discussed: Generate target category labeling information from virtual BIM models; Generate a 3D point cloud model from a virtual BIM model; The experimental analysis of the data input and output of the point cloud neural network proves that its output data foundation is complete, and a BIM model can be generated through simple parameter modeling. The implementation method of a deep learning based 3D point cloud generation BIM model proposed by the author unifies the steps of point cloud data preprocessing, component recognition, segmentation, localization, and modeling in traditional methods. Through the deep learning backpropagation mechanism, global optimization can be achieved for the entire process. The core process is the generation of 3D point cloud datasets and the implementation of 3D object detection algorithms.

REFERENCES

- [1] Hsieh, M. C., Huang, G. H., Dmitriev, A., & Lin, C. H. (2022). Deep learning application for classification of ionospheric height profiles measured by radio occultation technique. *Remote. Sens.*, 14(123), 4521.
- [2] Alam, M., Zhao, E. J., Lam, C. K., & Rubin, D. L. (2023). Segmentation-assisted fully convolutional neural network enhances deep learning performance to identify proliferative diabetic retinopathy. *Journal of clinical medicine*, 12(1),79.
- [3] Wang, H., Li, C., Zhang, Z., Kershaw, S., Holmer, L. E., & Zhang, Y., et al. (2022). Fossil brachiopod identification using a new deep convolutional neural network. *Gondwana research: international geoscience journal*88(105-), 105.
- [4] B, H. W. A., A, W. G., A, H. N., B, J. S. A., & A, M. Z. (2022). Classification of giemsa staining chromosome using input-aware deep convolutional neural network with integrated uncertainty estimates. *Biomedical Signal Processing and Control*, 71(1123), 103120-.
- [5] Sang, X., Zhou, R. G., Li, Y., & Xiong, S. (2022). One-dimensional deep convolutional neural network for mineral classification from raman spectroscopy. *Neural processing letters*457(1), 54.
- [6] Gao, F., Li, F., Fu, Y., You, S., Zhang, S., & Cao, K., et al. (2022). Marine aquaculture mapping using gf-1 wfv satellite images and full resolution cascade convolutional neural network. *International Journal of Digital Earth*, 15(1), 2047-2060.
- [7] Joshi, A., Lachure, J.,& Doriya, R. (2023). Drone security for precision agriculture by using one-dimensional convolutional neural network. *Journal of Uncertain Systems*, 16(04),24.
- [8] Vinolin, V., & Sucharitha, M. (2022). Taylor-rider-based deep convolutional neural network for image forgery detection in 3d lighting environment. *Data technologies and applications*687(1), 56.
- [9] Kyeong-Beom, P., & Yeol, L. J. (2022). Swine-net: hybrid deep learning approach to novel polyp segmentation using convolutional neural network and swin transformer. *Journal of Computational Design and Engineering*56655(2), 2.
- [10] Aljohani, N. R., Fayoumi, A., & Hassan, S. U. (2023). A novel focal-loss and class-weight-aware convolutional neural network for the classification of in-text citations:. *Journal of Information Science*, 49(1), 79-92.
- [11] Shin, H. (2022). Deep convolutional neural network-based signal quality assessment for photoplethysmogram. *Computers in Biology and Medicine*, 145(123), 105430-.
- [12] Radman, A., Mahdianpari, M., Brisco, B., Salehi, B., & Mohammadimanesh, F. (2023). Dual-branch fusion of convolutional neural network and graph convolutional network for polsar image classification. *Remote Sensing*, 15(1), 75-.
- [13] Fan, Y., Xu, B., Zhang, L., Song, J., Zomaya, A., & Li, K. C. (2023). Validating the integrity of convolutional neural network predictions based on zero-knowledge proof. *Inf. Sci.*, 625(345), 125-140.
- [14] Puneet, Kumar, R., & Gupta, M. (2022). Optical coherence tomography image based eye disease detection using deep convolutional neural network. *Health information science and systems*, 10(1), 13.
- [15] Ho, C. C., & Yu, C. W. (2022). Deep convolutional neural network based fabric color difference detection. *2022 25th International Conference on Mechatronics Technology 547(ICMT)*, 1-5.
- [16] Liu, X., Zhou, H., Wang, Z., Liu, X., Li, X., & Nie, C., et al. (2022). Fully convolutional neural network deep learning model fully in patients with type 2 diabetes complicated with peripheral neuropathy by high-frequency ultrasound image. *Computational and mathematical methods in medicine*, 2022(3325), 5466173.
- [17] Yang, X. B., & Zhang, W. (2022). Heterogeneous face detection based on multi-task cascaded convolutional neural network. *IET image processing*8967(1), 16.
- [18] Korkmaz, D., & Acikgoz, H. (2022). An efficient fault classification method in solar photovoltaic modules using transfer learning and multi-scale convolutional neural network. *Eng. Appl. Artif. Intell.*, 113(13), 104959.
- [19] Liu, L., Prodanovi, M., & Pyrcz, M. J. (2023). Impact of geostatistical nonstationarity on convolutional neural network predictions. *Computational Geosciences*, 27(1), 35-44.
- [20] Yang, E., Kim, C. K., Guan, Y., Koo, B. B., & Kim, J. H. (2022). 3d multi-scale residual fully convolutional neural network for segmentation of extremely large-sized kidney tumor. *Computer Methods and Programs in Biomedicine*, 215(4), 106616-.

Edited by: Zhigao Zheng

Special issue on: Graph Powered Big Aerospace Data Processing

Received: Jan 5, 2024

Accepted: Feb 26, 2024



HUMAN-COMPUTER INTERACTION INTERFACE DESIGN IN THE CAB OF NEW ENERGY VEHICLES

YANXUE ZHANG^{*}; NANMEI ZHANG[†]; HUI YANG[‡] AND YANYU REN[§]

Abstract. A new energy vehicle cab control platform based on human-computer interaction is proposed from control device design and control signal acquisition design. A driving mode conversion algorithm based on the threshold is combined with the existing intelligent vehicle driving mode conversion. The MATLAB and Python hybrid program is used to realize the simulation experiment of the driving control algorithm of new energy vehicles. The system can conduct interactive simulations of various typical driving scenarios. Data acquisition and dynamic display functions can present the operating status of new energy vehicles most directly and efficiently. The system can integrate and intelligently process vehicle driving data, road environment data and external data from multiple dimensions of the "vehicle-road-network". The experimental results show that the system enhances the user experience of new energy vehicles and improves the efficiency of human-computer interaction.

Key words: New energy vehicles; Data accommodation; Autonomous collection; Human-computer interaction; Automatic driving

1. Introduction. In the face of energy shortages in all countries, the development of new energy vehicles has become the future direction. Because its energy supply system is different from conventional vehicles, it belongs to a new stage of transformation from electromechanical equipment to electronic equipment, so the design of human-machine interfaces for electric vehicles will also change. For example, the dashboard occupies less and less area, rather than being limited to the conventional form of multiple dashboards combined [1]. Drive-by-wire technology the instrument display architecture of electric vehicles is more diverse. Electronic LCDS can also be used to replace conventional vehicle dashboards. Therefore, human-computer interaction is a significant work when the vehicle is moving. The instrument panel interface of the electric vehicle contains the speed display, remaining power display, mileage display, and a variety of driving state displays [2]. Reasonable arrangement and design of these information interfaces is the key to the interactive interface experience of the instrument panel of electric vehicles. The instrument panel is the essential carrier and communication window for the communication between people and vehicles while driving electric vehicles, and its use experience directly affects whether the information transmission between people and vehicles is friendly, visible, controllable and operable [3]. To achieve the best user experience is the ultimate goal of human-computer interface design. The user interaction experience is a way to take the user's needs as the key and then translate them into the product image. It needs to meet the following points: the human-machine interface is easy to use, and you can learn to use it without particular research; The interface elements and shape model can intuitively reflect the functional characteristics and operation mode of the system, which can conform to the user's natural usage habits. The graphic logo used in the interface should be consistent with the standardization and beauty of the interface elements. The operation and display of the interactive interface should be coordinated, and the human-machine interaction design can be differentiated for different user types. This provides precise operational instructions to various users. At the same time, it can avoid the wrong operation and ensure the safety of the man-machine interface. This paper first studies the trajectory tracking control algorithm of autonomous driving of new energy vehicles. Construct a new energy vehicle driving simulation experiment

^{*}School of Intelligent Manufacturing, Anhui Wenda University of Information Engineering, Hefei, 231201, China (Corresponding author, btboa005@163.com)

[†]School of Intelligent Manufacturing, Anhui Wenda University of Information Engineering, Hefei, 231201, China

[‡]School of Intelligent Manufacturing, Anhui Wenda University of Information Engineering, Hefei, 231201, China

[§]School of Intelligent Manufacturing, Anhui Wenda University of Information Engineering, Hefei, 231201, China

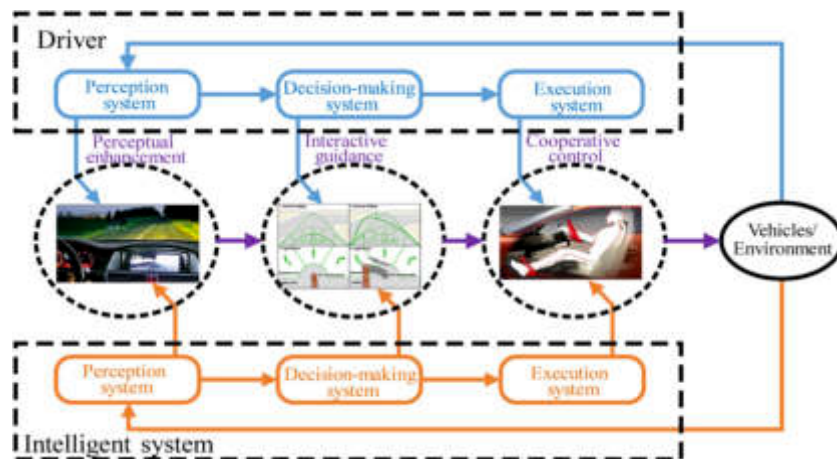


Fig. 2.1: Hardware architecture of human-computer interaction system for autonomous driving of new energy vehicles.

platform based on multi-modal information [4]. Under extreme conditions, it will lay a foundation for vehicle performance analysis and human-vehicle-environment interaction research.

2. Human-machine interaction requirements for driving new energy vehicles.

2.1. Physical Requirements. The human-computer interaction platform for intelligent management and control of new energy vehicles adopts a hierarchical structure composed of three levels: vehicle terminal, road test equipment/edge control unit, and intelligent management and control cloud platform [5]. Build an intelligent network-connected vehicle cloud platform that integrates "human-vehicle-road-network-cloud." Multi-mode communication networks such as 5G are supported on the network to achieve high-speed, low-latency data access and information transfer between vehicles, roadside devices and the cloud. Realize real-time scheduling and network management in specific unmanned driving application scenarios to ensure network security [6]—construction of traffic network architecture system to realize vehicle-road cooperation. Achieve effective vehicle-device-data collaboration. To realize the central scheduling and multi-objective optimal control for all sections and regions.

The comfort problem in the interactive simulation system of driver operation is systematically studied based on the actual vehicle driving situation. While ensuring the accuracy of data signal acquisition as much as possible, people pay more attention to the actual driving experience [7]. The steering wheel and seats are the same material as the vehicle. The clutch, throttle, and brakes are identical to the actual vehicles. In the autonomous driving mode, the design of the autonomous driving mechanism and the transformation of the driving mode are mainly studied [8]. The overall architecture composition of the system is shown in Figure 2.1 (the picture is quoted in IET Intelligent Transport Systems, 2019, 13(6): 960-966).

The platform operating equipment operates the manual driving mode, and the obtained detection information is transmitted to the signal acquisition board [9]. Through the collection and processing of the data, the data is finally transmitted to the computer. In the process of autonomous driving through the operation interface, the vehicle can dynamically adjust the driving path during the driving process. These two driving modes can complete the vehicle's driving simulation, laying a solid foundation for future intelligent research.

2.2. Functional Requirements.

2.2.1. Data integration and intelligence services. Eventually, multiple industry interconnection standards related to the vehicle will be formed, covering all scenarios of the entire intelligent network travel. Standardize information exchange among all links of the travel ecosystem, such as vehicles, traffic, roads, first aid, and meteorology [10]. To create an autonomous and controllable industrial environment for intelligent travel.

Multi-source information is integrated by establishing multi-source information such as road equipment status, road traffic events, and road traffic participants. The vehicle-cloud communication protocol consists of MQTT and TCP channels. Through the way of business reservation, it can realize dynamic lane planning, traffic control information, traffic speed limit information, real-time traffic flow prediction, dangerous road condition reminder, real-time status of signal light, future status of signal light, meteorological reality, conventional weather forecast, catastrophic weather warning, minute level precipitation forecast, parking lot information and other functions.

2.2.2. Real-time monitoring of road network. The lane and lane-level congestion operating conditions on both sides of the road are collected. Collect abnormal conditions such as landslides and large pits on the road surface [11]. Collect traffic lights, speed limits, road hazards, and other information. Browse the car accident list and display the information on the electronic map. The data on road construction and temporary traffic control are collected to realize real-time monitoring of vehicle collisions. The functions of "parking," "right turn," and "acceleration and deceleration" are realized. It can publish traffic events such as road congestion, traffic accidents, construction, temporary road closures and other information [12]. In-depth mining and analysis are carried out to build decision support for driving behavior management, traffic situation monitoring, road network optimization and other aspects combined with the historical data of vehicle-road integration.

3. Overall technical framework of the platform. "Vehicle-road-network" intelligent vehicle control cloud platform for cross-scale real-time monitoring, intelligent decision-making and collaborative control. Intelligent control cloud platform is the center of intelligent road networking cloud computing and the hub of data collection, processing, integration and application [13]. It can capture, convert, process and store advanced information in real-time from driverless vehicles, intelligent connected vehicles, facilities and roadside sensing. The integration of new energy vehicle driving information is achieved through the integration, fusion, and analysis of these data. The human-computer interaction platform combines modern cloud storage, cloud computing, big data, 5G, holographic intelligent perception, data communication transmission, electronic control and computer processing and other technologies into the system [14]. The overall technical architecture of the intelligent management and control cloud platform is shown in Figure 3.1 (the picture is quoted in the Novel ITS based-on Space-Air-Ground collected big-data).

The construction of the human-computer interaction architecture of mobile edge computing centers can reduce the signal delay in the automatic driving mode of new energy vehicles. At the same time, it can give full play to the geographical coverage characteristics of edge computing networks and support the cooperative operation of vehicles and roads with regional characteristics. The Computing architecture of human-computer interaction cloud-edge collaboration for new energy vehicles is shown in Figure 3.2 (the picture is quoted in Vehicular Edge Computing and Networking: A Survey).

3.1. Control signal acquisition. The joystick collects and processes this information in real-time as a critical component linking the platform control equipment to the computer simulation. The selected data acquisition board comprises 16 digital input modules and eight analog input modules and is interfacing with the simulation computer through RS232 serial communication, which can well meet the driving and signal acquisition requirements on the analog platform [15]. The communication module processes the signal received by the control device to realize the connection and data transmission with the vehicle vision. Using serial communication technology and the MSComm control program, the time class of serial communication is set to record and collect data in real time. Its control signal is the throttle, brake, steering, brake, and so on for its input. When the sensor installed on the vehicle body sends out the corresponding change, it will cause a change in the displacement, angle, and other electrical signals connected to it. According to the dynamics model of the vehicle, it is processed, and the environment is monitored in real-time.

3.2. Driverless. The system comprises steering wheel control, throttle control and brake control. The steering system mainly comprises the steering servo motor, the universal joint, and the relevant parts between the universal joint and the steering shaft. To facilitate and accurately control the brake and accelerator and ensure that the driver's brake and accelerator will not be interfered with under manual driving mode, a rolling screw sliding platform is specially selected to pull the brake and accelerator pedals [16]. The ball screw can make the rotation linear. The measured travel of the accelerator pedal and the brake pedal is 80 mm, and the

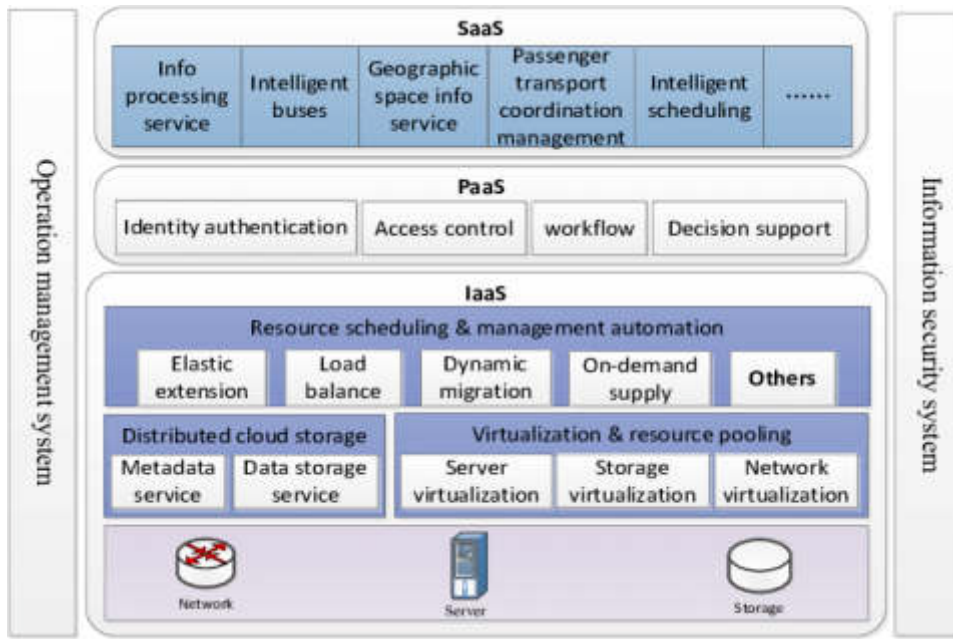


Fig. 3.1: New energy vehicle driving interactive platform system architecture.

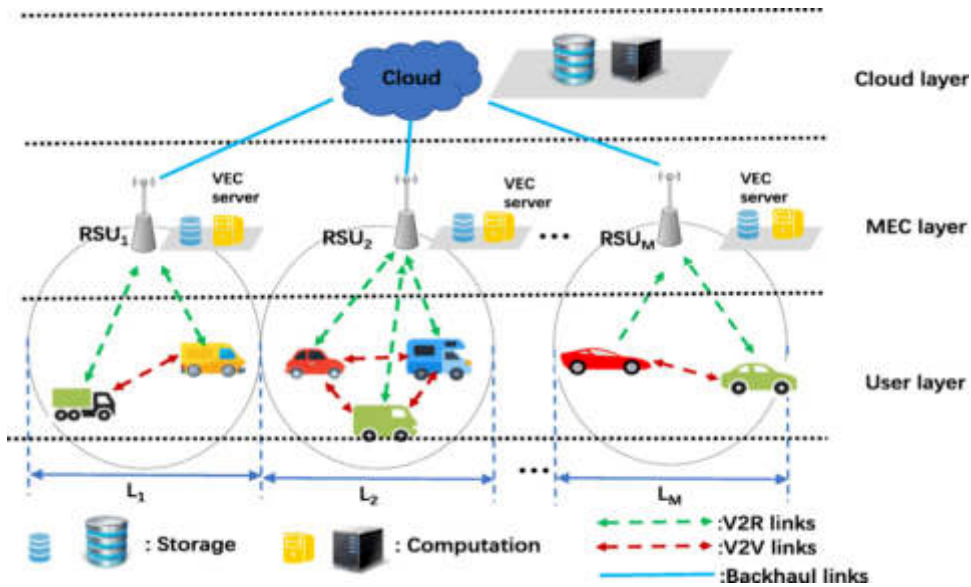


Fig. 3.2: Computing architecture based on cloud edge collaboration.

travel of the brake pedal is 90 mm, where the travel is the length required from the release of the pedal to press the drag chain fully. Through the motion control card to complete the mode of autonomous driving, the motion control card is used to transmit the real-time pulse signal to the servo motor, and then the received pulse signal starts rotating to make the slider rod on the lead screw move. Then, drive the platform for steering, acceleration, deceleration and other actions. In the driverless mode, the user can write the vehicle's trajectory

tracking function into the motion control card for various simulation situations. Let the motion control card perform the corresponding driving action according to the route.

4. Adaptive cruise control algorithm. The combination of PC and low-level hardware is used to collect and process the vehicle distance in front of the vehicle. Pass the desired acceleration command to the underlying device [17]. The onboard sensor at the bottom of the vehicle collects all the information needed when driving, and the motor completes the action, such as an accelerator and brake, to complete the vehicle’s driving. The car must be well-spaced during the journey. Use the front distance to determine the optimal workshop distance:

$$y_{a,y}(x) = \tau_2 u_l(x) + y_{a,0}$$

τ_2 is the distance between cars ahead. $y_{a,0}$ represents the safe distance between parked vehicles. The state quantity $\gamma(x) = [e(x), u_a(x), \delta_l(x), w(x-1)]^T$ of the system is defined. An interference observer that can accurately estimate the system is proposed.

$$\begin{cases} \gamma(x+1) = \lambda\gamma(x) + \eta_w \Delta w(x) + K_\gamma(p(x) - \hat{p}(x)) + \eta_s s(x) \\ p(x) = \mu\gamma(x) \\ \omega(x+1) = Z\omega(x) + K_s(p(x) - \hat{p}(x))s(x) \\ s(x+1) = U\omega(x) \end{cases}$$

$e(x), u_a(x), \delta_l(x), w(x)$ is the distance deviation between vehicles at the time x , the relative speed of vehicles, the acceleration of vehicles, and the acceleration command. The state variables formed by it conform to:

$$\gamma(x) \in \Gamma, \forall x = 0, 1, \dots$$

$$\Gamma = [-\tau_2 u_{l, \max}] \times [u_{a, \min}, u_{a, \max}] \times [\delta_{l, \min}, \delta_{l, \max}] \times [w_{\min}, w_{\max}] . \lambda, \eta, \mu, Z \text{ and}$$

U are constant matrices. K_γ, K_s is the gain of the observatory. $p(\gamma)$ and $\hat{p}(\gamma)$ represent the calculated deviation of workshop spacing and the actual deviation. The acquisition of $\hat{p}(\gamma)$ is detected by sensors on the car to obtain real-time data. s is an input of uncertain bounded interference. Use $\Delta w(x) = w(x) - w(x-1)$ to determine the incremental control input. Through vehicle networking communication and interaction between participants, perception and control of complex scenes are realized. The requirements for traceability and multiple economic objectives required in driving are given.

$$\begin{aligned} K_1(\gamma(x)) &= z_e \gamma_1^2(x) + z_u \gamma_2^2(x) \\ K_2(\gamma(x)) &= z_\delta \gamma_3^2(x) + z_{\delta c} \gamma_4^2(x) \\ K_3(w(x)) &= z_{sw} \Delta w^2(x) \end{aligned}$$

z_e, z_u, z_δ and z_{sw} are the weight factors corresponding to vehicle time error, speed error, acceleration and acceleration instruction in vehicle-road cooperative driving. It indicates the importance of each indicator. The following are sorted out and combined with the proposed performance indicators:

$$H(\gamma, \Delta w) = \sum_{i=0}^{j-1} \{ \gamma^T(i | x) Z_\gamma(x) \gamma(i | x) + z_{sw} \Delta w^2(i | x) \}$$

The state matrix is represented by $Z_\gamma = \text{diag} \{z_e, z_u, z_\delta, z_{sw}\}$. p is the predicted time interval. There exists $[x, x+j](x = 0, 1, L)$ in the control time domain of model predictive control. The vehicle acceleration parameter $\Delta w(i | x)$ is parameterized as follows:

$$\Delta w(i | x) = \varphi^i \Delta w(0 | x), i = 1, 2, L, j - 1$$

$i | x$ refers to the future value predicted for $x + i$ at sampling time x . And the compliance coefficient φ , appropriate to $[0, 1]$, is a parameter used to adjust the smooth output. Due to the high computational complexity

Table 5.1: *Test parameters of simulation test.*

Argument	Default value
Scene size /m x m	3000×3000
Simulation time /h	24
Node communication radius /m	150
Message size /Mb	5
Message generation interval /s	510
MAC layer protocol	IEEE802.11p
Application layer data flow	CBR
Transmission model	Two Ray Groun

of model prediction, its application is restricted, so the control sequence $\Delta w(i | x)$ is parameterized by the current decision variable $\Delta w(0 | x)$. This can be compensated by increasing the step size of the prediction time domain j , because a more significant value of j can lead to better performance. When the control gain of an interval $\Delta \bar{w}_{\min}(x) \leq \Delta w(x) \leq \Delta \bar{w}_{\max}(x)$ is given, the physical properties and driving comfort of the vehicle are considered comprehensively. The constraint is:

$$\left\{ \begin{array}{l} \Delta \bar{w}_{\min}(x) = \max \left\{ \Delta w_{\min}, (w_{\max} - w(x-1)) / \sum_{i=0}^{j-1} \varphi^i \right\} \\ \Delta \bar{w}_{\max}(x) = \min \left\{ \Delta w_{\max}, (w_{\min} - w(x-1)) / \sum_{i=0}^{j-1} \varphi^i \right\} \\ \gamma(i | x) = \lambda^i \gamma(0 | x) + \\ \sum_{j=1}^i \varphi^{i-j} \lambda^{j-1} \eta_w \Delta w(0 | x) \\ + \sum_{j=1}^i \alpha^{i-j} \lambda^{j-1} \Delta \hat{p}(0 | x) \\ + \sum_{j=1}^i \lambda^{j-1} \eta_s \hat{s}(i-j | x) \end{array} \right.$$

Since neither the present nor the future disturbance $s(i-j | x)$ is known, and the values of past disturbances at various times are used as present forecasts, there is $s(j-j | x) = s(x-1)$, $j = 1, 2, \dots, j$.

5. System inspection. The open-source highway traffic simulation programs SUMOSUMOs and NS 3 are tested and studied. The actual mathematical model of vehicle driving is established using the SUMO method. Use NS 3 to simulate packets' transmission success rate and average transmission delay with different nodes and TTL values. The test was conducted on Ubuntu16.04. The test parameters of the simulation test are shown in Table 5.1. The scenario and operation of the simulation test are shown in Figure 5.1 (the picture is quoted in System architecture for installed-performance testing of automotive radars over-the-air).

The simulation of transmission success rate under different node motion rates is shown in Figure 5.2. When the transmission rate is not high, the system's topology will change slowly, and the communication connection is relatively smooth. When the network transmission rate increases, the instability of the link will decrease, and the transmission success rate of various methods will decrease. At the same time, RSSM can dynamically adjust according to the change in network topology caused by fast movement to ensure the transmission channel's maximum connectivity and improve the transmission success rate.

6. Conclusion. Interaction mode selection, information identification, multi-purpose design and beautiful appearance design will all impact the interactive experience of new energy vehicle users. The purpose of this study is to discuss how to overcome the difficulties and satisfaction of the users in the human-machine interaction design of new energy electric vehicles from the theoretical and practical aspects. The acquisition of operation equipment and signals is studied under manual driving of new energy vehicles. The design of the electric control circuit is proposed, and the main parts and electrical components are selected and tested. It has been found that the system can reproduce the whole car process in real-time through human-computer interaction.

7. Acknowledgements. The work was supported by the Anhui Wenda University of Information Engineering Natural Science Research Project "Research on Human Machine Intelligent Interaction System for New Energy Vehicles" (Project Number: XZR2023A02).

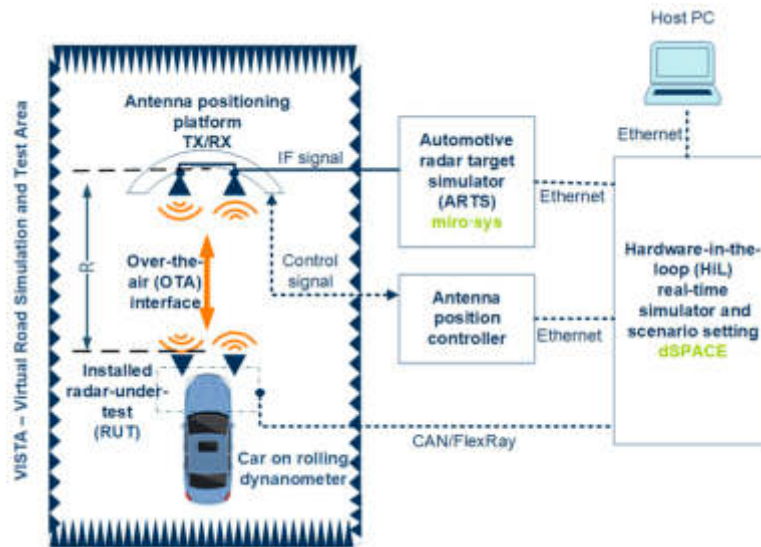


Fig. 5.1: Operating diagram of simulation test.

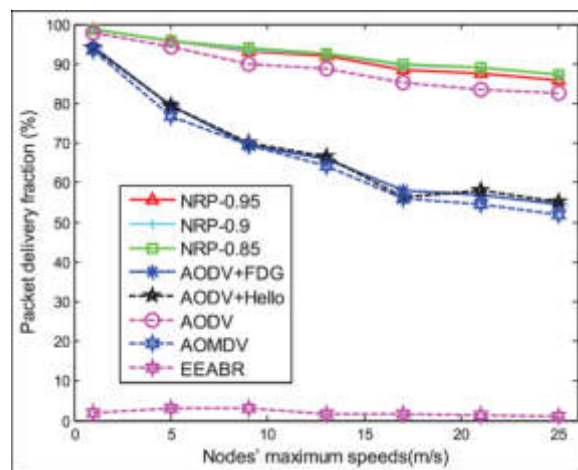


Fig. 5.2: Simulation of the relationship between network node rate and delivery success rate.

REFERENCES

- [1] Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495-504.
- [2] Detjen, H., Faltaous, S., Pflöging, B., Geisler, S., & Schneegass, S. (2021). How to increase automated vehicles' acceptance through in-vehicle interaction design: A review. *International Journal of Human-Computer Interaction*, 37(4), 308-330.
- [3] McDonnell, A. S., Simmons, T. G., Erickson, G. G., Lohani, M., Cooper, J. M., & Strayer, D. L. (2023). This is your brain on Autopilot: Neural indices of driver workload and engagement during partial vehicle automation. *Human factors*, 65(7), 1435-1450.
- [4] Bindhu, V. (2020). An enhanced safety system for auto mode E-vehicles through mind wave feedback. *Journal of Information Technology*, 2(03), 144-150.
- [5] Dargahi Nobari, K., Albers, F., Bartsch, K., Braun, J., & Bertram, T. (2022). Modeling driver-vehicle interaction in automated driving. *Forschung im Ingenieurwesen*, 86(1), 65-79.
- [6] Wang, X., Zheng, X., Chen, W., & Wang, F. Y. (2020). Visual human-computer interactions for intelligent vehicles and intelligent transportation systems: The state of the art and future directions. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, Man, and Cybernetics*, 50(12), 1-14.

- Cybernetics: Systems, 51(1), 253-265.
- [7] Tan, H., Sun, J., Wenjia, W., & Zhu, C. (2021). User experience & usability of driving: A bibliometric analysis of 2000-2019. *International Journal of Human-Computer Interaction*, 37(4), 297-307.
 - [8] Wintersberger, P. (2023). Team at Your Service: Investigating Functional Specificity for Trust Calibration in Automated Driving with Conversational Agents. *International Journal of Human-Computer Interaction*, 39(16), 3254-3267.
 - [9] Sevchenko, N., Appel, T., Ninaus, M., Moeller, K., & Gerjets, P. (2023). Theory-based approach for assessing cognitive load during time-critical resource-managing human-computer interactions: an eye-tracking study. *Journal on Multimodal User Interfaces*, 17(1), 1-19.
 - [10] Qian, X., Ju, W., & Sirkin, D. M. (2020). Aladdin's magic carpet: Navigation by in-air static hand gesture in autonomous vehicles. *International Journal of Human-Computer Interaction*, 36(20), 1912-1927.
 - [11] Miller, L., Kraus, J., Koniakowsky, I., Pichen, J., & Baumann, M. (2023). Learning in Mixed Traffic: Drivers' Adaptation to Ambiguous Communication Depending on Their Expectations toward Automated and Manual Vehicles. *International Journal of Human-Computer Interaction*, 39(16), 3268-3287.
 - [12] Le Guillou, M., Prévot, L., & Berberian, B. (2023). Bringing together ergonomic concepts and cognitive mechanisms for human-AI agents cooperation. *International Journal of Human-Computer Interaction*, 39(9), 1827-1840.
 - [13] Zhou, F., Yang, X. J., & Zhang, X. (2020). Takeover transition in autonomous vehicles: A YouTube study. *International Journal of Human-Computer Interaction*, 36(3), 295-306.
 - [14] Wang, D., Yin, G., & Chen, N. (2021). Optimisation of dynamic navigation system for automatic driving vehicle based on binocular vision. *International Journal of Industrial and Systems Engineering*, 39(3), 411-428.
 - [15] LUO, W., CAO, J., ISHIKAWA, K., & JU, D. (2021). Experimental Validation of Intelligent Recognition of Eye Movements in the Application of Autonomous Vehicle Driving. *International Journal of Biomedical Soft Computing and Human Sciences: the official journal of the Biomedical Fuzzy Systems Association*, 26(2), 63-72.
 - [16] Hancock, P. A. (2022). Avoiding adverse autonomous agent actions. *Human-Computer Interaction*, 37(3), 211-236.
 - [17] Riegler, A., Riener, A., & Holzmann, C. (2021). Augmented reality for future mobility: Insights from a literature review and hci workshop. *i-com*, 20(3), 295-318.

Edited by: Zhigao Zheng

Special issue on: Graph Powered Big Aerospace Data Processing

Received: Jan 9, 2024

Accepted: Feb 3, 2024



APPLICATION OF CLUSTER ANALYSIS ALGORITHM IN SUPPLY CHAIN RISK IDENTIFICATION

QINGPING ZHANG* AND YI HE†

Abstract. The risk control model of the power supply chain system is established. A fault information identification method based on fuzzy clustering is proposed. This method fully considers the power grid's characteristics and uses terrible data. A risk assessment model based on fuzzy set theory is established by the COWA operator weight method and grey cluster evaluation method. The security risk identification model of power grid enterprises uses insufficient data. The security risk identification data are normalized and classified. Empirical analysis determines various risk factors that may appear in power projects. The applicability and feasibility of the index system and evaluation model are verified.

Key words: Fuzzy clustering; Power supply chain; Security risk monitoring; Risk identification

1. Introduction. The rapid development of the electric power industry can benefit the people and greatly promote the development of the national economy. In the early stage of development, because the power industry has a high degree of monopoly, many operating entities such as power generation, electricity sales and transmission are concentrated in one company. This results in an industrial governance model similar to the corporate governance model, which has a high degree of monopoly but also causes a lot of resource waste, low efficiency, and high operating costs. At the same time, there are problems such as rent-seeking, price discrimination and network barriers. China has launched a series of power system reforms and market-oriented plans by introducing market competition to change the monopoly situation of the whole industry. This can reduce the operating cost and realize the rationality of resource allocation. It can not only effectively promote market competition but also reduce the waste of energy. The domestic power grid has separated power plants from the grid and established a perfect quotation model for power generation and other industrial sectors. It has broken through the monopoly model of the past and established a complete power grid supply chain. However, some potential security risks cannot be ruled out in the domestic power supply chain. Against this background, it is an important research direction for power network security risk monitoring.

The FMEA model was analyzed in the literature [1]. A complex diffusion network model of fault mode is established by using a complex network analysis method. The influence of the correlation between fault modes and fault modes is studied. Taking a typical supply chain as the research object, FMEA and complex network methods are studied to test their application value in fault-type evaluation. Literature [2] proposed the generation mechanism of risk factors in the construction stage of hydropower projects based on the perspective of the supply chain. The sub-indexes of 5 categories and 18 categories were established. The final risk assessment value is consistent with the actual risk status of the project. The effectiveness and feasibility of this method are proven. Literature [3] constructs the framework of an inter-provincial power trading system based on RPS. Clarify power supply and demand issues among various market participants. The customer's subjective choice is introduced, and the utility function describes the customer's purchasing behavior. The optimal decision problem of multiple trade participants in the supply chain based on maximum return is constructed. By using the method of reverse derivation, the optimal trade decision problem of each trader is solved [4]. It has particular reference significance to the cross-province power market in our country. However, the power supply in these ways is unstable. Therefore, this paper uses the fuzzy clustering method to monitor the security risks of each

*Business School, Shunde Polytechnic, Foshan, Guangdong, 528333, China (Corresponding author, 10370@sdpt.edu.cn)

†School of Economics and Management, Guangdong Vocational College of Post and Telecom, Guangzhou, Guangdong Province, 510630, China

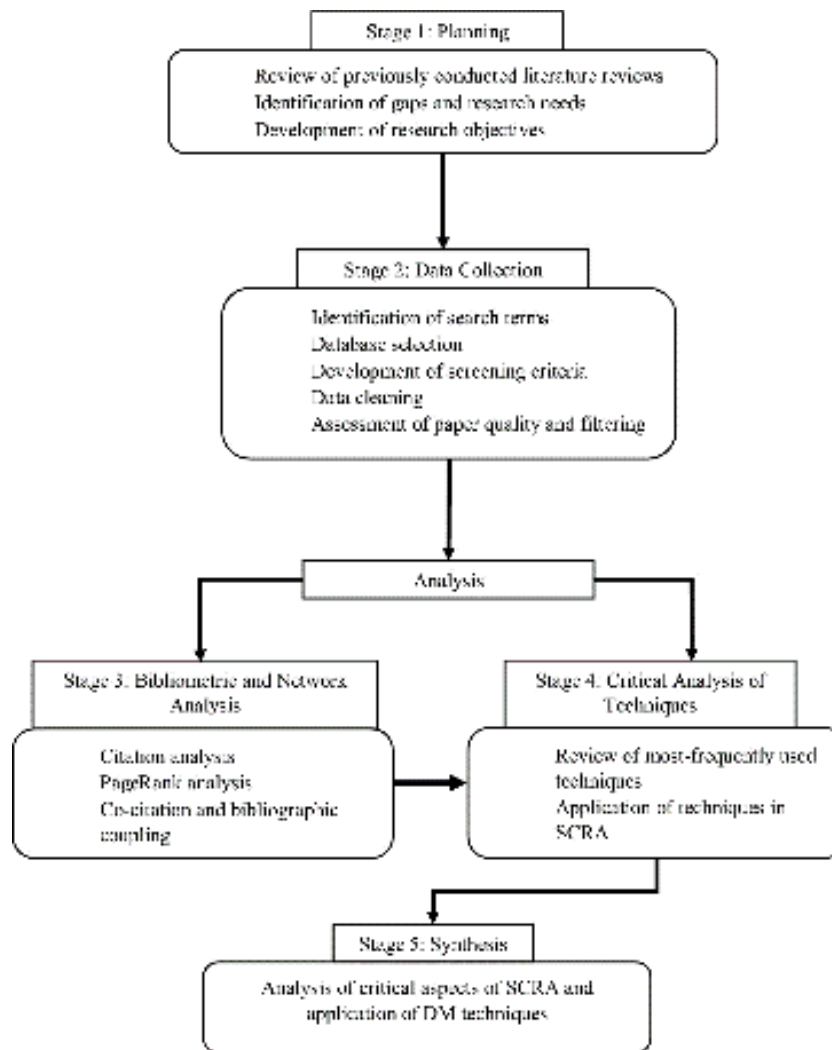


Fig. 2.1: *Specific construction process.*

link enterprise in the power grid to achieve the optimal allocation and supply chain management of each link enterprise in the power grid. The aim is to improve the stability of the power grid.

2. Fuzzy cluster monitoring in the power supply chain.

2.1. Security Risk Identification. The wrong data of the power grid system is analyzed according to the characteristics of the power grid itself based on the identification of power grid fault information. The security risks of power grid enterprises can be effectively identified [5]. The power grid operation's risk identification mechanism is obtained. First, the security of insufficient data must be collected and classified. The security risk index of the power supply chain system based on the network is constructed. Then, a preliminary framework of a fuzzy comprehensive evaluation system is constructed [6]. The detailed construction process is shown in Figure 2.1 (image cited in *Annals of Operations Research*, 2023, 322(2): 565-607). Through the identification of security risks of power grid enterprises, they are divided into internal and external categories. These two risk factors are listed in Table 2.1.

Table 2.1: *Safety risk factors.*

	Evaluation index	Level 1 security risk factor	Secondary security risk factors
Power supply chain security risk factors	Natural hazard index The closeness of national policy Industrial climate index Overall economic condition Customer satisfaction Inventory turnover Profit margin on sales Production cost Research and development phase Product flexibility Order fulfillment rate Qualified rate of finished product Order preparation time	External security risk of the power supply chain Internal security risk of the power supply chain	Ecological environment risk Policy and regulation risk The risk of public security hazards Economic and environmental risks Seller's security risks Manufacturer's safety risk Supplier security risk

2.2. Security risk monitoring. After obtaining the relevant information on the risk factors inside and outside the electricity market, the next step is to process the data [7]. The first step is to limit the scope of the relevant data. Monitoring the power grid security risk is to normalize and classify the collected information. The main contents include daily power threshold security risk monitoring, harmonic limit security risk monitoring, and parameter setting security risk monitoring. The security risk monitoring of the daily load threshold is mainly used to monitor the stability and security risks of the power grid load during the power grid operation. Power grid side security risk monitoring is mainly through the power grid voltage level and change amplitude to monitor the power grid operation [8]. If the voltage level and the range of change cannot meet the electricity demand, it must be modified. A power metering device is designed to measure the running state of the power system. If the parameter of the meter is set to 0.2, then the allowable deviation of the meter must be controlled within 0.2%. If the end indicator of the meter exceeds the control value, it indicates a security risk for the customer's meter in the entire grid. The power data collected during power grid operation is a vital link. It also includes the power meter precision, power rate, and other parameters set by the security risk monitoring function [9]. Prevent users from arbitrarily changing the user's power parameters.

3. A comprehensive risk assessment model of a power supply chain is established by combining C-OWA with grey clustering.

3.1. The COWA algorithm is used to calculate the weights of each evaluation index. The essence of the OWA algorithm is to sort the data incrementally, and the weighting depends only on the space. A ranked, weighted mean C-OWA algorithm is proposed. The detailed process is like this: 1) An expert is invited to evaluate the importance of the evaluation indicators at all levels, and the scores constitute the initial evaluation data set $(\eta_1, \eta_2, \dots, \eta_i, \dots, \eta_n)$ of the evaluation indicator H , and $\theta_0 \geq \theta_1 \geq \theta_2 \geq \dots \theta_j \geq \dots \theta_{n-1}$ is obtained from θ in the order from high to low. 2) The weighting λ_{j+1} of data θ_i is directly determined by the number of combinations Z_{n-1}^j :

$$\lambda_{j+1} = \frac{Z_{n-1}^j}{\sum_{t=0}^{n-1} Z_{n-1}^t} = \frac{Z_{n-1}^j}{2^{n-1}}, j = 0, 1, 2, \dots, n-1$$

Formula: $\sum_{j=0}^{n-1} \lambda_{j+1} = 1$. 3) Weight the evaluation data with weight vector λ to obtain the absolute weight $\bar{\delta}$ of index η_i :

$$\bar{\delta} = \sum_{j=1}^n \lambda_j \cdot \theta_j \in [0, 1], j \in [1, n]$$

4) Calculate the relative weight value δ_i for the exponential factor η_i

$$\delta_i = \frac{\bar{\delta}}{\sum_{i=1}^m \bar{\delta}_i}, i = 1, 2, \dots, m$$

3.2. Grey cluster analysis of the overall risk of the power supply chain. Many factors affect the integration risk of the power supply chain, and most of the existing evaluation methods rely on experts' subjective experience, knowledge level and subjective preferences [10]. So, the grey clustering method is used to study the risk of the power supply chain.

3.2.1. Grey category judgment and whitening weight. The complexity of risk assessment indicators will directly affect the refinement of grey categories [11]. It is assumed that the target to be assessed is divided into s grey categories, and the interval of its secondary index is also divided into s grey categories. It is divided into very low [2,0], low [4,2], average [6,4], high [8,6], and high [10,8]. People can get the point vector $U = (9, 7, 5, 3, 1)$ by applying the traditional grey system theory.

3.2.2. Grey cluster evaluation steps. 1) Construction of evaluation model. This paper classifies the risk degree of the power supply chain based on q class of power industry experts, supply chain research experts, mahagement experts and construction experts [12]. Then the evaluation matrix $S_i = [s_{ijt}]_{s \times q}$ is constructed according to the score of index H_{ij} . s is the number of exponents of the matrix. 2) Construction of grey cluster weight matrix. The clustering factor of Grade H_{ij} and Grade e gray level is grade $U_{ije} = \sum_{n=1}^q g_e [s_{ijt}]$, and the overall level evaluation factor is grade $U_{ij} = \sum_{e=1}^5 U_{ije}$. Then the weight vector of the gray cluster can be calculated as $c_{ije} = \frac{U_{ije}}{U_{ij}}$, and the weight matrix of the gray cluster can be obtained as follows

$$C_i = \begin{bmatrix} c_{i11} & c_{i12} & c_{i13} & c_{i14} & c_{i15} \\ c_{i21} & c_{i22} & c_{i23} & c_{i24} & c_{i25} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{ij1} & c_{ij2} & c_{ij3} & c_{ij4} & c_{ij5} \end{bmatrix}$$

3) Comprehensive cluster evaluation matrix. Evaluate each significant index cluster according to formula (3.5):

$$V_i = \delta \cdot C_i$$

Take $V_0 = [V_1, V_2, \dots, V_n]^T$ as the comprehensive evaluation matrix of the above index, and then use formula (3.6) to conduct a comprehensive cluster evaluation of the index:

$$Y = \delta \cdot V_0 = [Y_1, Y_2, \dots, Y_n]$$

4) Synthesize the evaluation values at all levels. Formula (3.7) is used to integrate weight Y and weight O to obtain the risk level of the power supply chain to prevent secondary losses in the evaluation process.

$$\Lambda = Y \cdot O^T$$

4. Experimental detection.

4.1. System Validity Check. A fault information identification method based on fuzzy clustering is proposed. Data was processed using the Xon (R) Server with Windows Server 2012R2 [13]. The network topology of the power system is established, and the total load is obtained. A risk identification model based on load level is proposed. Different test schemes are compared and analyzed [14]. The 46,890 labeled samples were classified. Among them, 70% are taken as training samples, 10% as confirmation samples and 20% as test samples (Table 4.1).

The correctness of the proposed algorithm is verified by comparing it with fuzzy clustering, k nearest neighbor classification, SVM, convolutional neural network and recurrent neural network. Each class of algorithms is based on a training set. Run ten times in one test set. Finally, the average of 10 samples is the final prediction

Table 4.1: Data set division unit.

Training set	Validation set	Test set
34191	4884	9769

Table 4.2: Accuracy rates of different models.

Model	Accuracy rate / %
Fuzzy clustering method	78.96
KNN	79.48
SVM	85.83
CNN	89.38
LSTM	90.21
CPE	93.54

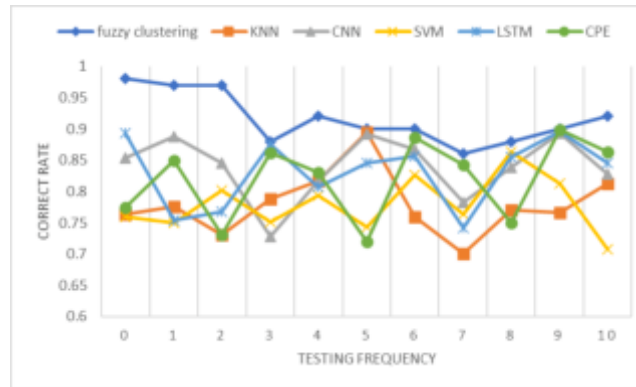


Fig. 4.1: Test accuracy.

accuracy [15]. Table 4.2 shows that the mathematical model proposed in this paper has good test results. It has advantages over classical machine learning algorithms such as fuzzy clustering, KNN, and support vector machines. Performance is better than CNN LSTM.

The model was trained during the experiment. Each trained model is run ten times on the same data set. The method is predicted ten times, and the final prediction accuracy is obtained [16]. The following conclusions are drawn through the analysis of the experimental data: 1) The accuracy of the fuzzy cluster analysis model in this paper can reach 89.8% by comparing the accuracy of different models. This algorithm exceeds the conventional machine learning algorithm and is better than the widely used deep learning algorithms such as neural networks and LSTM. Because the initial value limits the selection of the KNN initial value, it isn't easy to obtain an ideal initial value. However, support vector machines have substantial limitations in selecting kernel functions and are unsuitable for large-scale data. Although CNN has a good feature extraction function, there is often an aggregation process after extraction. And through pooling, resulting in more significant information loss. The short-term memory method is ineffective in feature extraction [17]. The clustering method has the advantages of better feature extraction, not being easy to lose, not being affected by the initial value, and having a greater demand for data. 2) As seen in Figure 4.1, the variance of the cluster analysis model is the least, while the variance of other models is more significant, indicating that the clustering method is robust. 3) The weights of high-level hazard factors and low-level factors are shown in Figure 4.2. The results show that the factors causing harm are the largest in the distribution system, and the planning links occupy a large proportion.

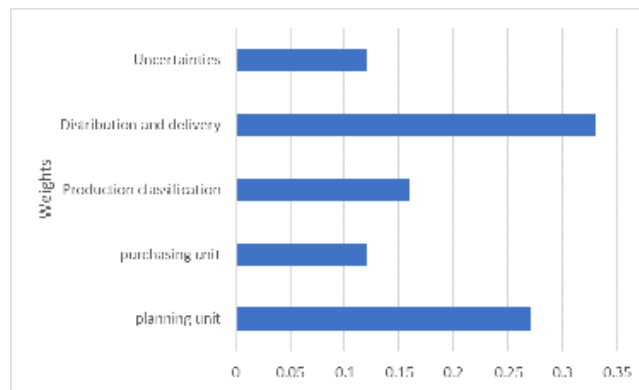


Fig. 4.2: Risk weight values.

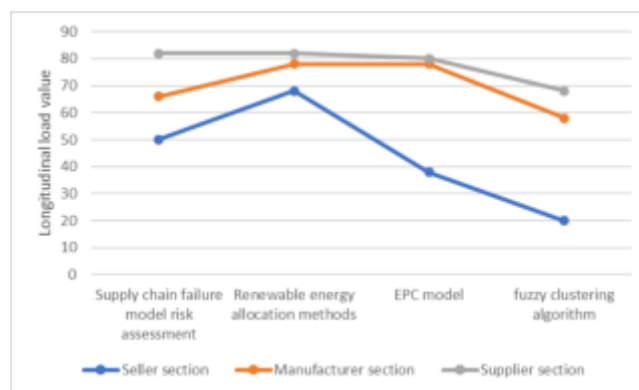


Fig. 4.3: Comparison of power supply chain stability.

4.2. System stability analysis. The supply chain failure mode risk assessment method, EPC mode method, renewable energy quota method and other methods are combined with this method, and the fuzzy cluster analysis method is compared [18]. The results show that the model in this paper can reflect the stability of the power grid well (Figure 4.3).

A risk assessment model based on fault types of supply chain is proposed. The resulting power grid stability is about 72.2% [19]. The simulation results show that the stability of the model is about 42.6%. The results show that the proposed algorithm has good stability. The fundamental reason is that the research idea proposed in this project is based on identifying insufficient data and integrating it with the characteristics of the power grid to identify the security risks of power grid enterprises. It guarantees the stability of the supply chain.

5. Conclusion. The fuzzy clustering model of power supply chain system risk is established to ensure the high stability of the power grid. This research result has important practical significance for developing China's power market. It is also relatively easy to implement. Its implementation process is simple and time-consuming is short, so it has good promotion value.

REFERENCES

- [1] Shishodia, A., Sharma, R., Rajesh, R., & Munim, Z. H. (2023). Supply chain resilience: A review, conceptual framework and future research. *The International Journal of Logistics Management*, 34(4), 879-908.
- [2] Babu, H., Bhardwaj, P., & Agrawal, A. K. (2021). Modelling the supply chain risk variables using ISM: a case study on Indian manufacturing SMEs. *Journal of Modelling in Management*, 16(1), 215-239.

- [3] Singh, S., & Kumar, K. (2021). A study of lean construction and visual management tools through cluster analysis. *Ain Shams Engineering Journal*, 12(1), 1153-1162.
- [4] Han, Y., Yan, X., & Piroozfar, P. (2023). An overall review of research on prefabricated construction supply chain management. *Engineering, Construction and Architectural Management*, 30(10), 5160-5195.
- [5] Zaridis, A., Vlachos, I., & Bourlakis, M. (2021). SMEs strategy and scale constraints impact on agri-food supply chain collaboration and firm performance. *Production Planning & Control*, 32(14), 1165-1178.
- [6] Li, G., Kou, G., & Peng, Y. (2021). Heterogeneous large-scale group decision making using fuzzy cluster analysis and its application to emergency response plan selection. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(6), 3391-3403.
- [7] Ivanov, D. (2023). The Industry 5.0 framework: Viability-based integration of the resilience, sustainability, and human-centricity perspectives. *International Journal of Production Research*, 61(5), 1683-1695.
- [8] Yang, J., Xie, H., Yu, G., & Liu, M. (2021). Antecedents and consequences of supply chain risk management capabilities: An investigation in the post-coronavirus crisis. *International Journal of Production Research*, 59(5), 1573-1585.
- [9] Chen, Z. S., Zhang, X., Rodríguez, R. M., Pedrycz, W., Martínez, L., & Skibniewski, M. J. (2022). Expertise-structure and risk-appetite-integrated two-tiered collective opinion generation framework for large-scale group decision making. *IEEE Transactions on Fuzzy Systems*, 30(12), 5496-5510.
- [10] Akbari, M., & Do, T. N. A. (2021). A systematic review of machine learning in logistics and supply chain management: current trends and future directions. *Benchmarking: An International Journal*, 28(10), 2977-3005.
- [11] Kang, H., & Jung, E. H. (2021). The smart wearables-privacy paradox: A cluster analysis of smartwatch users. *Behaviour & Information Technology*, 40(16), 1755-1768.
- [12] Neuburger, L., & Egger, R. (2021). Travel risk perception and travel behaviour during the COVID-19 pandemic 2020: A case study of the DACH region. *Current issues in tourism*, 24(7), 1003-1016.
- [13] Budianto, E. W. H. (2023). Research Mapping on Credit Risk in Islamic and Conventional Banking. *AL-INFAQ: Jurnal Ekonomi Islam*, 14(1), 73-86.
- [14] Um, J., & Han, N. (2021). Understanding the relationships between global supply chain risk and supply chain resilience: the role of mitigating strategies. *Supply Chain Management: An International Journal*, 26(2), 240-255.
- [15] Akin Ateş, M., Suurmond, R., Luzzini, D., & Krause, D. (2022). Order from chaos: a meta-analysis of supply chain complexity and firm performance. *Journal of Supply Chain Management*, 58(1), 3-30.
- [16] Izaguirre, C., Losada, I. J., Camus, P., Vigh, J. L., & Stenek, V. (2021). Climate change risk to global port operations. *Nature Climate Change*, 11(1), 14-20.
- [17] Xu, X., Zhu, D., Yang, X., Wang, S., Qi, L., & Dou, W. (2021). Concurrent practical byzantine fault tolerance for integration of blockchain and supply chain. *ACM Transactions on Internet Technology (TOIT)*, 21(1), 1-17.
- [18] Matiza, T., & Kruger, M. (2021). Ceding to their fears: A taxonomic analysis of the heterogeneity in COVID-19 associated perceived risk and intended travel behaviour. *Tourism Recreation Research*, 46(2), 158-174.
- [19] Modgil, S., Gupta, S., Stekelorum, R., & Laguir, I. (2021). AI technologies and their impact on supply chain resilience during COVID-19. *International Journal of Physical Distribution & Logistics Management*, 52(2), 130-149.

Edited by: Zhigao Zheng

Special issue on: Graph Powered Big Aerospace Data Processing

Received: Jan 10, 2024

Accepted: Feb 18, 2024



TRANSFORMER FAULT DIAGNOSIS AND LOCATION METHOD BASED ON FAULT TREE ANALYSIS

ZHIWU WU*, TIANFU HUANG†, CHUNGUANG WANG‡, XIANG WU§ AND YANZHAO TU¶

Abstract. Fiber optical current transformer (FOCT) is widely used in power systems for fault diagnosis and analysis, which can improve its operational reliability. Construct fault modes and fault trees based on fault data of all fiber current transformers, and construct a fault feature space. Constructing a fault diagnosis expert system using fault trees and fault feature space clustering centers to achieve accurate diagnosis of fault types, patterns, and components. The proposed method was validated using fault data and case studies of all fiber current transformers in a regional power grid, and the results showed that: The on-site fault case is closest to the cluster center of drift deviation fault, so it belongs to drift deviation fault. Further extract the on-site maintenance report, which indicates that the operating temperature of the all fiber current transformer is relatively high. The diagnostic results of the fault diagnosis expert system for the faulty all fiber current transformer are consistent with the actual results on site, verifying the accuracy and reliability of this method.

Key words: All fiber current transformer, Fault Mode and Effects Analysis Method, Fault tree, Expert system, fault diagnosis

1. Introduction. Transformer is a very important electrical equipment in the power system, whose main responsibility is to measure and transmit current and voltage information to ensure the safe operation of the power system. However, due to various reasons, transformers may malfunction, such as insulation damage, inter turn short circuits, open circuits, etc. These faults may lead to measurement errors, equipment damage, and even power system accidents.

Firstly, insulation damage is a common occurrence of transformer faults [1]. The insulation layer of the transformer plays a role in protecting the coil and magnetic core. Once the insulation layer is damaged, current may directly enter the coil of the transformer through the insulation layer, causing measurement errors and even equipment damage. Insulation damage may be caused by aging due to prolonged operation, or it may be due to harsh external conditions such as high temperature, humidity, etc. Therefore, it is very important to regularly inspect and maintain the insulation layer of the transformer to ensure its normal operation. Secondly, inter turn short circuit is also a common problem of transformer faults [2]. The coil of a transformer consists of many turns, and when a short circuit occurs between some of these turns, the current will bypass these turns, causing measurement errors. Short circuit between turns may be caused by damaged coil insulation, insufficient insulation distance between coils, and other reasons. In order to avoid inter turn short circuits, it is necessary to ensure that the coil insulation of the transformer is intact and that there is sufficient insulation distance between the coils during installation. In addition, the open circuit of the transformer is also a common fault situation. When the coil of the transformer is interrupted or poorly connected, it will result in the inability of current to flow through the coil and make accurate measurements. An open circuit may be caused by coil damage, loose wiring terminals, and other reasons. In order to avoid open circuit faults, it is necessary to regularly check whether the coils of the transformer are intact and ensure that the connection terminals are firmly and reliably connected [3]. The above mentioned fault situations are only a part of the possible occurrence of transformers, and the actual situation may be more complex. In order to ensure the safe operation of the power system, in addition to regular inspection and maintenance of transformers, other measures should also be

*Marketing Service Center of State Grid Fujian Electric Power Co., Ltd., Fuzhou, Fujian, 350013, China (Corresponding author, YanZhao_Tu@163.com)

†Marketing Service Center of State Grid Fujian Electric Power Co., Ltd., Fuzhou, Fujian, 350013, China

‡Marketing Service Center of State Grid Fujian Electric Power Co., Ltd., Fuzhou, Fujian, 350013, China

§Marketing Service Center of State Grid Fujian Electric Power Co., Ltd., Fuzhou, Fujian, 350013, China

¶Marketing Service Center of State Grid Fujian Electric Power Co., Ltd., Fuzhou, Fujian, 350013, China

taken, such as setting up overcurrent protection devices, installing backup transformers, etc., in order to cope with possible fault situations. In addition, with the development of the power system and the application of intelligent technology, the methods for detecting and preventing transformer faults are also constantly improving [4]. For example, using an infrared thermal imager can detect temperature anomalies in transformers, thereby detecting potential faults in advance. In addition, the emergence of intelligent transformers can achieve real-time monitoring and fault diagnosis of transformers, further improving the reliability and safety of transformers.

In short, as an important electrical equipment in the power system, the failure of transformers may have serious impacts on the power system. Therefore, we need to pay attention to the inspection and maintenance of transformers, timely detect and eliminate transformer faults, in order to ensure the safe operation of the power system. The author's purpose is to study the fault diagnosis and localization method for transformers based on fault tree analysis. By establishing a fault tree model for transformers, analyzing the mechanism and possible fault paths of transformer faults, and determining the probability and importance of faults occurring. Based on the results of fault tree analysis, the author proposes corresponding fault diagnosis and positioning strategies, providing technical support for the rapid diagnosis and accurate positioning of transformer faults. Through this study, the aim is to improve the efficiency and accuracy of transformer fault handling, ensuring the safe operation of the power system.

2. Principle and Fault Mode Analysis of All Fiber Optic Current Transformer.

2.1. Principle of all fiber optic current transformers . When linearly polarized light passes through certain optical materials in a direction parallel to the magnetic field, due to the influence of the magnetic field, the front side rotates. The formula for calculating the rotation angle θ of polarized light is

$$\theta = \mu\nu \int_{L_1}^{L_2} H(L)dL \quad (2.1)$$

In the formula, μ is the magnetic permeability of the material; V is the Verdet parameter of the optical material; L is the distance traveled by light in the material; $H(L)$ is the function of the spatially distributed magnetic field with respect to L [5]. If fiber optic or magneto optic block glass is used to close the optical path around the conductive conductor, then

$$\theta = \mu\nu N \oint H(L)dL = \mu\nu Ni \quad (2.2)$$

In the formula, N is the number of turns of the optical path and current intersection; I is the current flowing through the conductor. Therefore, the current can be calculated by detecting θ .

2.2. Fault mode analysis of all fiber optic current transformers . The process of inferring the form, location, and cause of a fault based on fault knowledge and a certain strategy is called the fault diagnosis process. In the diagnosis process, corresponding fault knowledge should be used as the basis, such as fault symptoms, fault detection methods, degree of fault harm, and maintenance measures. The external manifestation of faults is the fault mode, which can be observed through human senses or measuring instruments and meters [6]. When diagnosing equipment faults, it is necessary to first determine the fault mode through factual data, and then determine its impact on the system through the determined fault mode, and propose targeted maintenance measures and solutions. The Fault Mode and Effect Analysis (FMEA) method is used in multiple fields such as nuclear and aviation industries to eliminate equipment failure hazards. FMEA can provide a comprehensive qualitative analysis of system or equipment failure modes. In the design process of a system or equipment, the FMEA method analyzes the potential failure modes of its constituent units and their impact on the system or equipment, and classifies each potential failure mode according to its severity, proposing possible design, prevention, and improvement measures [7]. Before analyzing the all fiber current transformer using FMEA method, it is necessary to subdivide its various components to establish a reliability diagram of the all fiber current transformer. Existing research indicates that the main factors affecting the reliability of current transformers are as follows.

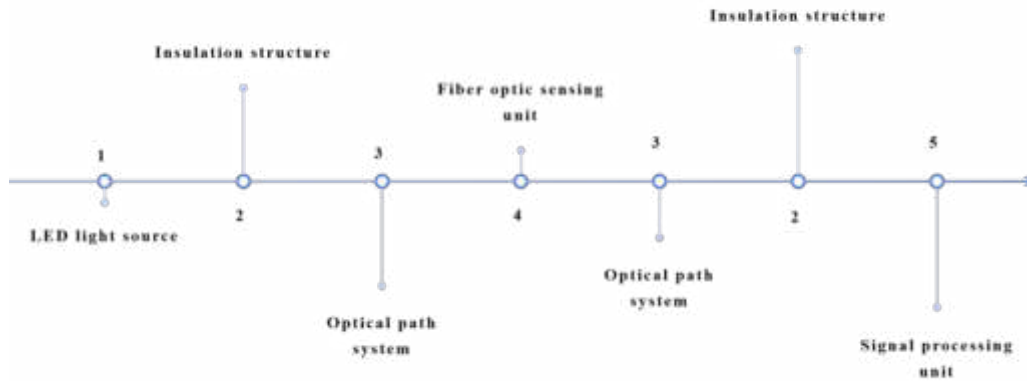


Fig. 2.1: Reliability Block Diagram of Full Fiber Current Transformer

- (1) The influence of sensor head structure. The optical material of the sensor head undergoes changes in its optical properties under the influence of external electric fields, temperature fields, and stress fields, resulting in various effects that affect the stability of the transformer system.
- (2) The impact of optical system. Due to the difference in surface smoothness and parallelism of optical components, as well as the inevitable small displacement during the bonding process, there will be deviation angles in the optical path system, which will affect the stability of the transformer.
- (3) The impact of signal processing circuits. The differences in component performance in signal processing circuits lead to certain deviations, such as low-pass filters and the dispersion of component parameters, which will affect the amplitude frequency response and cause errors in signal processing [8].
- (4) The influence of insulation structure. The insulation structure not only affects the linear range of system measurement, but also the selection of insulation materials affects the analysis of electric fields. The thermal expansion and contraction of materials cause changes in the distance between electrodes, thereby affecting the stability of transformers.
- (5) The impact of LED characteristics. The output of the sensor is a function of wavelength, and the characteristics of LED determine the stability of the wavelength. With the change of temperature, the wavelength of LED will change, thereby affecting the stability of the transformer. Analyze the various parts of the all fiber current sensor and establish its reliability diagram, as shown in Figure 2.1.

Figure 2.1 is based on a commonly used and effective total reflection fiber optic current transformer. Therefore, the insulation structure and optical path system through which the incident and reflected light pass are the same. Therefore, this reliability block diagram only contains 5 different parts. The above parts will individually or comprehensively affect the stability of the all fiber current transformer, and this impact on stability will be manifested through the output of the transformer [9]. Therefore, based on the abnormal situation of the output of the transformer and the influence of specific parts of the transformer, the following 5 different types of current transformer faults can be determined.

- (1) Fixed deviation fault. A fixed deviation fault occurs when there is always a fixed deviation between the measured value and the true value.
- (2) Drift deviation fault. The sensor head of an all fiber current transformer is easily affected by temperature, leading to a decrease in the performance of optical or electronic components, resulting in drift of measurement values over time. This drift is known as drift deviation fault.
- (3) Variable ratio deviation fault. The ratio of a transformer represents the proportional relationship between the true value and the measured value. In practical applications, sudden changes in the ratio may occur due to changes in the operating environment and an increase in operating time, resulting in distortion of the transformer output signal. This situation is called ratio deviation fault.

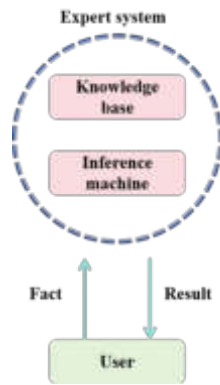


Fig. 3.1: Basic Principles of Expert Systems

- (4) Accuracy distortion fault. The failure of the signal processing module and transmission unit of the all fiber current transformer can cause distortion in the accuracy of the measured values. When accuracy distortion occurs, the average measurement value remains unchanged and the measurement variance changes [10].
- (5) Complete failure failure. Due to hardware circuit faults and optical component failures, the measured value of an all fiber current transformer does not change with the true value and always maintains a certain value (zero or maximum range value). This type of fault is called a complete failure fault.

3. Expert System for Fault Diagnosis of All Fiber Current Transformers . Expert system ES (Expert System) enables computer software systems to apply knowledge, facts, and reasoning mechanisms to solve complex problems that typically require human experts to solve. The expert system mainly consists of two parts: An expert knowledge base and an inference machine. Based on the facts provided by the user, the system uses certain inference methods to make inferences and judgments based on the knowledge base, and finally outputs the results [11]. The basic principle is shown in Figure 3.1.

3.1. Knowledge Base. The knowledge base is the fault knowledge base of all fiber current transformers, which is used to store domain expert knowledge and is a key factor in determining the performance of expert systems. The author's knowledge base consists of the following three parts:

1) *FMEA Table.* Using the FMEA table as expert knowledge, there is a certain connection between the component names, fault modes, fault consequences, and fault types in the table. By analyzing the output data of the transformer, the fault type and most of the fault consequence information can be obtained. Combined with a simple analysis of the transformer structure, the fault location can be determined, thus achieving fault diagnosis and providing corresponding response measures [12].

2) *Fault clustering center.* The fault situation of the all fiber current transformer can be reflected through the data obtained from its monitoring. For any all fiber current transformer, obtain the time-domain characteristics of its monitoring data, including rise time, fall time, pulse width, and duration; Frequency domain features refer to spectral peaks in the frequency domain; Shape parameters, namely skewness and kurtosis; And 11 feature quantities, including time center of gravity, equivalent duration, frequency center of gravity, and equivalent frequency width, are used to construct its feature vector for time-frequency joint features. Construct feature vectors for each fault case and classify them according to the fault type to form a feature space for that fault type. Finally, calculate the clustering center of this feature space and use it as a knowledge base.

3) *Fault Tree.* Using the fault tree as a knowledge base, after inferring the fault type of the tested current transformer, the FMEA table and fault tree can be combined to diagnose the fault of the tested current transformer, providing the fault location, fault consequences, and response measures.

3.2. Inference Machine. The inference machine provides results based on user input data, utilizing knowledge from the knowledge base and following certain inference rules. Rules are generally expressed as

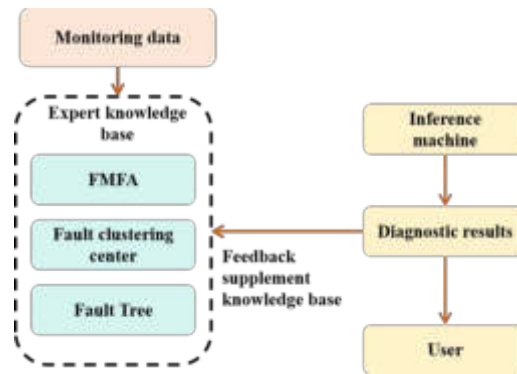


Fig. 3.3: Diagnosis process of expert system

IF-THEN, where IF is the premise and THE is the inference. When diagnosing faults in all fiber current transformers, the real-time monitoring values of the all fiber current transformer are taken as input, and then feature vectors are extracted to calculate the distance between this feature vector and each fault cluster center. The fault type with the smallest distance corresponds to the fault type of the all fiber current transformer [13]. Match the operating conditions (including operating conditions and environmental conditions) of existing all fiber current transformers with the basic events of the fault tree, and provide the final fault diagnosis result in combination with FMEA.

The functional structure and diagnostic process of the fault tree based all fiber current transformer fault diagnosis expert system based on fault mode and impact analysis method are shown in Figure 3.2 and Figure 3.3, respectively.

4. Fault diagnosis examples. Based on simulation data of mathematical models for various fault types and existing fault data of all fiber current transformers, a total of 5569 pieces of data were extracted. From these data, 11 quantities were extracted, including rise time, fall time, pulse width, duration, frequency domain spectral peak, skewness, kurtosis, time center of gravity, equivalent duration, frequency center of gravity, and equivalent frequency width, to construct feature vectors. Normalize the maximum and minimum values for each feature quantity to obtain the clustering centers of each fault type, as shown in Figure 4.1.

Extract the rise time, fall time, pulse width, duration, frequency-domain spectral peak, skewness, kurtosis, time center of gravity, equivalent duration, frequency center of gravity, and equivalent frequency width of all fiber current transformer fault cases obtained on site, and normalize them to obtain their corresponding feature vectors, as shown in Figure 4.2.

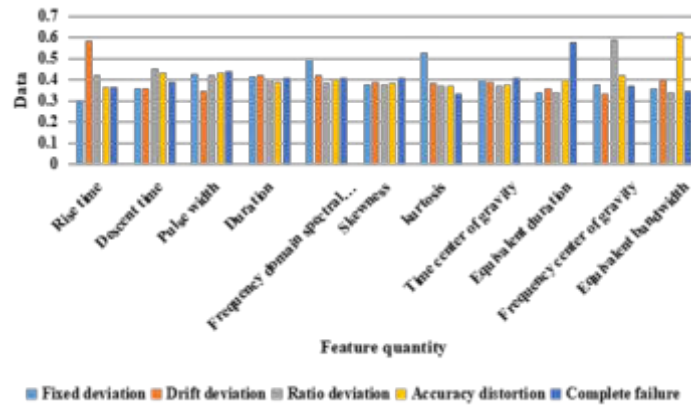


Fig. 4.1: Clustering center for each fault type

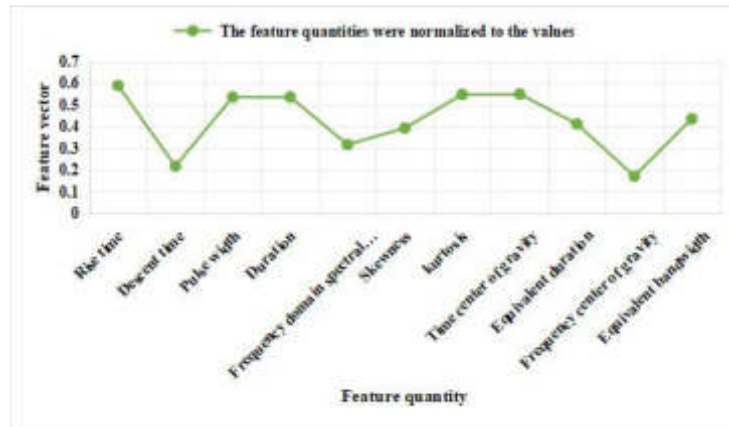


Fig. 4.2: Characteristic vector of the fault case

Table 4.1: Euclidean distance between the feature vectors of the fault cases and the cluster center of each fault type

Fixed deviation	Drift deviation	Change ratio deviation	Accuracy distortion	Complete failure
0.4914	0.4041	0.6112	0.5408	0.5042

Calculate the Euclidean distance between the feature vectors of the fault cases in Figure 4.2 and the clustering centers in Figure 4.1, as shown in Table 4.1.

According to the calculation results in Table 4.1, it can be seen that the on-site fault case is closest to the cluster center of the drift deviation fault, so it belongs to the drift deviation fault. Further extract the on-site maintenance report, which indicates that the operating temperature of the all fiber current transformer is relatively high[14,15,16,17]. Therefore, a fault diagnosis conclusion can be drawn: The drift deviation fault of the all fiber current transformer may be caused by a decrease in fiber temperature performance or a decrease in sensor unit temperature performance. It is recommended to replace the fiber or conduct performance testing on the sensor unit[18,19,20]. This conclusion is consistent with the conclusion given in the on-site report of the all fiber current transformer that high temperatures lead to a decrease in the performance of the optical

transmission system.

5. Conclusion. The author analyzed the impact of each part of the full fiber current transformer on overall stability, provided a reliability diagram of the full fiber current transformer, and analyzed the types of faults from the perspective of monitoring data, constructing a fault mode and impact analysis Table. A fault tree for all fiber optic current transformers was constructed based on the fault mode and impact analysis table for qualitative analysis of faulty current transformers. A fault diagnosis expert system for all fiber current transformers is proposed, which uses the clustering center, FMEA table, and fault tree of the faulty current transformer as the knowledge base of the expert system. Based on the corresponding inference rules, any all fiber current transformer is diagnosed, and the fault type, fault mode, fault component, and corresponding response measures are provided. The on-site fault case verified the usability of the method proposed by the author.

REFERENCES

- [1] Sui, X., Li, J., Wang, Z., Qi, Y., Li, G., & Zheng, N., et al. (2023). Research on power transformer fault diagnosis based on improved wavelet packet energy and hidden markov model. 2023 IEEE 6th International Electrical and Energy Conference (CIEEC),55(7), 3167-3172.
- [2] Yang, M. X., Dang, L., & Wen, T. (2022). Research on failure diagnosis method of a rocket borne micro electro-mechanical systems recorder based on fault tree. Journal of Nanoelectronics and Optoelectronics, 32(01), 37-70.
- [3] Yang, F., & Li, X. (2022). Research on fault diagnosis of spark discharge in transformer based on sound+improved bp neural network. Springer, Singapore, 293(3), 110363.
- [4] Zhang, R., Geng, L., & Liu, W. (2023). Research on static fault tree analysis method for inerting system safety based on random number generation. Aircraft engineering and aerospace technology, 34(1), 54-65.
- [5] Pang, J., Dai, J., Zhou, H., & Li, Y. (2022). A new fault diagnosis method for quality control of electromagnet based on t-s fault tree and grey relation. International journal of reliability, quality and safety engineering, 48(1), 111-121.
- [6] Liu, J., Tan, H., Shi, Y., Ai, Y., Chen, S., & Zhang, C. (2022). Research on diagnosis and prediction method of stator interturn short-circuit fault of traction motor. Energies, 27(2), 283-293.
- [7] Bai, R., Shen, F., Zhao, Z., Zhang, Z., & Yu, Q. (2023). The analysis of the correlation between spt and cpt based on cnn-ga and liquefaction discrimination research. Tech Science Press, 138(6), 104639.
- [8] Zhu, H. L., Liu, S. S., Qu, Y. Y., Han, X. X., He, W., & Cao, Y. (2022). A new risk assessment method based on belief rule base and fault tree analysis:. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability, 236(3), 420-438.
- [9] Hui, R., Jing, X., Jinling, L., Yunzhe, W., & Guoyu, X. (2023). Research on fault diagnosis of photovoltaic modules based on infrared images and improved mobilenet-v3. Acta Energiæ Solaris Sinica, 44(8), 238-245.
- [10] Wang, H. (2022). Research on vibration data-driven fault diagnosis for iron core looseness of saturable reactor in uhvdc thyristor valve based on cvae-gan and multimodal feature integrated cnn. Energies, 15(85),69-72.
- [11] Liu, G., & Li, W. (2022). Dynamic reliability analysis approach based on fault tree and new process capability index. Quality and Reliability Engineering International, 38(2), 800-816.
- [12] Chu, W., Liu, T., Wang, Z., Liu, C., & Zhou, J. (2022). Research on the sparse optimization method of periodic weights and its application in bearing fault diagnosis. Mechanism and Machine Theory: Dynamics of Machine Systems Gears and Power Trandmissions Robots and Manipulator Systems Computer-Aided Design Methods, 17(4), 697-710.
- [13] Takahashi, N., & Toda, S. (2022). Mapping active faults and folds of the nagamachi-rifu line fault system based on high-resolution dem and a borehole dataset. Active Fault Research, 2022(56), 1-12.
- [14] Thango, B. A., Nnachi, A. F., Dlamini, G. A., & Bokoro, P. N. (2022). A novel approach to assess power transformer winding conditions using regression analysis and frequency response measurements. Energies, 15(7),63-65.
- [15] Yang, D. (2022). Research on fault diagnosis of hot die forging multi-station feeding manipulator. Journal of Physics: Conference Series, 2383(1), 012062-.
- [16] Yan, P., Chen, F., & Kan, X. (2023). Research on transformer fault diagnosis based on an iwho optimized ms1dcnn algorithm and lif spectrum. Analytical methods, 26(1), 280-290.
- [17] Cheng, J., Feng, Z., & Xiong, Y. (2022). Transformer fault diagnosis based on an improved sine cosine algorithm and bp neural network. Recent advances in electrical & electronic engineering, 21(4), 1550.
- [18] Thango, B. A., Nnachi, A. F., Dlamini, G. A., & Bokoro, P. N. (2022). A novel approach to assess power transformer winding conditions using regression analysis and frequency response measurements. Energies, 10(1), 461-468.
- [19] Yang, Z., Cen, J., Liu, X., Xiong, J., & Chen, H. (2022). Research on bearing fault diagnosis method based on transformer neural network. Measurement Science & Technology,85(8), 33.
- [20] Zhu, X., Hu, W., & Fan, H. (2022). Research on circuit fault diagnosis method based on multi-feature information fusion. 2022 13th International Conference on Reliability, Maintainability, and Safety (ICRMS), 65(1),280-284.

Edited by: Zhigao Zheng

Special issue on: Graph Powered Big Aerospace Data Processing

Received: Jan 15, 2024

Accepted: Mar 1, 2024



ETHICAL EVALUATION AND OPTIMIZATION OF ARTIFICIAL INTELLIGENCE ALGORITHMS BASED ON SELF SUPERVISED LEARNING

RUOYU DENG* AND YANG ZHAO†

Abstract. Active learning solves the problem of requiring a large amount of manpower and resources due to the large size of training samples. The core problem is how to select valuable samples to reduce annotation costs. Using neural networks as classifiers, most methods choose samples with large amounts of information without considering the issue of information redundancy between the selected samples. Through the study of redundancy issues, the author proposes a sample selection optimization method to reduce information redundancy. Using uncertainty methods to select samples with high information content to form a candidate sample set, and using latent variable vectors calculated in the network to represent sample information, the cosine distance between candidate samples is calculated using this vector to select subsets with large interval distance and low information redundancy. Compared with several uncertainty methods in the Mnist, Fashion mnist, and Cifar-10 datasets, this method reduces labeled samples by a minimum of 11% with the same sample accuracy. The higher the dimensionality of the feature vector calculated by CNN, the more candidate samples it contains, and the more information it contains. After being improved by self supervised learning algorithms, the effect becomes more significant. The more candidate samples are selected, the stronger the information redundancy. The better the performance of self supervised learning algorithms.

Key words: Artificial intelligence algorithms, Information redundancy, Cosine distance, Uncertainty method

1. Introduction. With the rapid development of artificial intelligence (AI), intelligent algorithms are being applied more and more widely in various fields, such as autonomous driving, medical diagnosis, financial analysis, etc [1]. The introduction of these algorithms has brought great convenience and benefits, but at the same time, it has also raised concerns about ethical issues related to these algorithms. In the decision-making and behavior of AI algorithms, there may be significant impacts on individuals, society, and the environment [2]. For example, in the field of autonomous driving, intelligent algorithms are responsible for determining the trajectory and behavior of vehicles, which is directly related to driving safety and road traffic order. In medical diagnosis, AI algorithms can assist doctors in disease diagnosis and treatment decisions, but their accuracy and safety are also factors that need to be considered. In financial analysis, intelligent algorithms can assist in investment decision-making and risk assessment, but the fairness and transparency of their decisions are also controversial [3].

Therefore, in order to ensure the security, fairness, and trustworthiness of AI algorithms, ethical evaluation and optimization have become crucial. The purpose of ethical evaluation is to examine whether the decisions and behaviors of algorithms comply with ethical and ethical standards, and to predict their potential impacts. This requires consideration of ethical considerations in algorithm design and training, as well as the potential risks and impacts that may arise during algorithm application. Firstly, in the process of algorithm design and training, it is necessary to consider whether the collection and use of data are legal and compliant [4]. For the processing of sensitive information, such as personal identity information, medical records, etc., relevant laws and regulations should be strictly followed to protect the privacy rights of users. At the same time, it is necessary to ensure the quality and reliability of the data, and avoid erroneous decisions made by algorithms due to data bias or incompleteness. In addition, the training process of the algorithm should also follow the principles of fairness and equality to avoid discriminatory results. Secondly, in the process of algorithm application, it is necessary to consider the transparency and interpretability of the algorithm. Intelligent algorithms are usually built based on machine learning and deep learning techniques, and their decision-making process is often black

*School of Marxism, Chengdu Technological University, Sichuan, Chengdu, 611731, China

†School of Automation Engineering, University of Electronic Science and Technology of China, Sichuan, Chengdu, 611731, China
(Corresponding author, dyixy1025@163.com)

box like, difficult to explain and understand. This is a challenge for users and relevant stakeholders as they are unable to understand how algorithms make decisions [5]. Therefore, the interpretability of algorithms has become an important ethical issue. In order to address this issue, researchers are working hard to develop interpretable AI algorithms and promoting the development of relevant standards and specifications. In addition, the fairness of algorithms is also an ethical issue that needs to be considered. The decision-making of intelligent algorithms may be affected by data bias and unfairness, such as discriminatory outcomes for certain specific groups. Therefore, it is necessary to review and calibrate the training data of the algorithm to avoid bias and discrimination. At the same time, it is necessary to establish supervision and feedback mechanisms to promptly correct unfair decisions made by algorithms and protect the rights and interests of users [6]. Finally, risk assessment and management of algorithms are also important aspects of ethical assessment. Intelligent algorithms may bring some potential risks, such as privacy breaches, security vulnerabilities, ethical conflicts, etc. Therefore, it is necessary to conduct a comprehensive assessment and management of these risks, take corresponding measures to reduce risks, and establish supervision and monitoring mechanisms to timely identify and solve problems [7].

The sample selection strategy is a core issue in the process of artificial intelligence algorithms. In the pool sample mode, different classifiers have different selection strategies. Under the condition of using SVM as a classifier, the information content of samples is clearly measured by the distance between samples and support vectors, such as SVM's batch pattern artificial intelligence algorithm method. However, considering the distribution problem between samples, Maximum Mean Discrepancy (MMD) is used to measure the distribution difference between sample sets, thereby ensuring the consistency of distribution between unlabeled and labeled sets, for example, the Batch Mode Active Learning (BMAL) method and the Discriminative and Representative Model Active Learning (DRMALS) method [8]. Furthermore, the Similarity based Sparse Modeling Representative Selection (DSMRS) and mutual information method were used to measure the similarity between sample sets, thereby reducing the redundancy between sample sets. For example, adaptive active learning methods, Convex Programming Active Learning (CPAL) methods, etc.

Under the condition of using neural networks as classifiers, similar to SVM, the closer the sample information is to the classification boundary, the greater the amount of information. Currently, most uncertainty methods are used for measurement. Due to the lack of clear explanation from neural networks, some scholars believe that the current method does not select samples close enough to the classification boundary and needs to reselect samples near the classification boundary, such as: The Bayesian Active Learning (DBAL) algorithm calculates the mean of the results after multiple dropouts as the final classification result; Attack the Unmarked Sample (DFAL) algorithm using the Deepool algorithm that generates adversarial samples; Considering the sample distribution relationship, in order to ensure the consistency of distribution between labeled and unlabeled sample sets, a Cost Effective Active Learning (CEAL) method with classifier self labeling is used to transform the sample selection problem into a K-center problem using Euclidean distance, these methods ensure distribution consistency and allow for the selection of samples with large amounts of information, but they do not solve the problem of information redundancy. The author proposes a self supervised learning algorithm to reduce redundancy and achieve better results.

2. Problem Description. This section defines the active learning problem, which indirectly measures the amount of sample information and the redundancy of sample information through latent variables in CNN with MLP. Combining the process of artificial intelligence algorithms, the problem of minimizing redundancy is defined. Assuming there are n samples of m classes [9]. The problem of artificial intelligence algorithm based on pool sample selection sample pattern is as follows:

Question 1. Assuming the number of labeled samples is n_L and $L = \{x_i | i = 1, 2, \dots, n_L\}$; The number of unlabeled samples is n_U , $U = \{x_i | i = 1, 2, \dots, n_U\}$, $x_i \in R^k$ and $n_L + n_U = n$; Sample label set: $Y = \{y_i | i = 1, 2, \dots, n_L\}$ and $y_i \in R^m$. The loss function of the CNN model is $l(L, Y; f(\theta))$, which is mapped to $R^{n_L \times k} \times R^{n_L \times m} \rightarrow R^{n_L}$.

Each time k samples are selected from U to form a S_i set and placed in L . The active learning problem is:

$$\min_L E[l(L, Y; f(\theta))] - E[;(L, Y; f(\theta))] \text{ s.t. } |S_i| = k \quad \text{and} \quad L = \bigcup_{i=1}^T S_i \quad (2.1)$$

The loss function is the information cross entropy function, which is:

$$l(L, Y, \theta) = -\frac{1}{n_L} \sum_{i=1}^{n_L} \sum_{j=1}^m \{y_i = j\} \log p(y_i = j | x_i; \theta) \quad (2.2)$$

In the formula, T is the number of iterations; $1 \{\cdot\}$ is the indicator function, which is 1 when CNN predicts correctly, otherwise $p(y_i = j | x_i; \theta)$ is the output result of class j after the Softmax process.

Due to the lack of clear interpretability of CNN, it is difficult to determine what specific information the sample has about CNN. Given this issue, it is assumed that the CNN is connected to an MLP fully connected layer after passing through a convolutional layer, and the output vector of the hidden layer in MLP is specified as a latent variable, abstracting sample information through latent variables. The modulus of latent variable vectors represents the amount of information, and information redundancy is represented by calculating the distance between samples. The commonly used distance measures include Euclidean distance and cosine distance. The cosine distance is more effective in calculating and yields the same conclusion, so cosine distance is chosen[10]. Furthermore, the inner product measures the redundancy of information between samples, meaning that a larger inner product indicates a higher similarity of latent variables and a higher redundancy of information between the two samples being compared.

Definition 1. For samples x_i and x_j , the latent variables calculated by CNN are x'_i and x'_j , and $x'_i, x'_j \in R^n \in R^n$, the information redundancy matrix R and information quantity I between n samples are:

$$R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1k} \\ r_{21} & r_{22} & \dots & r_{2k} \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ r_{k1} & r_{k2} & \dots & r_{kk} \end{pmatrix} \quad (2.3)$$

$$I = (I_1, I_2, \dots, I_n) \quad (2.4)$$

In order to solve problem 1, the goal is achieved by reducing redundancy by selecting a sample set with high information content and low redundancy. Question 1 becomes how to select the sample set S_i to minimize the redundancy between samples[11].

Question 2. Assuming that there is already a redundancy matrix R, select the sample set S_i from U to minimize the mean of R, that is:

$$\min_L \text{average}(R) \text{ s.t. } |S_i| = k \quad \text{and} \quad L = \bigcup_{i=1}^T S_i \quad (2.5)$$

3. Redundancy methods. The redundancy problem mainly occurs between multiple selections of S_i , between sample sets, or between samples in a single selection of S_i . Assuming that after each selection of S_i , the CNN converges on the set L after labeling, there is less information redundancy between the S_i sample sets selected in each iteration, and the redundancy problem mainly lies between the samples in the S_i set [12].

The analysis of the redundancy problem of S_i set is shown in Figure 3.1, where: circle and triangle represent two types of samples m_1 and m_2 ; The dashed line represents the initial classification boundary; The solid line is the classification boundary after selecting samples; Grid points are candidate sample sets; The real point is the selected sample set. The samples are divided into two categories: m_1 and m_2 , where $m_j^{(i)}$ represents the i-th subset of samples belonging to the j-th category. Assuming that some samples have been selected and labeled based on uncertainty methods. As shown in Figure 3.1 (a), several samples were selected from the $m_1^{(1)}, m_2^{(2)}$ sample set near the CNN classification boundary. After iterative training, the original dashed boundary became a solid boundary, and the result did not completely separate the two types of samples. It can be seen that the sample set selected based on uncertainty methods has information redundancy. In response to this issue, the author proposes a self supervised learning algorithm [13].

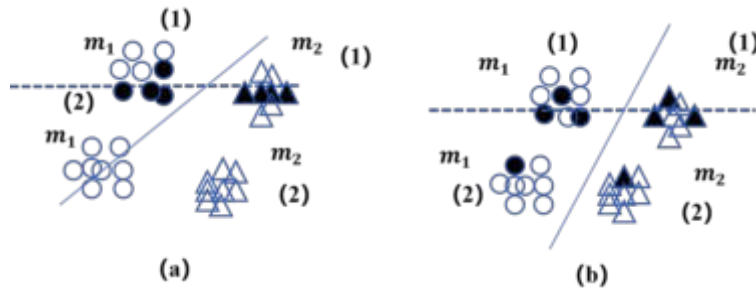


Fig. 3.1: Analysis of Redundancy Issues in S_i Sets

3.1. Uncertainty methods. The principle of sample selection based on the pooling sample pattern is mainly to select samples with high information content in the unlabeled pool to enable CNN to quickly fit the samples. High information content means that after CNN calculation, the probability of unlabeled samples in each classification is close to $\frac{1}{m}$, or they are uncertain in the most likely classification. Such samples close to the classifier boundary are the selected samples in Figure 3.1 (a). The current methods of uncertainty are as follows:

(1) Low credibility:

$$x^* = \operatorname{argmax}_x [1 - p(y_{max}|x_i; \theta)] \tag{3.1}$$

In the formula: $y_{max} = \max(y_i = j|x_i; \theta)$. Sort the maximum values of samples in various classification probabilities from small to large, and select the top K samples.

(2) Information entropy:

$$x^* = \operatorname{argmax}_x - \sum_i p(y_i = j|x_i; \theta) \log(p(y_i = j|x_i; \theta)) \tag{3.2}$$

Sort the top K samples in descending order of information entropy.

(3) Bayesian estimation:

$$p = \frac{1}{T} \sum_{t=1}^T p(y_i = j|x_i; \theta, dropout_t) \tag{3.3}$$

Average the classification results under multiple dropout values, and then select samples based on Equations 3.1 and 3.2.

3.2. Redundancy algorithm. As shown in Figure 3.1(a), the above uncertainty method only selects some samples as S_i sets based on the given calculation indicators, without considering sample redundancy, thus selecting meaningless samples. A self supervised learning algorithm is proposed based on this problem, as shown in Figure 3.1(b), which constructs a candidate sample set consisting of samples located near the CNN classification boundary, which includes all categories. Finally, select a subset of samples with low redundancy from the candidate sample set.

Select samples located near the CNN classification boundary based on uncertainty methods to form a candidate sample set. Calculate the cosine distance matrix between latent variables for all candidate samples:

$$D = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1k} \\ d_{21} & d_{22} & \dots & d_{2k} \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ d_{k1} & d_{k2} & \dots & d_{kk} \end{pmatrix} = \frac{R}{2\sqrt{I^T \times I}} \tag{3.4}$$

Using the distance matrix to obtain a sample set with low redundancy, that is, selecting the sample from the candidate set that is least similar to the labeled set each time. If L is an empty set, selecting the sample that is most similar to the candidate set[14].

3.3. Self supervised learning algorithms. From the above algorithms, it can be seen that the purpose of using cosine distance and redundancy algorithms is to select sample sets with category diversity and low redundancy from the candidate sample set.

Assuming that K samples are selected in an unlabeled sample pool in one round, and the LeNet5 network is used as a classifier with S convolutional kernels, with a single convolutional kernel computation time of t_s , the time complexity analysis of self supervised learning algorithms and uncertainty methods is as follows:

Uncertainty methods:

$$T_U = T_s + T_k \quad (3.5)$$

Self supervised learning algorithms:

$$T_D = T_s + T_k + T_R \quad (3.6)$$

In the formula, T_s is the time taken for network training; T_k is the sorting selection sample time; T_R is the redundancy calculation time.

$$T_s = t_s K S = O(t_s S) \quad (3.7)$$

$$T_k = K = O(1) \quad T_R = N K = O(N) \quad (3.8)$$

Obviously, $t_s \gg N > 1$, then $T_s \gg T_R > T_k$, the time is mainly spent on training the neural network. Therefore, Equations 3.1 and 3.2 have the following relationship:

$$T_U \approx T_D \approx O(t_s S) \quad (3.9)$$

From Equation 3.5, it can be concluded that the running time of the two methods is roughly the same[15].

4. Experiments and Result Analysis.

4.1. Datasets and Network Structures. Multiple experiments were conducted on Mnist, Fashion mnist, and Cifar-10 on the Lenet and NIN models. The neural network structure is shown in Tables 4.1 and 4.2, and the description of the dataset is as follows:

- (1) Mnist: 28×28 grayscale images, totaling 10 categories. Used for recognizing handwritten digit datasets, including 50000 for training, 5000 for validation, and 10000 for testing. 10000 samples were used as unlabeled sample pools in the experiment.
- (2) Fashion latest: 28×28 grayscale images, totaling 10 categories. Used to identify fashion clothing datasets, including 50000 training sets and 10000 testing sets. Due to its higher complexity than Mnist, the experiment used 20000 samples as an unlabeled sample pool.
- (3) Cifar-10: $32 \times \text{thirty-two} \times 3$ color images, totaling 10 categories. A small dataset for identifying universal objects, including 50000 training sets and 10000 testing sets. 20000 samples were used as unlabeled sample pools in the experiment[16].

4.2. Experimental parameters. In order to reduce the impact of experimental randomness, each dataset experiment had an average of 5 results. In each experiment, in order to avoid the model being biased during the training process, the validation set selected for each iteration is to randomly and uniformly extract 2% of samples from each class in the existing labeled set, and the CNN initialization network parameters are the same in each dataset experiment[17]. The experiment used the Keras toolkit on the Python platform to compare self supervised learning algorithms with low credibility, information entropy, and Bayesian methods in uncertainty methods multiple times. For the Mnist dataset, $n_0=100$, $T=15$, $N=3$, $K=100$, feature vector length 128, and network structure are shown in Table 4.1. For the Fashion mnist dataset, $n_0=200$, $T=20$, $N=10$, $K=150$, the feature vector length is 256, the network structure is shown in Table 4.1, and the output length of the Fc layer is changed to 256[18]. Considering the complexity of the Cifar-10 dataset and network training issues, in order to reduce overfitting and result stability, the dropout value is reduced, $n_0=1\ 000$ $T = 20$ $N = 30$ $K = 150$ dropout = 0.25 the length of the feature vector is 512.

Table 4.1: Mnist and Fashion mnist experimental network structures

type	Core size/step size	Output Size
convolution	$3 \times 3 / 1$	$28 \times 28 \times 32$
Pooling	$2 \times 2 / 2$	$13 \times 13 \times 64$
convolution	$3 \times 3 / 1$	$13 \times 13 \times 64$
Pooling	$2 \times 2 / 2$	$6 \times 6 \times 64$
Fc(dropout50%)	—	$1 \times 1 \times 128$
Fc	—	$1 \times 1 \times 10$
Softmax	—	$1 \times 1 \times 10$

Table 4.2: Cifar-10 Experimental Network Structure

type	Core size/step size	Output Size
convolution	$5 \times 5 / 1$	$32 \times 32 \times 192$
Batch normalization	—	$32 \times 32 \times 192$
convolution	$1 \times 1 / 1$	$32 \times 32 \times 160$
convolution	$1 \times 1 / 1$	$32 \times 32 \times 96$
Pooling	$3 \times 3 / 2$	$15 \times 15 \times 96$
convolution	$5 \times 5 / 1$	$15 \times 15 \times 192$
Batch normalization	—	$15 \times 15 \times 192$
convolution	$1 \times 1 / 1$	$15 \times 15 \times 192$
convolution	$1 \times 1 / 1$	$15 \times 15 \times 192$
Pooling	$3 \times 3 / 2$	$7 \times 7 \times 192$
convolution	$3 \times 3 / 1$	$7 \times 7 \times 192$
BN	—	$7 \times 7 \times 192$
convolution	$1 \times 1 / 1$	$7 \times 7 \times 192$
convolution	$1 \times 1 / 1$	$7 \times 7 \times 64$
Fc(dropout25%)	—	$1 \times 1 \times 512$
Fc	—	$1 \times 1 \times 10$
Softmax	—	$1 \times 1 \times 10$

4.3. Analysis of experimental results. In order to verify the effectiveness of self supervised learning algorithms, different datasets use different network structures due to their varying complexity. The results are shown in Figure 4.1.

It can be clearly seen that self supervised learning algorithms have significant improvements in the existing uncertainty methods. Compared to the samples required for the highest accuracy of the uncertainty method, in Mnist, when the uncertainty method reaches 98%, the entropy method reduces the maximum number of samples by 28%, while the Bayesian method reduces the minimum number of samples by 16%; In Fashion mnist, when the uncertainty method reaches 85%, the maximum decrease in lead is 30%, and the minimum decrease in entropy is 14%; In Cifar-10, the three methods achieved a maximum lead reduction of 22% and a minimum sample reduction of 11% for Bayesian when achieving an accuracy of 52%. From the above results analysis, it can be seen that the Self supervised learning algorithm (SSL) reduces samples by up to 30% and the lowest by 11% in the three methods.

From these results, it can be found that the information of the samples selected by the original three uncertainty methods is redundant for the classifier. Because self supervised learning algorithms mainly select sample sets with low redundancy through redundancy algorithms. The accuracy in Figure 4.1 increases, and under the same number of samples, the amount of information increases, resulting in a decrease in redundancy. Self supervised learning algorithms can effectively enhance uncertainty methods at the cost of time [19].

In order to reduce the impact of latent variable feature vector length and candidate sample number on sample redundancy research, experiments were conducted on latent variable feature vector length and candidate

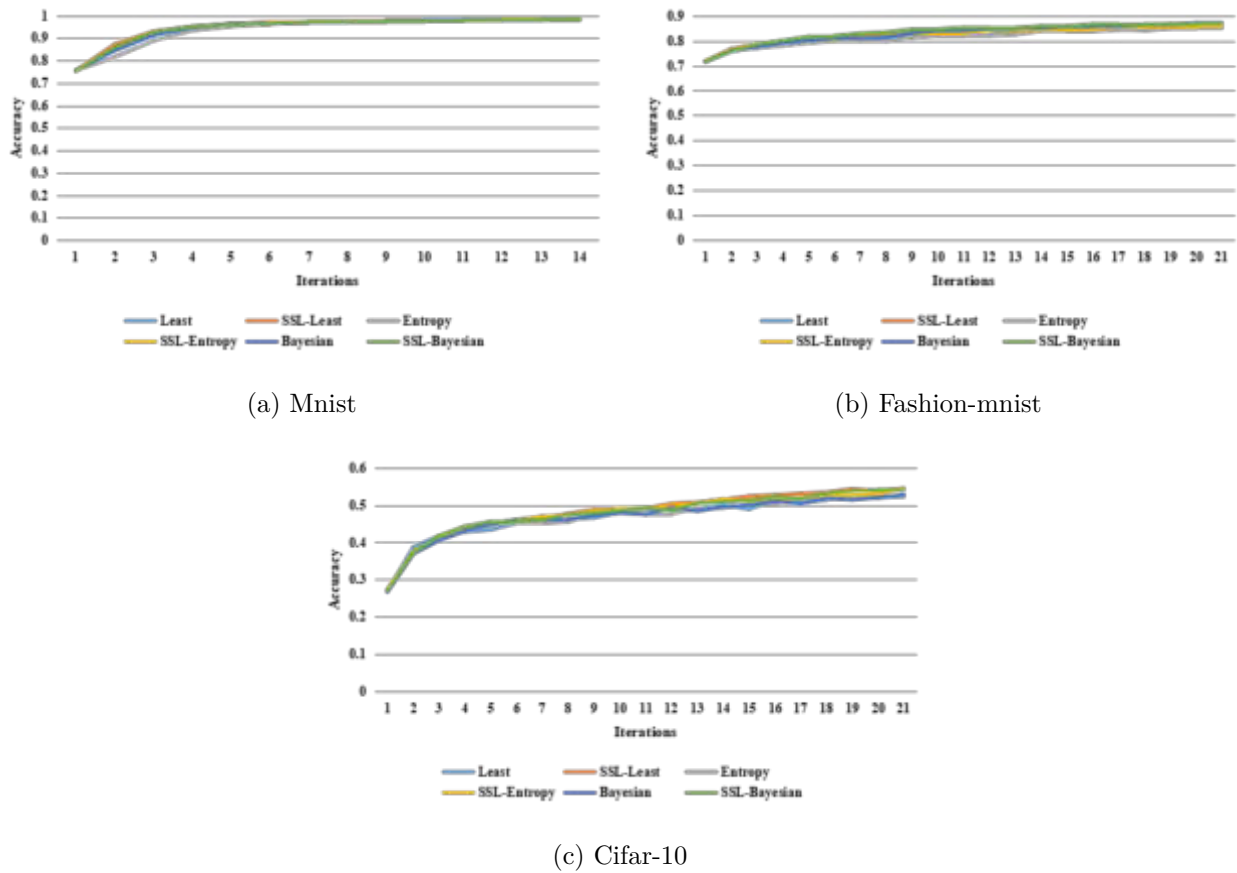


Fig. 4.1: Experimental results of SSL algorithm on different datasets

sample number. Under the conditions of 128 and 256 feature vectors, as well as 300 and 1000 candidate samples, in Fashion minimum, two sets of experiments were conducted using the self supervised learning algorithm - fast method. The experimental results on the relationship between information redundancy and the length of feature vectors and the number of candidate samples are shown in Figures 4.2 and 4.3.

The experiment in Figure 4.2 is conducted on the Fashion mnist dataset, with a candidate sample size of 1000 and feature vector lengths of 128 and 256, respectively. Due to the different widths of CNN in the last layer, the initialization network parameters are inconsistent. To address this issue, this experiment aims to ensure that the initialization network performs equally on the test set as much as possible. It can be seen that when the model accuracy reaches 80%, the former has 100 more labeled samples than the latter. So, the longer the feature vector, the more information it carries. The experiment in Figure 4.3 is conducted on the same dataset, with $n_0=200$, $T=20$, $K=100$, and a feature vector length of 256. The sample size of the candidate set is 300 and 1000, respectively, and the initialization network parameters are the same. Similarly, when the model accuracy reaches 80%, the former uses 100 more labeled samples than the latter. Through this experiment, it can be compared that if the number of candidate samples is small, the information contained in the candidate sample set is insufficient, but the redundancy is small, which will be affected by the number of samples[20]. Through the above experiments, it can be seen that the higher the dimensionality of the feature vectors calculated by CNN, the more candidate samples contain more information, and the more obvious the effect is after being improved by self supervised learning algorithms. The more candidate samples are selected, the stronger the information redundancy. The better the performance of self supervised learning algorithms.

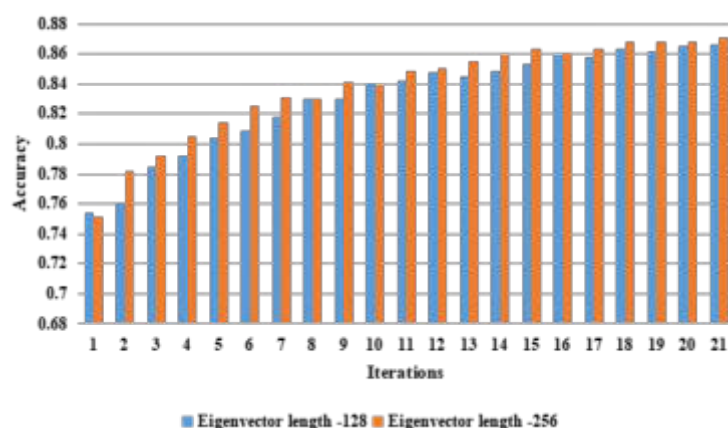


Fig. 4.2: The length of feature vectors in self supervised learning algorithms

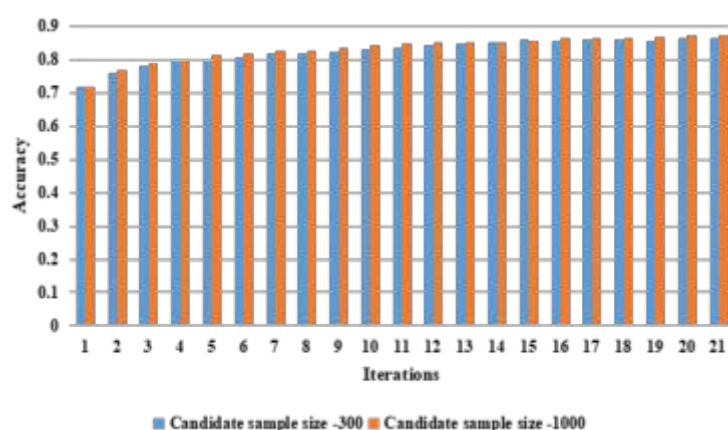


Fig. 4.3: Number of candidate samples in self supervised learning algorithms

5. Conclusion. In the sample selection mode of artificial intelligence algorithm pool, the author proposes a self supervised learning algorithm to reduce sample information redundancy. Using uncertainty methods, a large number of candidate samples are selected by CNN to form a candidate set. In the candidate set, the cosine distance relationship of the samples is used for a second screening, resulting in a sample set with large information content and low redundancy. This method can effectively reduce sample data redundancy and further reduce the number of labeled samples required by the model. In the future, uncertainty methods can be further optimized.

6. Acknowledgement.

1.2023 The third batch of excellent ideological and political work projects in universities in Sichuan Province & 2022 School-level excellent ideological and political work project: Reconstruction and exploration of the all-media practical education system deeply integrated by "E-Preaching"

2.2022 Project of College of Traditional Chinese Culture of Chengdu Technological University: "Research on the Educational Mechanism of Integrating Industrial Cultural values into College Curriculum Ideology and Politics" Project number: ZHY202205"

REFERENCES

- [1] Zeng, Y., & Xu, X. (2022). The construction and optimization of an ai education evaluation indicator based on intelligent algorithms. *International journal of cognitive informatics and natural intelligence*, 187(7), 7-11.
- [2] Wang, Y., & Na, K. S. (2022). Innovative research on english teaching model based on artificial intelligence and wireless communication. *International journal of reliability, quality and safety engineering*, 26(1), 63-87.
- [3] Song, J. Y., Li, Z. X., Li, Y. X., & Han, D. S. (2022). The optimization method of mechanical fault diagnosis based on artificial intelligence technology. *Journal of Physics: Conference Series*, 2158(1), 012033.
- [4] Wang, Y., Guo, J., Wang, J., Wu, C., Song, S., & Huang, G. (2023). Erratum to meta-semi: a meta-learning approach for semi-supervised learning. *CAAI Artificial Intelligence Research*, 2, null-null, 15(4), 308-314.
- [5] Loukas, C. (2022). Surgical gesture recognition in laparoscopic tasks based on the transformer network and self-supervised learning. *Bioengineering*, 9(7),52-56.
- [6] Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., & Eke, C. I., et al. (2022). A comprehensive survey of clustering algorithms: state-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110(47), 104743.
- [7] Alwi, S., Salleh, M., Abu, M. F., Ismail, A. F., Abbas, M. S., & Fadzilah, A. H. H. (2023). Concept of integration of blockchain and artificial intelligence. *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 15(7),1160-1163.
- [8] Jangmin, O. H. (2022). Short-term stock price prediction by supervised learning of rapid volume decrease patterns. *IEICE Transactions on Information and Systems*, E105.D(8), 1431-1442.
- [9] Xueqin Lü, Deng, R., Li, X., & Wu, Y. (2022). Comprehensive performance evaluation and optimization of hybrid power robot based on proton exchange membrane fuel cell. *International Journal of Energy Research*, 46(2), 1934-1950.
- [10] ZhaoHongyu, LyuFang, & LuoYalan. (2022). Research on the effect of online marketing based on multimodel fusion and artificial intelligence in the context of big data. *Security and Communication Networks*, 40(3), 422-440.
- [11] Mubarak, M. A. (2023). Sustainably developing in a digital world: harnessing artificial intelligence to meet the imperatives of work-based learning in industry 5.0. *Development and Learning in Organizations: An International Journal*, 37(3), 18-20.
- [12] Mei, Z. (2023). 3d image analysis of sports technical features and sports training methods based on artificial intelligence. *Journal of Testing and Evaluation: A Multidisciplinary Forum for Applied Sciences and Engineering*, 34(20), 18125-18141.
- [13] Zhu, W., Zhang, T., Wu, Y., Li, S., & Li, Z. (2022). Research on optimization of an enterprise financial risk early warning method based on the ds-rf model. *International review of financial analysis*(May), 16(1), 79-93.
- [14] Li, H., Xu, Q., Wang, Q., & Tang, B. (2023). A review of intelligent verification system for distribution automation terminal based on artificial intelligence algorithms. *Journal of Cloud Computing*, 12(1),85-89.
- [15] Nguyen, N., & Chang, J. M. (2022). Csnas: contrastive self-supervised learning neural architecture search via sequential model-based optimization. *IEEE Transactions on Artificial Intelligence*,65(4), 3.
- [16] Rekawek, P., Herbst, E. A., Suri, A., Ford, B. P., Rajapakse, C. S., & Panchal, N. (2023). Machine learning and artificial intelligence: a web-based implant failure and peri-implantitis prediction model for clinicians. *The International journal of oral & maxillofacial implants*,11(03), 176-191.
- [17] Yang, W., Cheng, X., Zhao, Y., Qian, R., & Li, J. (2023). Depth edge and structure optimization-based end-to-end self-supervised stereo matching. *International Journal of Pattern Recognition and Artificial Intelligence*, 37(13),63-65.
- [18] Villar, A., & Andrade, C. R. V. D. (2024). Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study. *Discover Artificial Intelligence*, 4(1),78-79.
- [19] Leng, J., Wang, D., Ma, X., Yu, P., Wei, L., & Chen, W. (2022). Bi-level artificial intelligence model for risk classification of acute respiratory diseases based on chinese clinical data. *Applied intelligence (Dordrecht, Netherlands)*, 52(11), 13114-13131.
- [20] Yin, J., Qiu, J. J., Liu, J. Y., Li, Y. Y., Lao, Q. C., & Zhong, X. R., et al. (2023). Differential diagnosis of dcis and fibroadenoma based on ultrasound images: a difference-based self-supervised approach. *Interdisciplinary Sciences: Computational Life Sciences*, 15(2), 262-272.

Edited by: Zhigao Zheng

Special issue on: Graph Powered Big Aerospace Data Processing

Received: Jan 15, 2024

Accepted: Mar 3, 2024



MACHINE LEARNING ALGORITHMS IN SUPPLY CHAIN COORDINATION SIMULATION AND OPTIMIZATION

QINGPING ZHANG* AND YI HEI†

Abstract. In response to the current situation of poor adaptive learning performance in Agent production and sales negotiation and dynamic changes in negotiation environment, the author proposes a method based on machine learning algorithms. Consider the impact of conflict level, cooperation possibility, and negotiation remaining time on negotiations in a dynamic negotiation environment, and use the entropy method to determine the weights of three influencing factors and perform linear weighting. Based on the differences in current negotiation topics, a concession amplitude prediction model based on dynamic selective ensemble learning is constructed, and an optimization strategy for supply chain production and sales negotiation is proposed. The experimental results indicate that, in the adaptive negotiation strategy of a regular SVM single learning machine, the joint utility of the most successfully negotiated agents falls within the interval $[0.55, 0.70]$, while the author's ensemble learning strategy mainly focuses on $[0.6, 0.8]$, the author's strategy is relatively superior to ordinary learning strategies in terms of both the number of successfully negotiated agents and the joint utility. Compared with the single learning machine negotiation strategy, this strategy improves the success rate and joint utility of Agent adaptive learning, and ensures the benefits of both production and sales in the supply chain, achieving a mutually beneficial situation for both parties in cooperation.

Key words: Dynamic selective ensemble learning, Dynamic negotiation environment, Agent production and sales negotiation, Adaptive learning, Entropy method

1. Introduction. Supply chain coordination is a key issue in supply chain management, which involves collaboration and coordination among different stakeholders. In the supply chain, due to the different goals and constraints of each participant, problems such as information asymmetry, order lag, and inventory backlog are prone to occur, leading to inefficiency and instability of the supply chain. In order to achieve efficient operation of the supply chain and maximize overall benefits, researchers have proposed various supply chain coordination models and algorithms. The core goal of supply chain coordination is to ensure effective coordination and cooperation among various links in the supply chain through reasonable decision-making and collaboration mechanisms, in order to optimize the efficiency of the entire supply chain system. In actual supply chain management, in order to solve the problem of information asymmetry, some measures can be taken to improve the flow and sharing of information. For example, establishing an information platform to improve the coordination and flexibility of various links in the supply chain through information sharing and transmission [1]. At the same time, adopting appropriate reward and punishment mechanisms to encourage all parties involved to work together and reduce the problems caused by information asymmetry.

In addition, order lag and inventory backlog are common issues in supply chain coordination. In order to address these issues, some supply chain coordination models and algorithms can be adopted. For example, by establishing a supply chain coordination model based on demand forecasting, it is possible to accurately predict demand and make corresponding adjustments according to changes in demand, avoiding the problems of order lag and inventory backlog [2]. In addition, reasonable inventory management strategies such as first in, first out (FIFO) and regular inventory can be adopted to control inventory backlog and improve the operational efficiency of the supply chain.

Supply chain coordination can also be achieved by optimizing logistics transportation and distribution. In the supply chain, logistics transportation and distribution play a crucial role. By optimizing logistics transportation and distribution plans, transportation costs can be reduced, transportation efficiency can be improved,

*Business School, Shunde Polytechnic, Foshan, Guangdong, 528333, China (Corresponding author, 10370@sdpt.edu.cn)

†School of Economics and Management, Guangdong Vocational College of Post and Telecom, Guangzhou, Guangdong Province, 510630, China

and supply chain coordination and optimization can be achieved [3]. For example, centralized distribution can be adopted to reduce the frequency and distance of transportation, and lower transportation costs; At the same time, utilizing logistics technology and information systems to achieve visual management of logistics transportation and distribution, improving the efficiency and service quality of logistics transportation.

In addition to the above methods, supply chain coordination can also be achieved through reasonable partner selection and supplier management. In supply chain management, selecting suitable partners and suppliers is crucial for the coordination and stability of the supply chain. By evaluating and managing suppliers, we can ensure their quality and delivery time, and reduce the risks and problems they bring [4]. At the same time, establish long-term and stable cooperative relationships, strengthen communication and collaboration with suppliers, and improve the efficiency and stability of the supply chain. In short, supply chain coordination is a key issue in supply chain management. Through reasonable decision-making and collaboration mechanisms, problems such as information asymmetry, order lag, and inventory backlog can be solved, improving the efficiency and stability of the supply chain. In actual supply chain management, various means and methods can be adopted to achieve coordination and optimization of the supply chain, including information sharing, demand forecasting, inventory management, logistics transportation and distribution optimization, partner selection, and supplier management. Through continuous research and practice, supply chain coordination models and algorithms can be further improved, promoting the development and progress of supply chain management [5].

The author aims to study the application of machine learning algorithms in supply chain coordination simulation and optimization. By using machine learning algorithms to analyze a large amount of data in the supply chain, we can better understand the operational mechanisms and optimization methods of the supply chain. Specifically, the research objectives include the following aspects:

1. Analyze the characteristics and challenges of supply chain coordination problems: Through in-depth research on supply chain coordination problems, analyze the relationships and interactions between various links in the supply chain, and reveal the essence and challenges of supply chain coordination.
2. Explore the application of machine learning algorithms in supply chain coordination simulation: Using machine learning algorithms, construct a supply chain coordination simulation model, simulate the impact of different coordination strategies on supply chain performance, and further study the behavior and decision-making of all parties involved in the supply chain.
3. Propose a supply chain optimization method based on machine learning algorithms: Based on the analysis results of machine learning algorithms, design a supply chain optimization algorithm to improve the efficiency and stability of the supply chain by coordinating the decisions and actions of all parties involved.

Through the implementation of the above research objectives, the author aims to provide new ideas and methods for supply chain management, promote coordination and optimization of the supply chain, and improve overall operational effectiveness.

2. Agent Supply Chain Production and Sales Negotiation Model.

2.1. Production and sales negotiation framework. In negotiation, the entropy method is used to calculate the weight of the negotiation environment, and the impact factor of the negotiation environment on the concession amplitude of the issue is obtained through linear weighting. Using optimized strategies to predict the concession range of opponents and measuring the impact of other negotiating opponents on the current negotiation based on global utility, the proposal values for each issue in the next round are obtained[6]. After the negotiation, the negotiation agent selects the best partner based on their own wishes, as shown in Figure 2.1.

2.2. Negotiation parameters. The specific steps of a negotiation strategy based on dynamic selection ensemble learning are: 1) Description of the negotiation environment; 2) Define negotiation issues and essential elements; 3) Support adaptive learning, update concession amplitude, and propose counter proposals[7]. Therefore, an octet representation negotiation model is proposed, and the negotiation parameters are explained in Table 2.1.

$$NM = \{A, I, P, w, T, C, Lc, U\} \quad (2.1)$$

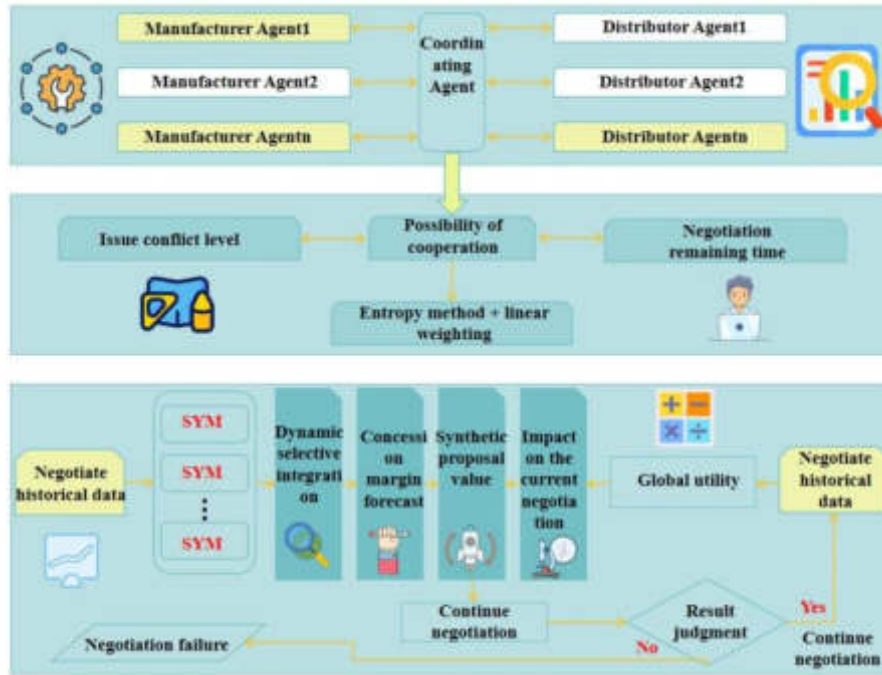


Fig. 2.1: Agent production and marketing negotiation framework

Table 2.1: Description of the negotiation parameters

Parameter	Describe
A	distributor manufacturer coordinator
I	Negotiation topic set
P	The agenda value for each round of negotiation
T	The remaining time for negotiation gradually decreases during the negotiation process
L_c	Conflict level between both parties
w	Participate in negotiating agent's weight vectors for each issue
C	The likelihood of cooperation between manufacturer and distributor agents
U	Evaluation of the proposed value of the opponent in the t-th round of negotiation on issue j

2.3. Negotiation Environment. For the expression of the negotiation environment, the author characterizes it using three factors: the level of issue conflict, remaining negotiation time, and the possibility of the best partner. Among them, the level of issue conflict is a positive indicator; The greater the remaining negotiation time and the possibility of the best partner, the smaller the concession made, which is a negative indicator.

Issue conflict level. The degree of conflict between the negotiating agent and the opponent on issue j, as shown in Equation 2.2.

$$Lc_t = \sum_{j=1}^n ep_j w_j \sqrt{|P_{t,j}^s - p_{t,j}^{opp}|^2} \quad (2.2)$$

Among them, $P_{t,j}^s$ represents the proposal value of our agent for issue j in round t, $P_{t,j}^{opp}$ represents the proposal value of our opponent for issue j in round t; ep_j represents the proportion of the number of opponents who are in conflict with the Agent regarding issue j in the total number of negotiations. Best partner possibility: decreases with the increase of competitors. A_i^t has A_c^t competitors and A_p^t trading parties in round t negotiation, and the possibility of A_i^t being considered as the preferred trading partner of the trading parties is shown in Equation 2.3.

$$c(A_i^t, A_p^t) = 1 - [(A_p^t - 1)/A_p^t]^{A_c^t} \tag{2.3}$$

Remaining negotiation time. The remaining time in the negotiation of round t is calculated as shown in Equation 2.4.

$$T(t, \tau, \lambda) = 1 - (t/\tau)^\lambda \tag{2.4}$$

Among them, τ is the deadline; λ is the optimal time limit for MDA.

2.4. Integration of negotiated environmental factors . The legitimate value method determines the weights of three factors in the negotiation environment, assuming that there are r agents participating in the negotiation, Amd calculates the conflict level of the manufacturer and distributor’s own issues, the possibility of the best partner, and the remaining negotiation time in each round of negotiation[8]. The positive and negative indicators are dimensionless according to Equations 2.5 and 2.6, forming a matrix as shown in Equation 2.7.

$$s_{r,i} = (s_{r,i} - \min\{s_i\})/(\max\{s_i\} - \min\{s_i\}) \tag{2.5}$$

$$s_{r,i} = (\max\{s_i\} - s_{r,i})/(\max\{s_i\} - \min\{s_i\}) \tag{2.6}$$

$$R = \begin{bmatrix} s_{11} & s_{12} & s_{13} \\ s_{21} & s_{22} & s_{23} \\ \dots & & \\ \dots & & \\ \dots & & \\ s_{r1} & s_{r2} & s_{r3} \end{bmatrix} \tag{2.7}$$

The legitimate weight of the jth environmental indicator is shown in Equation 2.8, with a legitimate weight of π_j . The entire negotiation environment should make concessions to the t-round negotiation agent θ_t . As follows:

$$H_j = -(1/\ln r) \sum_{i=1}^r (\frac{s_{ij}}{\sum_{i=1}^r s_{ij}}) \tag{2.8}$$

$$\pi_j = \frac{(1 - H_j)}{r - \sum_{i=1}^r H_j} \tag{2.9}$$

$$\theta_t = \pi_1 l c_t + \pi_2 c_t + \pi_3 T_t \tag{2.10}$$

3. Adaptive negotiation optimization strategy based on multi-agent.

3.1. Concession amplitude learning based on dynamic selective ensemble SVM. The predictive performance of each sub SVM learning machine varies for different data, and it is not advisable to use the same model function to estimate the concession amplitude for different issues. Based on the current issue value in the negotiation, use the nearest neighbor sample set as the evaluation sample to evaluate the performance of each sub model, and retain the sub models with better performance for integration[9]. In the negotiation, the K-means nearest neighbor search algorithm is used for each negotiation topic. The validation dataset is used to find k subsets of samples that are closest to the current value of the topic to be predicted, and the root mean square error is used as the evaluation criterion for the predictive performance of each sub model. Some sub models with poor predictive performance are eliminated, and the combined weights of each sub model are calculated to establish the final dynamic selective ensemble SVM model.

1. Generate an evaluation dataset using K-means. In order to predict the negotiation sequence P_q , let the number of nearest negotiation samples in the validation dataset P_L be k, calculate the distance between P_q and each negotiation data sample point P_i in P_L , and obtain the first k sample sets P_k .

$$P_D(P_q, P_i) = \sqrt{\sum_{i \in L} (P_q - P_i)^2} \quad (3.1)$$

2. SVM sub learning machine filtering. Input P_k sample sets, use root mean square error as the screening criterion, and select the corresponding top \bar{k} sub learning machine as the ensemble sub model of the predicted set P_q . The root mean square error of the i-th sub model is shown in Equation 3.2.

$$E_{ij} = \sqrt{\frac{\sum_{i=1}^k (\widetilde{c}_{ij} - C_{ij})^2}{k}} \quad (3.2)$$

Among them, \widetilde{c}_{ij} represents the predicted concession amplitude of the i-th sub learning machine for the next round of issue j; C_{ij} represents the actual concession amount for the next round of agenda item j.

3. Calculate the combined weights of each sub model. According to the root mean square error E_{ij} of the i-th sub model, the combined weight of this sub model is:

$$a_i = \left(\frac{1}{E_{ij}^2} \right) / \left(\sum_{i=1}^{\bar{k}} \frac{1}{E_{ij}^2} \right) \quad (3.3)$$

When all h sub learning machines are successfully trained, combined with the \bar{k} sub learning model with the smallest selection error for the current issue, four variables are inputted: the average concession amplitude value of the manufacturer and distributor agents in the first t rounds to hedge against the sudden issue j, and the difference in the proposed values of the manufacturer and distributor agents in the t round[10]. The predicted concession amplitude values for A_{fac} and A_{dis} in the t+1 round are obtained. The predicted output for each issue's concession amplitude is:

$$C_{t+1,j}^{fac/dis} = a_1 C_{1j} + a_2 C_{2j} + \dots + a_{\bar{k}} C_{\bar{k}j} \quad (3.4)$$

3.2. Utility Function Optimization. The global utility indicates that for positive issues, the larger the opponent's issue value is, the better, while for negative issues, the opposite is true. The utility evaluation functions for the negotiation object's issue value during t-round negotiation are shown in Equations 3.5 and 3.6, respectively.

$$U_{t,all}^+ = \sum_{j=1}^n w_{t,j} \left(\frac{p_{t,j}^{opp} - p_{t,j}^{min}}{p_{t,j}^{max} - p_{t,j}^{min}} \right) \quad (3.5)$$

$$U_{t,all}^- = \sum_{j=1}^n w_{t,j} \left(\frac{p_{t,j}^{max} - p_{t,j}^{opp}}{p_{t,j}^{max} - p_{t,j}^{min}} \right) \tag{3.6}$$

$$U_{t,all} = U_{t,all}^+ + U_{t,all}^- \tag{3.7}$$

Among them, $p_{t,j}^{opp}$ represents the current proposal value of the other party; $p_{t,j}^{max}$ represents the maximum value of the current proposal. Taking A_{fac} as an example, in the t -round negotiation, the global utility with each A_{dis} is calculated based on Equation 3.6. larger the $U_{t,all}$, the greater the utility obtained from the current A_{dis} negotiation, and the smaller the impact on the concession amplitude, making a larger concession[11].

The difference in local utility between the two rounds of negotiation is used to determine whether to stop the current negotiation process, as shown in Equation 3.8. According to the predicted concession range of A_{fac} on issue j in round $t+1$, the proposal value of distributor BB on issue j in round t of $C_{t+1}^{dis \rightarrow fac}$ negotiation is $p_{t,j}^{dis \rightarrow fac}$. The predicted proposal value of distributor $A_{t+1,j}^{dis \rightarrow fac}$ on issue j in round $t+1$ is shown in Equation 3.9.

$$U_{t,area} = \sum_{j=1}^j w_j^j p_j^{dis \rightarrow fac} \tag{3.8}$$

$$p_{t+1,j}^{dis \rightarrow fac} = p_{t,j}^{dis \rightarrow fac} + C_{t+1,j}^{dis \rightarrow fac} \tag{3.9}$$

Coordinate Amd with Equations 3.8 and 3.9 to calculate the difference between A_{dis} 's predicted utility value in round $t+1$ and the actual utility value in round t . When the difference is $\Delta U_{t+1,t} > 0$, continuing to negotiate A_{fac} 's utility will increase, but the utility has not been maximized yet. Otherwise, end the negotiation[12].

3.3. Topic proposal. Taking A_{fac} as an example, in multilateral adaptive negotiation, not only should the impact of the negotiation environment on the degree of concession be considered, but also the impact of other negotiation objects on the current negotiation. Therefore, the next round of proposal values for topic j are proposed by Equations 3.10 and 3.11.

$$p_{t+1,j}^{fac \rightarrow dis} = p_{t,j}^{dis \rightarrow fac} - p_{t,j}^{fac \rightarrow dis} \times (w_j) \times (a\theta + \beta C_{t+1}^{dir} + \frac{U_{t,all}^i - U_{t,all}^{min}}{U_{t,all}^{max} - U_{t,all}^{min}}) \tag{3.10}$$

$$p_{t+1,j}^{fac \rightarrow dis} = p_{t,j}^{dis \rightarrow fac} + p_{t,j}^{fac \rightarrow dis} \times (w_j) \times (a\theta + \beta C_{t+1}^{dir} + \frac{U_{t,all}^i - U_{t,all}^{min}}{U_{t,all}^{max} - U_{t,all}^{min}}) \tag{3.11}$$

Among them, Equation 3.10 represents cost based issues; Equation 3.11 represents profit oriented issues; θ Indicates the extent of concessions made under the influence of the negotiation environment; C_{t+1}^{fac} represents the concession amount of the opponent in the next round based on the ensemble learning algorithm; $U_{t,all}^i$ represents the global utility obtained from negotiating with the current negotiating party. The larger the value, the greater the concession, and vice versa[13].

3.4. Best Partner Selection. After the negotiation, the A_{fac} manufacturer makes a decision on the negotiation results, selects the appropriate A_{dis} , calculates the similarity of topics based on common neighbors according to the needs of the negotiation, and selects cooperation partners that are more suitable according to the similarity of topics, as shown in Equation 3.12.

$$S_{fac,dis} = (1 + e^{-D_{fac,dis}}) \times ||I_{fac} \cap I_{dis}|| \tag{3.12}$$

Among them, $D_{fac,dis}$ represents the issue gap between A_{fac} and A_{dis} ; $I_{fac} \cap I_{dis}$ represents the number of topics that A_{fac} and A_{dis} are satisfied with after successful negotiations, and selects the best partner based on similarity.

Table 4.1: Issue range and corresponding weights of manufacturers and distributors

Parameter type	Manufacturer Issue Range	Distributor Issue Range	Manufacturer weight	Distributor weight
Price/yuan	[3 000,3 800]	[3 000,3 300]	0 25	0 25
Quantity	[800,1 000]	[850,1 200]	0 25	0 25
Delivery time/month	[2 5,3]	[1,3]	0 25	0 25
Warranty period/month	[12,24]	[15,48]	0 15	0 10
Defective rate/%	[80,95]	[0,95]	0 10	0 15

4. Adaptive negotiation steps and examples.

4.1. Negotiation Steps.

- The specific steps for adaptive negotiation are as follows:
- Step 1: Negotiate initialization. Based on the negotiation targets of A_{fac} and A_{dis} , determine issue I, initialize issue weight W, maximum negotiation time T, and acceptable range of the issue, and normalize the issue.
- Step 2: A_{md} determines if the remaining negotiation time T has been exceeded. If it has been exceeded, the negotiation will be concluded; On the contrary, proceed to step 3.
- Step 3: A_{md} evaluates the negotiation environment and calculates according to Equation 2.10 θ and provide feedback to A_{fac} and A_{dis} who are currently negotiating [14].
- Step 4: A_{md} adds the new proposal to the negotiation history database. A_{fac} and A_{dis} divide the negotiation data into multiple samples, adjust the parameters of each sub learning model, and calculate the root mean square error E_{ij} of each model according to equation (12).
- Step 5: A_{fac} , A_{dis} calculates the weight of each sub learning model according to Equation 3.3 α , output the final strong combination learner, combined with the current proposal, to determine the next round of concession amplitude and measure $C_{t+1}^{fac/dis}$.
- Step 6: A_{md} calculates the current global utility based on Equation 3.7 and provides feedback to A_{fac} and A_{dis} who are currently negotiating.
- Step 7: Calculate the impact of the negotiation environment on the concession level of A_{fac} and A_{dis} being negotiated θ , based on the predicted concession amplitude in step 5, derive the counter proposal for each issue according to Equations 3.10 and 3.11, update the current utility value, negotiate the environmental conditions, and provide feedback to A_{md} .
- Step 8: A_{fac} , A_{dis} proposes a counter proposal. If the negotiating opponent accepts it, proceed to Step 9; Otherwise, proceed to step 3.
- Step 9: A_{md} will write the successfully negotiated agent to the database. After the quick negotiation activity, A_{fac} and A_{dis} determine their partners based on Equation 3.11.

4.2. Example analysis. In order to demonstrate the differences between the two algorithms, simulation experiments were conducted. Assuming that there are conflicts between multiple manufacturers and distributors in the mobile phone production and sales chain when formulating collaborative plans, negotiations should be conducted according to common learning strategies and integrated optimization strategies. Consider the price, quantity, delivery time, warranty time, and defect rate of the plan as negotiation topics [15]. According to expert experience, manufacturers are most concerned about price, quantity, and delivery time, followed by warranty time and defect rate; Distributors are most concerned about price, quantity, and delivery time, followed by defect rate and warranty time. Therefore, the issue range and weight of manufacturers and distributors are shown in Table 4.1.

Generate 50 manufacturers and 50 distributors for the two adaptive negotiation strategies proposed by the author, and simulate the conflict resolution of collaborative plans. Based on existing experience, it is assumed that the maximum negotiation time τ_{fac} is 25 and τ_{dis} is 20.

In Figures 4.1(a)-(b), the x-axis and x-axis represent respectively α, β . The y-axis represents the average joint utility value of manufacturers and distributors at the time of successful negotiation, with different combinations of values [16]. It can be seen that as Increase in size The difference between the average joint utility

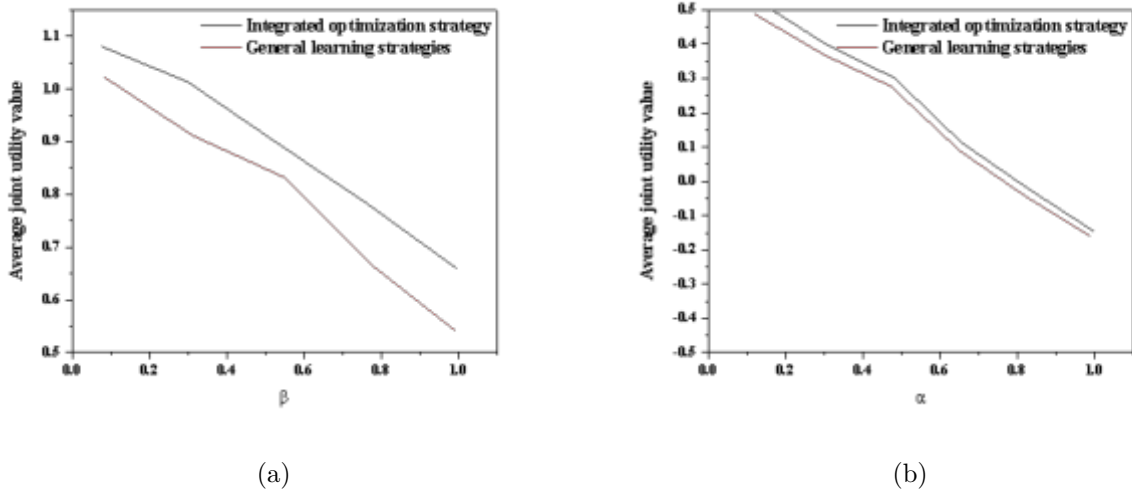


Fig. 4.1: Meverage joint utility simulation results for 2 strategies

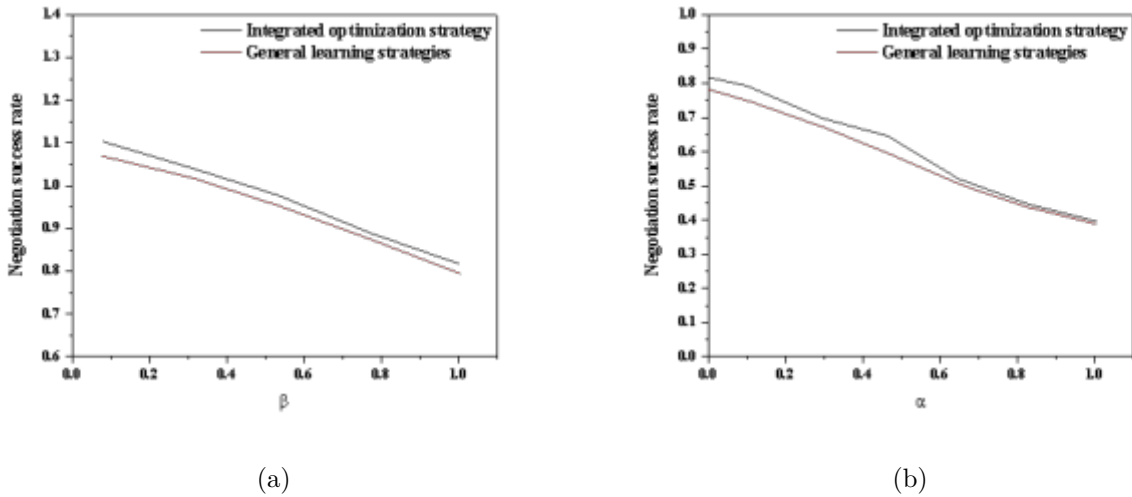


Fig. 4.2: Negotiated success rate simulation results for 2 strategies

of the two negotiation strategies is increasing as the decrease in β . When it is 0.6, the average combined utility of the two reaches its maximum. Therefore, the ensemble learning optimization strategy proposed in this article has better negotiation effectiveness than ordinary single learning machine adaptive strategies.

In Figures 4.2(a)-(b), the x-axis and y-axis represent respectively β and α . The y-axis represents the negotiation success rate of manufacturers and distributors with different value combinations. It can be seen that in general, the integrated learning optimization strategy proposed in this paper has a higher negotiation success rate than the ordinary single learning machine adaptive strategy. Therefore, the integrated learning optimization strategy can improve the success rate of production and sales negotiations to a certain extent [17].

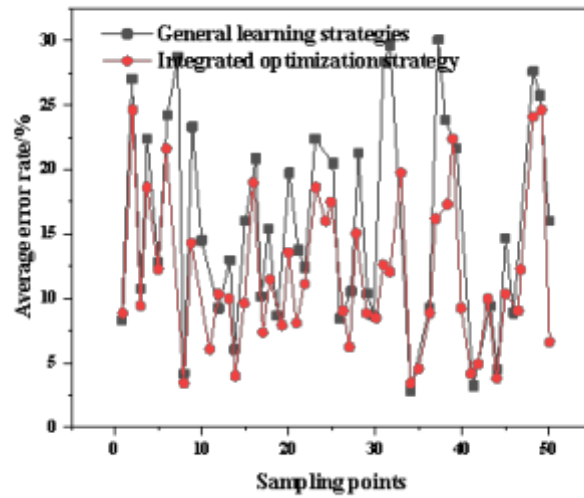


Fig. 4.3: Meverage error rate simulation results for 2 strategies

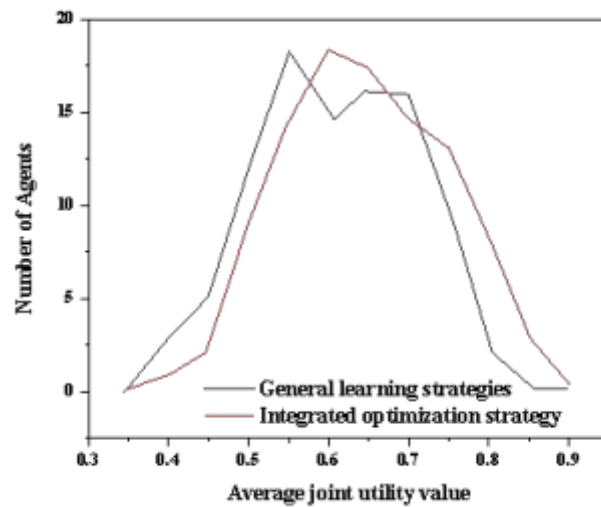


Fig. 4.4: Relationship between joint utility and successfully negotiated Agent

Figure 4.3 selects 50 manufacturers in the experiment to predict the average error rate of the opponent's concession amplitude on the same issue. Comparing the performance of the two strategies, it can be seen that in most cases, the author's ensemble learning strategy has lower error rates than the ordinary SVM single learning machine adaptive strategy [18].

From Figure 4.4, it can be seen that in the adaptive negotiation strategy of a regular SVM single learning machine, when the most successful agents are negotiated, their joint utility falls within the interval $[0.55, 0.70]$, while the author's ensemble learning strategy mainly focuses on $[0.6, 0.8]$, the author's strategy is relatively

superior to ordinary learning strategies in terms of both the number of successfully negotiated agents and the joint utility [19,20]. The conclusion drawn from the above is that the negotiation strategy based on dynamic selective ensemble learning performs relatively better than ordinary single learning machine adaptive negotiation strategies in terms of joint utility, negotiation success rate, average error rate, etc.

5. Conclusion. Resolving conflicts in supply chain production and sales collaboration is beneficial for improving the operational efficiency of the supply chain. On the basis of considering the impact of the environment on negotiation, the author proposes an adaptive negotiation strategy based on dynamic selective ensemble SVM, which can reduce the error of opponent prediction information and also consider the impact of other negotiation processes in multilateral negotiation. The experimental results show that compared with the adaptive negotiation strategy of ordinary single learning machines, this strategy can to some extent improve the negotiation success rate, conflict resolution efficiency, and the joint utility of manufacturers and distributors. The next step will be to study adaptive negotiation methods for resolving conflicts in supply chain production and sales collaboration based on multilateral negotiations, in order to improve the intelligence level of the supply chain.

REFERENCES

- [1] Ng, C. S. W., Amar, M. N., Ghahfarokhi, A. J., & Imsland, L. S. . (2023). A survey on the application of machine learning and metaheuristic algorithms for intelligent proxy modeling in reservoir simulation. *Computers & Chemical Engineering*, 170(8), 108107.
- [2] Zhang, W., Gu, X., Tang, L., Yin, Y., Liu, D., & Zhang, Y. . (2022). Application of machine learning, deep learning and optimization algorithms in geoenvironment and geoscience: comprehensive review and future challenge. *Gondwana research: international geoscience journal*, 38(99), 2434-2440.
- [3] Rajabi, M. M., & Chen, M. . (2022). Simulation-optimization with machine learning for geothermal reservoir recovery: current status and future prospects. *Advances in Geo-Energy Research*, 6(6), 451-453.
- [4] Shavaki, F. H., & Ghahnavieh, A. E. . (2022). Applications of deep learning into supply chain management: a systematic literature review and a framework for future research. *Artificial intelligence review*, 52(7), 1-43.
- [5] Keynia, F., & Memarzadeh, G. . (2022). A new financial loss/gain wind power forecasting method based on deep machine learning algorithm by using energy storage system. *IET generation, transmission & distribution*, 96(5), 16.
- [6] Liu, J., & Yeo, J. . (2023). Predicting the fracture propensity of amorphous silica using molecular dynamics simulations and machine learning. *International Journal of Applied Mechanics*, 15(10), 26-28.
- [7] Dingjun, H., Hong, F., & Jianchang, F. . (2023). Research on corporate social responsibility and product quality in an outsourcing supply chain. *Journal of Industrial and Management Optimization*, 19(4), 2485-2506.
- [8] Kim, T., Zhou, X. S., & Pendyala, R. M. . (2022). Computational graph-based framework for integrating econometric models and machine learning algorithms in emerging data-driven analytical environments. *Transportmetrica*, 74(8), 765-776.
- [9] Basu, B., Morrissey, P., & Gill, L. W. . (2022). Application of nonlinear time series and machine learning algorithms for forecasting groundwater flooding in a lowland karst area. *Water Resources Research*, 63(2), 58.
- [10] Ahmad, R. M., Ali, B. R., Fatma, A. J., Sinnott, R. O., Noura, A. D., & Saberi, M. M. . (2023). A review of genetic variant databases and machine learning tools for predicting the pathogenicity of breast cancer. *Briefings in Bioinformatics*, 74(1), 1.
- [11] Marousi, A., & Kokossis, A. . (2022). On the acceleration of global optimization algorithms by coupling cutting plane decomposition algorithms with machine learning and advanced data analytics. *Computers & Chemical Engineering: An International Journal of Computer Applications in Chemical Engineering*, 895(163), 163.
- [12] Ferdowsi, A., Valikhan-Anaraki, M., Farzin, S., & Mousavi, S. F. . (2022). A new combination approach for optimal design of sedimentation tanks based on hydrodynamic simulation model and machine learning algorithms. *Physics and chemistry of the earth*, 15(4), 7687-7713.
- [13] Momenitabar, M., Dehdari, E. Z., & Ghasemi, P. . (2022). Designing a sustainable bioethanol supply chain network: a combination of machine learning and meta-heuristic algorithms. *Industrial Crops and Products*, 147(1), 381-382.
- [14] Raza, S. A., Govindaluri, S. M., & Bhutta, M. K. . (2023). Research themes in machine learning applications in supply chain management using bibliometric analysis tools. *Benchmarking: An International Journal*, 30(3), 834-867.
- [15] Chen, C., Chen, C., Yaari, Z., Yaari, Z., Apfelbaum, E., & Apfelbaum, E., et al. (2022). Merging data curation and machine learning to improve nanomedicines. *Advanced Drug Delivery Reviews*, 183(7), 114172.
- [16] Yan, X. T., & Shang, Z. L. . (2022). Urban intelligent traffic signal coordination control system based on machine learning. *Advances in Transportation Studies*, 14(4), 413-430.
- [17] Minbashi, N., Sipil, H., Palmqvist, C., Bohlin, M., & Kordnejad, B. . (2023). Machine learning-assisted macro simulation for yard arrival prediction. *J. Rail Transp. Plan. Manag.*, 25(7), 100368.
- [18] Lee, Y., Park, B., Jo, M., Lee, J., & Lee, C. . (2022). A quantitative diagnostic method of feature coordination for machine learning model with massive data from rotary machine. *Expert Syst. Appl.*, 214(85), 119117.
- [19] Lim, H. G., Rychel, K., Sastry, A. V., Bentley, G. J., Mueller, J., & Schindel, H. S., et al. (2022). Machine-learning from

pseudomonas putida kt2440 transcriptomes reveals its transcriptional regulatory network. *Metabolic engineering*, 72(63), 297-310.

- [20] Yang, J., Chen, Y., Yao, H., & Zhang, B. . (2022). Machine learning-driven model to analyze particular conditions of contracts: a multifunctional and risk perspective. *Journal of management in engineering*, 21(5), 136-141.

Edited by: Zhigao Zheng

Special issue on: Graph Powered Big Aerospace Data Processing

Received: Jan 15, 2024

Accepted: Feb 18, 2024



SIMULATION OF SEGMENTED CLUSTERING OF CLOUD STORAGE DATA BASED ON NEURAL NETWORK MODELS AND PYTHON

GUOQING XIA* AND HUAZHEN CHEN[†]

Abstract. In order to improve the operational efficiency of traditional cloud storage data segmentation clustering methods, the author proposes a machine learning based cloud storage data segmentation clustering method. Reasonably extract multiple small datasets from the cloud storage database, which contain all natural clusters in the cloud storage database. Construct a similarity matrix based on the definition of similarity. Using nonlinear kernel principal component algorithm to measure the similarity of data in the similarity matrix, data with the same features are grouped together through similarity measurement, and a mixed Gaussian distribution probability density model is used to calculate the posterior probability of different categories of data, implement segmented clustering of cloud storage data by comparing probability sizes. The experimental results show that the proposed method can shorten the clustering running time, reduce the clustering variation to 29%, and effectively improve the smoothness of the clustering results.

Key words: Neural network models, Cloud storage, Segmented clustering of data, A mixed Gaussian distribution probability density model

1. Introduction. Neural networks are widely used in nonlinear system identification and control due to their strong learning ability. However, practical objects are full of uncertain factors, and many problems cannot be described with accurate mathematical models. Fuzzy systems utilize the knowledge and experience of experts to solve mathematical problems through natural language. Since the proposal of Adaptive Network Based Fuzzy Reference System (ANFIS), the research and application of this theory have made significant progress [1]. For a long time, fuzzy design networks were based on traditional Type 1 fuzzy logic systems. With the development of fuzzy set theory and the shortcomings of Type 1 fuzzy logic systems in describing the uncertainty of the objective world, the theory and application of Type 2 fuzzy logic systems have become a research hotspot in fuzzy theory in recent years. With the deepening of research on the identification and control of type 2 fuzzy logic systems in nonlinear systems, research on the identification and control of type 2 fuzzy neural networks in nonlinear systems is gradually increasing. At present, the structure of type-2 fuzzy neural networks mainly includes fixed network structure, self adjusting type-2 fuzzy neural network, self evolving type-2 fuzzy neural network, self-organizing interval type-2 fuzzy neural network, etc. Once the structure of interval type-2 fuzzy neural networks is determined, the next step is to learn the network parameters. Currently, the most commonly used parameter learning algorithm is the backpropagation (BP) algorithm. However, the BP algorithm is sensitive to initial values, and inappropriate initial values can cause the algorithm to diverge or converge to non optimal solutions. The structure of interval type-2 fuzzy logic systems is similar to that of traditional fuzzy logic systems, but it requires a key step, which is order reduction. The order reduction process first reduces the type-2 fuzzy set to a type-1 fuzzy set, and then obtains the precise output of the final interval type-2 fuzzy system using the conventional method of defuzzification of type-1 fuzzy sets. Currently, most interval type-2 fuzzy neural network systems use the Karnik Mendel (KM) reduction algorithm, which is an iterative optimization algorithm, firstly, it is necessary to sort the discrete points by size, so that the corresponding membership degree needs to be modified accordingly. The order can be reduced to obtain two switching points, and each switching point may not be the same. Therefore, when using the BP algorithm to learn parameters, the process is relatively cumbersome and there is no unified learning formula.

*School of Information Engineering, Guangdong Polytechnic, Foshan, Guangdong, 528041, China (Corresponding author, 13926104089@163.com)

[†]Department of Electronics, Software Engineering Institute of Guangzhou, Guangzhou, Guangdong, 510990, China

Clustering is the process of distinguishing and classifying things according to certain requirements and rules. In this process, there is no guidance from teachers or any prior cognitive information about classification, but only the similarity between things is used as the standard for their classification. Therefore, it belongs to the research content of unsupervised learning. Cluster analysis is the use of mathematical methods to process and study things that need to be classified [2-3]. Birds of a feather flock together. Clustering method is the process of grouping a collection of physical or abstract objects into multiple classes composed of similar objects, and clustering is an ancient problem. Since the emergence of human society, with the continuous development of human society, research on clustering has also been deepening. The continuous exploration of the world by humans requires distinguishing things that belong to different categories and recognizing the similarities of things in the same category. Multivariate statistical analysis, as a branch developed from classical statistics, is also an important branch of mathematical statistics, and cluster analysis belongs to multivariate statistical analysis. As one of the important research directions in statistics, cluster analysis has a profound theoretical foundation and has formed a systematic methodological system. In the field of pattern recognition, cluster analysis is also an important research direction for unsupervised pattern recognition.

Unlike classification, clustering does not rely on pre-defined classes and signed training practices, so clustering analysis is observational learning rather than example based learning. Through cluster analysis, a sample set without any prior knowledge is divided into several subsets based on specific classification rules. The samples within these subsets maintain high similarity, while the samples between subsets maintain low similarity as much as possible. In other words, samples in the same cluster should be as close as possible, while the distance between different cluster centers should be as far as possible. In many applications, data objects in certain classes can be treated as a whole. There are many clustering methods, and their principle is to divide the sample set that needs to be classified into several different categories based on similarity to represent the different characteristics of the system.

The minimum overlap between categories should be used to avoid repetition, which means that each category should contain as many similar samples as possible and have significant differences from each other. The purpose of clustering algorithms is to find several least similar sample centers that contain a set of similar samples, in order to maximize the representation of different features of the system. At the same time, each cluster center should contain a sufficient number of samples to ensure that it uses as few cluster centers as possible to represent the system. Clustering is a technique that studies the logical or physical interrelationships between data. Its analysis results not only reveal the inherent connections and differences between data, but also provide important basis for further data analysis and knowledge discovery. The purpose of clustering is to discover the essential clustering properties between samples, and it is an important component of data mining techniques. Data clustering is flourishing, and contributing fields include data mining, statistics, machine learning, spatial database technology, biology, and marketing. Nowadays, data clustering analysis has become a very active research topic [4]. Traditional clustering analysis is a hard division that strictly divides the objects to be analyzed into certain categories, with the property of either this or that. Therefore, the boundaries of this type of classification are clear. However, in reality, most objects do not have strict attributes, and they have intermediary properties in terms of form and category, with the nature of "this is also that". For example, people are classified according to their height as "tall people", "short people", and "not tall but not short people". However, as tall as possible, and as short as possible, this classification discrimination is a problem that classical classification cannot solve, so it is suitable for soft partitioning. The proposal of fuzzy set theory provides a powerful analytical tool for this soft partitioning, and people have begun to use fuzzy methods to handle clustering problems, namely fuzzy clustering analysis. Fuzzy clustering analysis extends the values of membership relationships from binary logic of 0,1 to the interval of [0,1], thereby more reasonably representing the mediating nature between things. Due to the uncertainty level of the sample belonging to each category obtained by fuzzy clustering, which expresses the "this is also that" property of the sample belonging to different categories, that is, the fuzziness of the sample's membership relationship to different categories, the description and expression of the real world are more reasonable, and have made significant progress in the theory of clustering analysis. The implementation process of data clustering is shown in Figure 1.1.

Fuzzy clustering theory has been widely used in the real world, promoting the improvement of social productivity. With the continuous development of practice, fuzzy clustering theory is also constantly improving

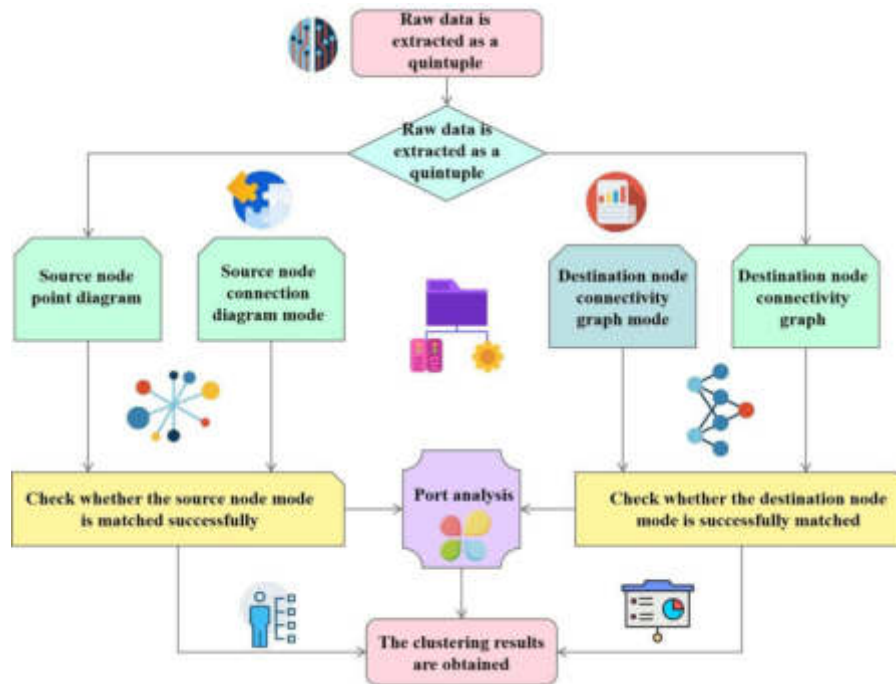


Fig. 1.1: Implementation process of data clustering

and enriching in practice. With the development of practice and the improvement of theory, fuzzy clustering has been widely applied in many fields, and has achieved satisfactory results and objective benefits. Its application scope involves many fields such as channel equalization in communication systems, codebook design in vectorized coding, time series prediction, neural network training, nonlinear system identification, parameter estimation, medical diagnosis, weather forecasting, food classification, water quality analysis, etc, the author mainly studies its application in nonlinear system identification. In nonlinear system identification, fuzzy clustering algorithms can be used for feature space partitioning and fuzzy rule extraction to construct fuzzy classifiers based on fuzzy if then rules.

The amount of data on the internet is showing an explosive growth trend with the rapid development of computers, leading to increased data storage costs, reduced reliability of data storage, and difficulties in managing large amounts of data, which have long plagued users. Users hope to obtain useful information from complex and diverse data, therefore, cloud storage data segmentation and clustering technology has emerged. However, in traditional cloud storage data segmentation and clustering methods, the use of Ethernet and TCP/IP network communication protocols improves and simplifies network protocols to reduce clustering delays, however, neglecting the diversity of data types can lead to problems such as long clustering time and large errors. In this context, researching accurate and efficient cloud storage data segmentation clustering methods has become a widely concerned focus in the current data clustering field, receiving widespread attention from industry insiders. At the same time, many good methods have also emerged. In response to the above issues, the author proposes a machine learning based segmented clustering method for cloud storage data. The experimental results show that the proposed method can shorten the running time of data segmentation clustering and improve the smoothness of clustering results.

2. Methods.

2.1. Overview of Spatial Data Cloud Storage. Space cloud storage mainly refers to the distributed storage and read/write of spatial data based on cloud computing technology. Currently, most of its research focuses on raster data storage and management, while there is relatively little research in the field of vector

data cloud storage. At the same time, research on spatial database systems and spatial databases as services (SDBaaS) in cloud environments based on cloud storage of spatial data is also in its early stages. We will elaborate on the theoretical research of spatial data cloud storage from two aspects: Raster and vector [5].

(1) *Overview of Cloud Computing.* Cloud computing, as a new type of distributed parallel computing model, can integrate computing power, storage space, and network resources from different regions, allowing any terminal (PC, Web, iOS, Android) user to access the cloud platform at any time and effectively use any resources on the platform. It fully utilizes the good scalability, powerful computing power, and cross regional information resource sharing function of distributed computing systems, by establishing a virtual and single system image for applications and data, users can easily and easily access all shared resources on the entire cloud platform, which is an excellent strategy to solve the deep sharing of current geospatial data and data processing services. Large cloud computing providers (Amazon, Google, Microsoft) encapsulate their infrastructure, platforms, frameworks, software, applications, and even data into network services, providing users and developers with standardized interfaces and on-demand billing models. Google's cloud computing platform provides basic technical support for applications such as Google App Engine, Google Map, and Google Trend. Google Cloud Platform was founded in 1996, when the founder of Google began attempting to combine multiple inexpensive PCs into a powerful computing platform to index billions of web pages; Nowadays, Google has a cloud computing platform consisting of over one million servers, providing different levels of services including IaaS, PaaS, and more.

Finally, Amazon created an elastic computing cloud EC2 based on this platform, providing users with on-demand online rental services for computing resources with surplus hardware resources. Microsoft also released its cloud computing platform product Azure in November 2009, which is a cloud application platform built on top of Microsoft's data center. It can manage and hook cloud application systems and provide a set of tools to facilitate developers to develop and debug cloud applications locally. So far, few have conducted research and development on cloud computing platforms as an integrated platform for spatial cloud storage and third-party service release and deployment. In order to achieve the goals of cloud computing and its high performance, low cost, and strong universality, cloud computing has developed a series of key technologies that support its functions such as data storage, data management, parallel computing, and concurrency control, including computer system virtualization technology, massive distributed storage technology, parallel programming mode, large-scale data management technology, distributed resource management technology, etc.

Virtualization technology is the underlying key to building an IaaS cloud platform. It quickly integrates and decomposes physical resources in a specific way, can dynamically organize multiple heterogeneous hardware, and achieve isolation between underlying physical hardware and other virtual machines on specific computers, achieving loose coupling between computing cluster hardware and software, and relieving severe dependencies between these architectures, thus meeting the scalability and scalability requirements of cloud computing for clusters. Its basic strategy is to build independent virtual machines, flexibly respond to the sudden increase or decrease in resource demands of cloud users, and improve the efficiency of platform resource utilization. Simulation is a process of continuously extracting dependencies, and its guest virtual machines will be managed and operated by virtual machine administrators (Hypervisors), providing personalized and diverse computing environments.

(2) *Spatial database.* Spatial databases optimize the storage and querying of spatial objects, including points, lines, and surfaces, based on traditional relational databases. A typical relational database typically only includes various numbers and characters, while a spatial database adds spatial data types and adds database functionality to handle these types of data [6-7]. OGC (OpenGeospatial Consortium) has developed the Simple Features Specification for geospatial data and corresponding standards to standardize the functionality of spatial databases in data processing. Due to the fact that the indexes of relational databases are not optimized for spatial queries, spatial databases need to develop their own spatial indexes to improve the efficiency of spatial database operations. Universal spatial databases typically support spatial operations such as spatial metrics, spatial functional functions, spatial predicates, constructors, and more. However, currently many NoSQL databases such as MongoDB and CouchDB, although they support spatial data types, do not fully support the aforementioned spatial operation functions.

Table 2.1: Comparison of NoSQL Database Types

Type	Product	Characteristic
Key value	Redis	Simple and easy to use, with direct values
Column Family	Bigtable, HBase	Flexible mode, allowing for arbitrary addition or deletion of column families
Document	MongoDB, CouchDB	Arbitrary pattern query, nested documents
Chart	Neo4, GraphLab	Suitable for complex data structures

(3) *NoSQL database*. Popular Web 2.0 applications typically have hundreds of millions of users, and these applications generate massive amounts of user data in a short period of time (ranging from a few months to a year), causing server loads to grow exponentially, resulting in extremely high demands for data storage scalability. In order to meet the requirements of data storage and management for such applications, web data must be partitioned and stored on thousands of processors. These new storage systems aim to provide distributed data storage, good horizontal scalability, and high-performance read and write operations. At the same time, traditional relational databases severely lack horizontal scalability, which limits the performance of single machines to handle data of such scale. NoSQL is a thriving non relational database technology in this context, which is usually classified into various types such as key values, column families, documents, and graphs based on different data models, as shown in Table 2.1.

The NoSQL database adopts a looser consistency model to provide a simple, lightweight, and efficient storage and retrieval mechanism to support better scalability and availability than traditional relational databases. The key technology for NoSQL to handle massive data is that it pre-set partitioning functions for the data. NoSQL typically automatically divides data into relatively small table units (MB level), stored on multiple different physical servers, and user programs access the servers where these table blocks are located through indexes or metadata. All updates generated by the user program will be aggregated to the main server, and then the updates to the data will be propagated to each replica server storing the table block through synchronization services.

Due to the high cost of the two-stage commit protocol used in traditional distributed databases, it may also fail during commit, leading to cluster congestion. NoSQL databases follow the CAP theorem and mostly provide a BASE concurrent transaction model that is looser than ACID, achieving basic availability, flexible state, and final consistency of the database in specific application areas. Therefore, NoSQL basically includes the following characteristics:

The ability to expand horizontally. NoSQL database can dynamically expand to multiple servers as business and data grow; Copy distribution capability. NoSQL database easily, effectively, and accurately replicates and propagates data to child node servers; Simple query interface and protocol. Compared to the complete SQL syntax of relational databases, NoSQL databases only provide lower level data query interfaces; Data storage can effectively utilize distributed indexes and memory; Ability to dynamically add new attributes to data records.

(4) *Figure Database*. In many fields including semantic analysis, geographic information systems, image processing, social networks, and biochemical informatics, graphs are a natural and suitable data model for domain data features. Semantic web information can be viewed as a collection of graphs that represent entities and explicit relationships; The transportation network in GIS is a typical graph; In chemical informatics, graphs can be used to represent the atoms and chemical bonds of compounds. The data in these fields are highly complex and large-scale, and existing data models, query languages, and database systems are difficult to support the modeling, querying, and management of these data. A graph database is a database system that represents and stores data in a graph structure with vertices, edges, and attributes, providing adjacency operations without indexing. The existing graph databases mainly use PropertyGraph as the data model, and some graph databases support HyperGraphModel. A property graph is a multigraph in which both vertices and edges have attributes stored in key value pairs, and all edges of the property graph are directed and asymmetric, as shown in Figure 2.1. A hypergraph is a superset of a graph, whose edges can be associated with any number

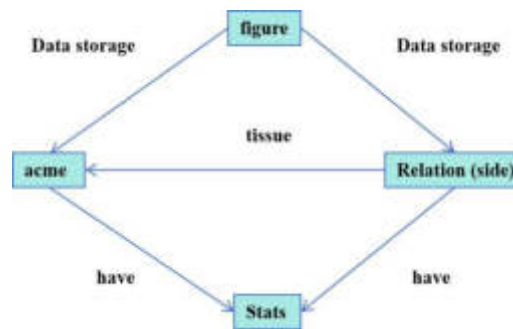


Fig. 2.1: Attribute Graph Data Model

of vertices.

Due to the fact that graph databases typically use attribute graphs as data models, native graph based data management systems have natural advantages in data management in the aforementioned fields. They can be used to store highly complex and large-scale non relational data in many fields, including semantic web, GIS, social network, bioinformatics, chemical informatics, and more. The author will map vector data that follows the OGC Simple Feature Specification (SFS) to an attribute graph model, where the objects and relationships of the vector data correspond to the vertices and edges of the graph data, respectively. Due to the fact that both the item points and edges in the attribute graph have attributes, there will be some storage redundancy. The spanning tree of the attribute graph model itself can serve as a natural database index, but it is not entirely applicable to spatial data. Spatial indexing can effectively improve the efficiency of spatial data queries, and in-depth research is needed on the theory of spatial indexing in order to select the appropriate spatial index.

(5) *Spatial index*. Spatial index is a spatial data structure that can be used in spatial databases to optimize spatial query indexing techniques by preserving the position, shape, and positional relationships of spatial entities. The components of spatial indexing include pointer objects pointing to spatial entities, bounding polygons of spatial entities, and object identifiers of spatial entities, which have very similar implementation principles. The spatial index will divide the spatial query area based on the principle of spatial segmentation by dividing the query dimension of the target. The spatial dimension will identify these partitioned subspaces with a tree structure and ensure index uniqueness through hash identification. The principle of spatial segmentation mainly includes two methods: Rule-based segmentation and object based segmentation. The former is based on the idea of computational geometry, while the latter is based on the independence of spatial objects. Rule based segmentation method, viewing a geographic plane as a multi-dimensional geometric body, segmentation is achieved through regular rectangles or irregular concave polygons. The integrity of a single spatial feature entity is ignored, and it will be divided into multiple parts of different units. However, this method of segmentation does not disrupt the logical consistency of spatial entities, but only changes the pointer object of the spatial index. The object segmentation method first determines the minimum bounding rectangle of the spatial entity, and then separates it according to the degree of entity independence, while ensuring the spatial and attribute consistency of the entity. However, this method has a high time complexity and produces a large number of spatial entities, resulting in significant storage and computational consumption. Therefore, it can be seen that spatial indexing, through preprocessing, can exclude objects unrelated to user query targets and quickly locate spatial entities that meet query syntax requirements.

2.2. Machine learning based segmented clustering method for cloud storage data.

(1) *Building a similarity matrix for cloud storage datasets*. Reasonably extract multiple small datasets from the cloud storage database, which should contain all natural clusters in the database. Divide the numerical data in the small-scale dataset into other types of data, extract them separately, and obtain independent datasets. Based on the data in each column, construct numerical and symbolic matrices, and combine them to obtain similarity matrices. Various types of data are mixed and stored in the database, and the dataset forms multiple natural clusters in the database. Extracting small-scale datasets from the database and performing segmented

clustering on the selected dataset can effectively reduce the computational error of numerical data clustering algorithms and simplify the algorithm steps. When extracting small-scale datasets, it is important to ensure the validity of the dataset and determine the number of selected natural clusters. Assuming that there are a total of n natural clusters in the selected dataset, in order to effectively simplify the complexity of mixed large-scale databases, a small-scale dataset sampling method is adopted to cluster the selected dataset, and a sample estimation method is designed to obtain the ideal sampling sample. Use S to represent the sampling sample, and let $\xi, \xi \in [0, 1]$ be the data extraction ratio of the dataset, due to the presence of n natural clusters in the dataset, the size of the extracted natural clusters is n . T represents the probability of extracting $\xi \times n$ data from the natural clusters in the dataset, and the resulting data samples are represented as:

$$S = \xi \times n + \frac{n}{n_i} \times \log \frac{1}{\tau} + \frac{n}{n_i} \sqrt{\log\left(\frac{1}{\tau}\right)^2 + 2\xi \times n \times \log \frac{1}{\tau}} \tag{2.1}$$

If the size and number of categories of the natural cluster are small or equal, it indicates that the size of the natural cluster is more than one layer, and the size of the natural cluster meets the requirements for extracting the dataset [8]. Set the extraction ratio of the dataset to ξ in such cases where the natural cluster size and number of categories are small, the extracted dataset size will also be smaller. Adopting equal scale sampling for mixed large-scale dataset A can enhance the convenience of dataset sampling in cloud storage databases and ensure the rationality of sampling. Set the small-scale dataset size as A_i , undergo m sampling, and meet the sampling control conditions as follows:

$$\begin{cases} A_i \cap A_j = \emptyset \\ m_i = m_j \end{cases} \tag{2.2}$$

Because the size of the sampled dataset is relatively small, when clustering the dataset, the aggregation level will quickly complete the clustering, which greatly improves the clustering speed. Moreover, due to the small size, the clustering accuracy is also improved. Further cluster the numerical data extracted from the dataset, strip out other types of data, and in order to simplify the algorithm steps, all other types of data are uniformly recorded as symbolic data. Extract numerical and symbolic data separately, construct two independent datasets, and construct the approximation matrix of the dataset as follows:

$$W_i = \frac{C_i}{\sum_{i=1}^n C_i} \tag{2.3}$$

In the formula, W represents an independent dataset.

Calculate the similarity between numerical and symbolic data separately, and construct a matrix of numerical data using a Gaussian function, assuming AA represents a numerical data matrix, and d represents the euclidean distance between data points, λ represents the characteristic parameter of the Gaussian function, then T_{ij} can be expressed as:

$$T_{ij} = \exp\left(-\frac{d}{2\lambda^2}\right), i, j = 1, 2, \dots, n \tag{2.4}$$

Symbolic data attributes can be set to:

$$T'(x_i, x_j) = \begin{cases} 0, & x_i \neq x_j \\ 1, & x_i = x_j \end{cases}, i, j = 1, 2, \dots, n \tag{2.5}$$

The numerical matrix $T_{i,j}$ and symbolic data $T'_{i,j}$ can be represented as follows:

$$T_{i,j} = \begin{bmatrix} 1, 0.331, 0.475, 0.358 \\ 0.331, 1, 0.331, 0.135 \\ 0.475, 0.331, 1, 0.216 \end{bmatrix} \tag{2.6}$$

$$T'_{i,j} = \begin{bmatrix} -1, 0, 0, 1, 0, 1 \\ 0, 1, 1, 0, 1, 0 \\ 0, 1, 1, 0, 1, 0 \\ 1, 0, 0, 1, 0, 1 \\ 0, 1, 1, 0, 1, 1 \end{bmatrix} \quad (2.7)$$

The similarity matrix constructed by combining numerical matrix $T_{i,j}$ with symbolic data $T'_{i,j}$ is:

$$T' = T_{i,j} + \sum_{i=1}^n P_i \times T'_{i,j}, j = 1, 2, \dots, n \quad (2.8)$$

In the formula, P represents the similarity weight.

(2) *Segmented clustering of cloud storage data based on machine learning.* Based on the similarity matrix provided in (1), a nonlinear kernel principal component algorithm is used to measure the similarity of data in the similarity matrix. A mixed Gaussian distribution probability density model is used, combined with the comparison of similarity measurement probabilities, to achieve segmented clustering of cloud storage data.

Based on the similarity matrix, a non-linear kernel principal component algorithm is used to obtain the matrix similarity measure, taking into account a set of variables with non-linear correlation in the cloud storage database $X_{i,j}(t), i = 1, 2, \dots, N, j = 1, 2, \dots, m, t = 1, 2, \dots, T$, N represents the data capacity in the cloud storage database, n represents the number of variables, and T represents the length of the time series. Assuming that the $N \times m$ sample data in cloud storage is represented as $X(t)(x_1(t), x_2(t), \dots, x_n(t))$, a nonlinear mapping function ω is used to project the sample data $X_i(t)$ from the input space R^N to the high-dimensional feature space F^N :

$$\omega : R^N \rightarrow F^N \quad (2.9)$$

The sample data projected onto the high-dimensional and high-dimensional feature space F^N is represented by $\omega(x_i(t), i = 1, 2, \dots, N)$, and the mapping process needs to satisfy the centralization condition of the feature space as follows:

$$\sum_{i=1}^N \omega(x_i(t)) = 0 \quad (2.10)$$

After projection, the covariance matrix of the sample data that meets the centralization condition is set as:

$$C = \frac{1}{n} \sum_{i=1}^n \omega(x_i(t)) \omega^T(x_i(t)) \quad (2.11)$$

In the formula, C represents the covariance matrix.

Assuming that the eigenvalues of C satisfy $0 \leq \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_N$, the corresponding feature data is $V(\nu_1, \nu_2, \dots, \nu_N)$. Due to the fact that the non-zero feature data V and the projection data $\omega(X(t))$ belong to the high-dimensional feature space F^N , there exists a set of coefficients $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_N]^T$ that enable V to be represented by a linear combination of $\omega(X(t))$ as:

$$V = \sum_{i=1}^N \alpha_i \omega(x_i(t)) \quad (2.12)$$

Set a kernel matrix K for $N \times N$, the specific process is as follows:

$$K_{i,j} = k(x_i, x_j) = \omega^T(x_i(t)) \omega(x_j(t)) \quad (2.13)$$

The vector form of its sum matrix is:

$$K\alpha = N\lambda\alpha \quad (2.14)$$

In the above equation, K is the kernel function matrix of $N \times N$; N is the data characteristic value of K ; $\alpha_1, \alpha_2, \dots, \alpha_N$ is the eigenvector corresponding to the eigenvalues of each data. The projection $P(x_j(t))$ of sample data $x_j(t)$ in the direction of feature vector $V_r (r = 1, 2, \dots, k)$:

$$P(x_j(t)) = V_r^T \omega(x_i(t)) \quad (2.15)$$

In the formula, the vector $(x_i(t))$ represents the j -th sample data.

After setting the non-linear kernel function $k(x_i, x_j)$, it is possible to perform non-linear projection of data in cloud storage databases based on mapping rules, reduce data dimensions, and extract data feature information according to non-linear related indicators. The weight of the extracted features is:

$$p_i = \frac{\lambda_i}{\sum_{i=1}^k \lambda_i} \quad (2.16)$$

In the formula, p_i represents the weight of the eigenvalues [9].

Based on similarity measurement, a mixed Gaussian distribution model is used to segment and cluster cloud storage data. The mixed Gaussian distribution model is a weighted linear combination of a finite number of multivariate Gaussian distributions. The probability density function $g(x)$ of the mixed Gaussian distribution is calculated as:

$$g(x) = \sum_{i=1}^k \alpha_i f_i(x; \mu_i) \quad (2.17)$$

In the formula, K represents the number of multivariate Gaussian distributions; $f_i(x; \mu_i)$, α_i is the probability density function and weight of the i -th multivariate Gaussian distribution. In the parameter estimation of the model, due to the large number of parameters to be estimated, the moment estimation method is obviously difficult to implement. Therefore, consider maximum likelihood estimation and construct the corresponding likelihood function L ;

$$L = \prod_{j=1}^N g(x_j; \theta) \quad (2.18)$$

According to the above equation, when there is a large amount of data, maximum likelihood estimation is more difficult.

Assuming the number of clusters k is a known exogenous variable, initialize all parameters of the mixed Gaussian model, assuming $\theta_0 = (\alpha_{i0}, \mu_{i0})$, among them, the mean vectors and covariance matrices of K multivariate Gaussian distributions can be obtained through other statistical algorithms, and the weights are initially set to $\frac{1}{k}$.

In the first step of updating weights, for any sample x_j , the probability $\omega_{j1}(k)$ of the k -th class is used, and the $\omega_{j1}(k) \times x_j$ part of its value is treated as generated by the k -th multivariate gaussian model, the k -th multivariate gaussian model produces $\omega_{j1}(k) \times x_j, (j = 1, 2, \dots, N)$, with a total of N data points, and the parameters belonging to this multivariate gaussian distribution replace the initial parameters. After the first iteration, the parameters of the k -th single Gaussian model are:

$$N_{k1} = \sum_{j=1}^N \omega_{j1}(k) \quad (2.19)$$

After a complete EM calculation, the updated value $\theta_1 = (\alpha_{i1}, \mu_{i1})$ of all cloud storage segment data can be obtained. Using θ_1 as the cloud storage segmentation data for the mixed Gaussian model, a second EM iteration can be performed. Given a sufficiently small threshold, after multiple iterations, when $|\ln(L)^{[n-1]} - \ln(L)^{[n]}| < \text{threshold}$, the EM loop iteration can be exited. Obtain convergent model parameters.

After achieving global convergence of model parameters, clustering of cloud storage segmented data samples x_j can be achieved by comparing the probability value $\omega_j(k), \omega_j(k)$ of any sample x_j belonging to different categories.

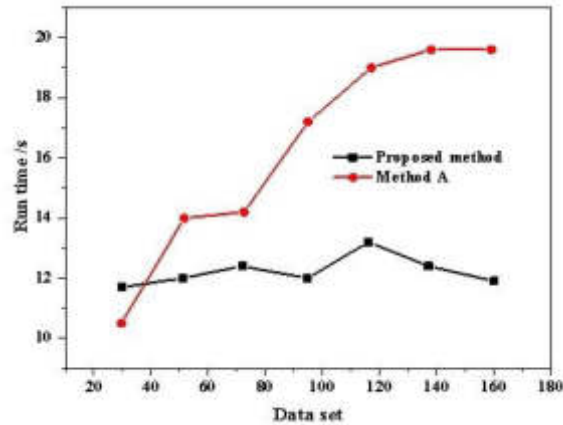


Fig. 3.1: Comparison of runtime (s) of different methods

Table 3.1: Clustering variation of different methods (%)

SJ/s	W7	W8	ST
2	59	66	43
4	63	69	47
6	57	70	39
8	62	73	32
10	60	83	30

3. Results and Analysis. In order to verify the comprehensive effectiveness of the proposed machine learning based cloud storage data segmentation clustering method, a simulation is required, and the simulation environment is: Intel (R) Core (TM) 2i5-3210M, CPU main frequency 2.3GHz, 2GB of memory, operating system Windows 7, software environment Anaconda3 (64 bit), simulation program using Python language. Cluster the proposed method with method A for cloud storage data, and compare the operational efficiency (%) of the two methods. The comparison results are shown in Figure 3.1. (Method A: Segmented clustering method for cloud storage data based on trend function space).

As shown in Figure 3.1, as the number of samples increases, the running time of the two methods also changes. Overall, the proposed method has a shorter running time than method A [10]; Specific analysis shows that when the number of data in the dataset is 30, the clustering time of method A is 11 seconds, which is 1 second shorter than the proposed method. However, when the number of data in the dataset is higher than 30, the running time of method A increases linearly, which is 8 seconds longer than the proposed method. The main reason is that method A needs to calculate the integrated spatial distance between a large number of data points, the proposed method effectively reduces the calculation of integrated spatial distance and shortens the clustering running time by dividing the data in the cloud storage database into multiple small datasets. Compare the clustering changes (%) at different times on the same dataset using the proposed clustering method, A clustering method, and B clustering method, respectively. The comparison results are shown in Table 3.1: In Table 3.1, SJ represents different time points in seconds, represented by (s); JL represents the degree of clustering change, in%; ST represents the method proposed; W7 represents method A; W8 represents method B. (Method A: Segmented clustering method for cloud storage data based on trend function space, Method B: Segmented clustering method for cloud storage data based on principal component analysis).

Compare the experimental data in Table 3.1 in the form of experimental graphs, as shown in Figure 3.2.

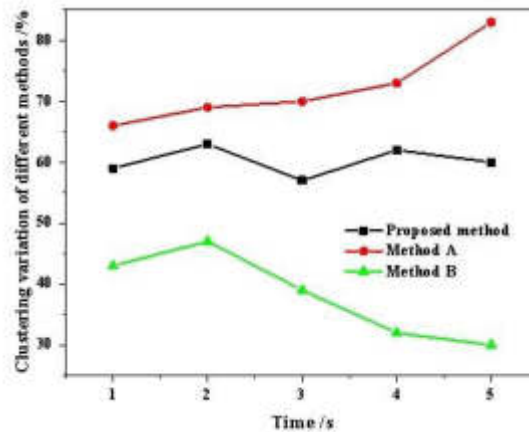


Fig. 3.2: Clustering variation of different methods

From Table 3.1 and Figure 3.2, it can be seen that the clustering degree of the three methods is not the same at different times. At 2 seconds, the clustering degree of the B clustering method is 65%, which is 23% higher than the proposed clustering method and 7% higher than the A clustering method; As the clustering time increases, the clustering degree of A and B clustering methods also increases. At 10 seconds, the clustering degree of A method is 30% higher than that of the proposed method, and the clustering degree of B method is 53% higher than that of the proposed method; Throughout the clustering process, only the proposed method showed a decreasing degree of clustering variation with increasing clustering time, decreasing from 42% to 29%. Overall, the proposed method effectively improved the smoothness of clustering results.

4. Conclusion. In response to the problems of low efficiency and insufficiently smooth clustering results of traditional cloud storage data segmentation clustering methods, the author proposes a machine learning based cloud storage data segmentation clustering method. The experimental results indicate that, the method proposed by the author has significantly improved the clustering variation compared to traditional methods. The proposed method reduces the clustering variation from 42% to 29%, effectively improving the smoothness of the clustering results. There are still many problems in using the proposed cloud storage data segmentation and clustering method in practical applications, and the network environment is becoming increasingly complex, with the types of data becoming more diverse, after the emergence of complex and diverse data, the proposed method cannot effectively cluster all types of data. In the face of a wide variety of data, it is necessary to improve the ability of segmented clustering.

5. Acknowledgement.

Source: University Level Scientific Research Projects;

Name: Research on the Application of Artificial Intelligence Algorithm in UAV Detection System; Number XJKY202209.

REFERENCES

- [1] Sabur, A., Chowdhary, A., Huang, D., & Alshamrani, A. (2022). Toward scalable graph-based security analysis for cloud networks. *Computer Networks*, 206(468), 108795-.
- [2] Zhao, H. (2023). Research on the recognition and localization of *momordica grosvenori* based on binocular vision and a convolutional neural network. *2023 IEEE International Conference on Control, Electronics and Computer Technology 223(ICCECT)*, 404-408.
- [3] Wang, P., Nie, S., Wang, J., Wang, C., Xi, X., & Du, M. (2022). Segmentation of the communication tower and its accessory

- equipment based on geometrical shape context from 3d point cloud. *International Journal of Digital Earth*, 15(1), 1547-1566.
- [4] Anna E. Sikorska-Senoner. (2022). Clustering model responses in the frequency space for improved simulation-based flood risk studies: the role of a cluster number. *Journal of Flood Risk Management*, 15(1), n/a-n/a.
 - [5] Saurabh, & Dhanaraj, R. K. (2023). Enhance qos with fog computing based on sigmoid nn clustering and entropy-based scheduling. *Multimedia Tools and Applications*, 83(1), 305-326.
 - [6] Gao, W., & Zhang, L. (2022). Semantic segmentation of substation site cloud based on seg-pointnet. *J. Adv. Comput. Intell. Intell. Informatics*, 26(546), 1004-1012.
 - [7] Zhang, Y., Gao, X., Bai, Y., Wang, M., & Tian, Q. (2022). Multi-condition identification of thermal process data based on mixed constraints semi-supervised clustering. *SN Applied Sciences*, 4(7), 1-19.
 - [8] Li, J., Xing, Y., & Zhang, D. (2022). Planning method and principles of the cloud energy storage applied in the power grid based on charging and discharging load model for distributed energy storage devices. *Processes*, 10(2), 194-.
 - [9] Chen, G., Bai, B., Mao, Z., & Dai, J. (2022). Real-time road object segmentation using improved light-weight convolutional neural network based on 3d lidar point cloud. *International Journal of Ad Hoc and Ubiquitous Computing*, 39(3), 113-.
 - [10] Singh, A., & Kumar, M. (2023). Bayesian fuzzy clustering and deep cnn-based automatic video summarization. *Multimedia Tools and Applications*, 83(1), 963-1000.

Edited by: Zhigao Zheng

Special issue on: Graph Powered Big Aerospace Data Processing

Received: Jan 16, 2024

Accepted: Mar 5, 2024



PLATEAU ALTITUDE DISASTER PREVENTION AND REDUCTION PLATFORM BASED ON BEIDOU SYSTEM

TONGWEI ZHANG*, YONGQI LIU† AND LINGYING LI‡

Abstract. This project is based on the Beidou satellite navigation and positioning technology developed by China. The integrated analysis method is used to obtain the monitoring data of high-precision power cables in high-altitude areas with high-precision millimeters to realize the whole process of monitoring various disasters. An early warning system for power line swing and tower bottom geological hazards is established. The track data of cable swing is obtained by using single-frequency dynamic data post-processing (PPK) based on synchronous power grid monitoring. The applicability and trajectory accuracy of the scheme are evaluated by combining theory with practical engineering. This provides essential technical support for the mechanism and prevention of cable sloshing. The experimental simulation shows that the monitoring efficiency of the system is high.

Key words: Beidou satellite; Transmission lines; Cable sloshing; Foundation strengthening system

1. Introduction. In recent years, the frequent occurrence of power line swaying events has seriously affected the regular operation of the power system. In January 2020, due to continuous low-temperature rainfall, the power systems of five provinces and cities, Anhui, Hubei, Hunan, and Jiangxi, appeared to have a wide range of swings. Among them, more than 20 high-voltage transmission lines of 220 kV and above appeared to have different degrees of oscillation. The oscillation period is more than 40 hours, and the oscillation amplitude can reach 5 meters. Transmission line oscillation is a kind of mechanical vibration caused by wind force, ice-covered wing lift and wire tension under particular weather conditions. It is determined by three factors: ice covering, wind excitation, line structure and parameters, and the whole process includes starting, maintenance, attenuation and other stages.

The power network monitoring and management system must monitor the transmission lines in the swing area online and measure the swing influence factors in real-time. Currently, there are two kinds of real-time monitoring methods for power line swaying at home and abroad: video image and accelerometer. The existing monitoring method based on video images has some problems, such as limited transmission of transmission lines, low resolution, low accuracy of swing detection, and inability to realize real-time monitoring. The monitoring mode of the acceleration sensor is to measure a single conductor's acceleration and then obtain its oscillation's velocity and displacement through mathematical integration. However, this method can only reflect the swing amplitude locally, and it is easy to produce cumulative errors and cannot reflect the direction characteristics of the swing. Literature [1] uses a novel FBG sensing mode to design a method for measuring the shaking of objects. Literature [2] established a new idea of regional sloshing monitoring and early forecasting based on measured data by combining sloshing sampling with weather data. Literature [3] proposes using independent differential GPS technology to monitor and forecast the swing of transmission lines in real time and dynamics. But the above methods have some defects. Therefore, a method of wide-area transmission line shaking monitoring based on the Beidou foundation strengthening system and Beidou high precision positioning technology is proposed in this paper. A new method of swing detection for a "wide area transmission line" is studied [4], and at the same time, related equipment development is studied. This lays a foundation for the application research of "swing" monitoring and "active early warning" in China's power system.

2. Design a power disaster prevention Beidou system in high-altitude areas.

*Diqing Power Supply Bureau, Yunnan Power Grid Co. Ltd., Diqing, Yunnan, 674400, China (Corresponding author, liyongqi108@126.com)

†Diqing Power Supply Bureau, Yunnan Power Grid Co. Ltd., Diqing, Yunnan, 674400, China

‡Diqing Power Supply Bureau, Yunnan Power Grid Co. Ltd., Diqing, Yunnan, 674400, China

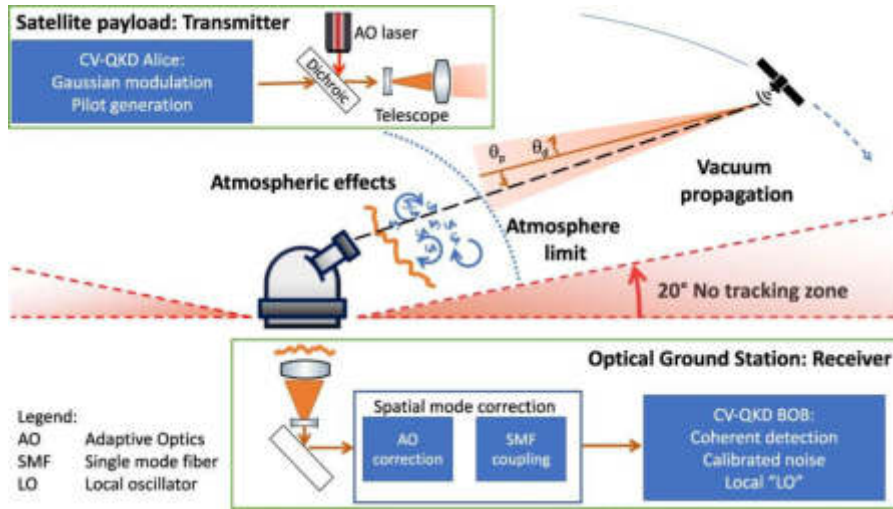


Fig. 2.1: Schematic diagram of satellite single point positioning.

2.1. Basic Theory of Satellite GPS. In practical application, the Beidou system (BDS) is complicated to achieve accurate distance measurement because of the uncertainty of inter-satellite and receiver clock differences. Therefore, in measuring the distance between the satellite and the receiver, the Beidou system includes multiple navigation satellites, user receiver equipment and ground measurement and control network [5]. They are both spatial position reference points and communication relay stations. In engineering practice, the receiver synchronously receives signals from multiple satellites and obtains more accurate navigation and position through corresponding algorithms [6]. The receiver then obtains a pseudo-range formula containing the unknowns. Since the unknowns include the user’s space coordinates (X, Y, Z) and the receiver’s time difference, the pseudo-range positioning is essentially to obtain the receiver’s geodetic coordinate system and receiver clock difference by using the pseudo-range measurements of multiple satellites [7]. Figure 2.1 shows the schematic diagram of satellite single point positioning (the picture is quoted in Feasibility of satellite-to-ground continuous-variable quantum key distribution).

Suppose the user’s position is (X, Y, Z) , the satellite coordinates on BDS are (X_r, Y_r, Z_x) , and the distance between the user and the i satellite is S_{ri} .

$$S_{ri} = \sqrt{(X_r - X)^2 + (Y_r - Y)^2 + (Z_r - Z)^2}$$

It is determined that there is an error between the pseudo distance of GPS and the actual distance. The pseudo-distance is represented by Q_{ri} ,

$$Q_{ri} = S_{ri} + v\Delta t_d$$

v is the speed of light. Δt_d represents the time error of the receiver. There are four unknowns X, Y, Z and Δt_d in equation (2.1) and (2.2). The receiver must receive at least four satellite position signals to find the uncertainty [8]. When more than four satellites are received, the least square method is used to determine the coordinates and clock differences of the receiver. The equation composed of $n(n \geq 4)$ satellite position formulas is shown in formula (2.3):

$$\begin{cases} Q_{r1} = \sqrt{(X_{r1} - X)^2 + (Y_{r1} - Y)^2 + (Z_{r1} - Z)^2} + \sigma W \\ Q_{r2} = \sqrt{(X_{r2} - X)^2 + (Y_{r2} - Y)^2 + (Z_{r2} - Z)^2} + \sigma W \\ Q_m = \sqrt{(X_m - X)^2 + (Y_m - Y)^2 + (Z_{rn} - Z)^2} + \sigma W \end{cases}$$

There is $\sigma W = v\Delta t_d$ in the formula. Because the prerequisite for calculating by the least square method is that the equation is linear, while the satellite navigation is non-linear, the term (2.3) should be linearized [9]. The above expression is malleable using the Taylor series.

$$Q_{rn} = Q_{rn} + \frac{X_m - X}{S_{rn}} \Delta X + \frac{Y_m - Y}{S_m} \Delta Y + \frac{Z_m - Z}{S_m} \Delta Z - \Delta \sigma W$$

Q_{rn} is the approximate distance between the receiver and the second n satellite. $(\hat{X}, \hat{Y}, \hat{Z})$ are the approximate azimuth coordinates of the receiver for each solution. $\Delta X, \Delta Y, \Delta Z, \Delta \sigma W$ is the displacement concerning the four unknowns. Convert all (2.3) to (2.4) and represent it with a matrix table

$$Q_r = \varphi H$$

$Q_r = [Q_{r1} \quad Q_{r2} \quad \cdots \quad Q_m]^T$ $H = [\Delta X \quad \Delta Y \quad \Delta Z \quad \Delta \sigma W]^T$, φ is the coefficient matrix, specifically

$$\varphi = \begin{bmatrix} \frac{X_{r1}-X}{S_{r1}} & \frac{Y_{r1}-Y}{S_{r1}} & \frac{Z_{r1}-Z}{S_{r1}} & -1 \\ \frac{X_{r2}-X}{S_{r2}} & \frac{Y_{r2}-Y}{S_{r2}} & \frac{Z_{r2}-Z}{S_{r2}} & -1 \\ \vdots & \vdots & \vdots & \vdots \\ \frac{X_{rn}-X}{S_m} & \frac{Y_m-Y}{S_{rn}} & \frac{Z_{rn}-Z}{S_m} & -1 \end{bmatrix}$$

Due to the interference of atmospheric transmission time delay and multipath effect, the position accuracy of pseudo distance positioning will decrease when the least square algorithm fixes the measured value.

2.2. System Design. This project relies on the national "Beidou" navigation and ground combined navigation technology to carry out high-precision differential positioning of multiple wobbler points in an extensive range (Figure 2.2 quoted in Journal of Revisualization and Spatial Analysis, 2020, 4:1-12.). Grasp the real-time information on wire swaying waveform, swaying amplitude, swaying ellipse Angle, and swaying frequency. The position information of each wobbly satellite is transmitted in real-time [10]. At the same time, combined with the field environment, it is dynamically adjusted in real-time to achieve the purpose of real-time detection and alarm of the security risk of the power grid. The obtained results will be pushed to the monitoring and early warning application center and related monitoring terminals.

The oscillating monitoring equipment based on Beidou is adopted. At the same time, it is integrated with the sensing equipment installed in the equipment to realize the online monitoring of multiple signal sources [11]. Through the public network communication module in the equipment, the obtained information is uploaded to the solution business platform of the "Beidou" ground enhancement system. Using the PPK solution, each observation point's X, Y and Z coordinates in vertical coordinates are dynamically post-processed. The corresponding road map is given, combined with the historical data of each monitoring point [12]. The heave curve's response equations of X, Y and Z coordinates are established, and spectral analysis is carried out.

2.3. System Communication. With the support of the Beidou application terminal, SMS transceiver, mobile communication terminal, emergency communication server and emergency communication service platform software, each component realizes the sending and receiving of Beidou short messages and mobile short messages under the support of the system network of Beidou Communication and mobile communication. The system architecture is shown in Figure 2.3 (image cited in Satellite Navigation, 2020, 1:1-23).

Beidou handheld/vehicle terminal equipment is used for people/vehicles to move in and out of specific areas. The "Beidou" system is used as the command terminal to provide support for communication support in emergencies. The RDSS link is used to realize short message communication between Beidou mobile phone/car users and Beidou command clients. Short message sending and receiving is a communication device with mobile phone short message sending and receiving, which can send and receive short messages to the mobile communication terminal through the communication base station. In an emergency, the mobile phone short message sending and receiving device is configured in the emergency communication support center [13]. The system is mainly used to manage the user terminal and short message-receiving device in the Beidou system and realize the system's data transmission and address transmission.

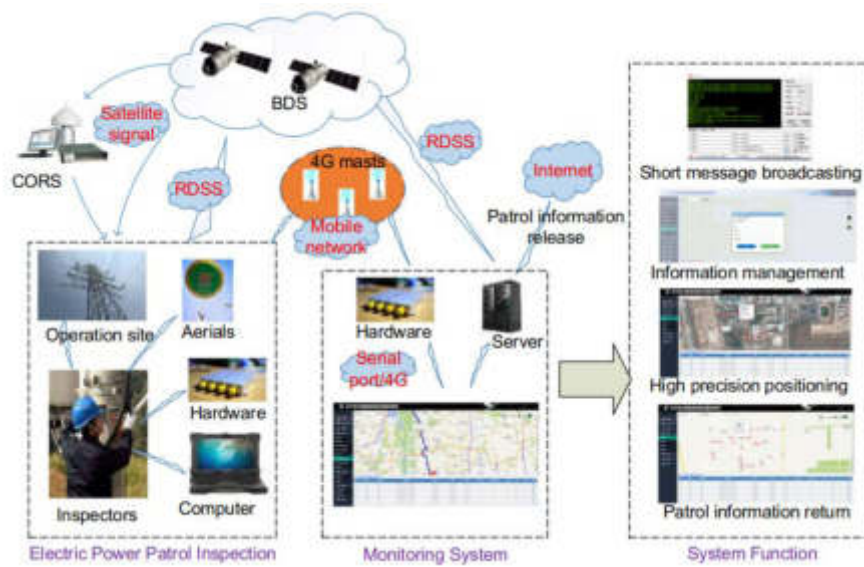


Fig. 2.2: Structure of BDS swing monitoring system.

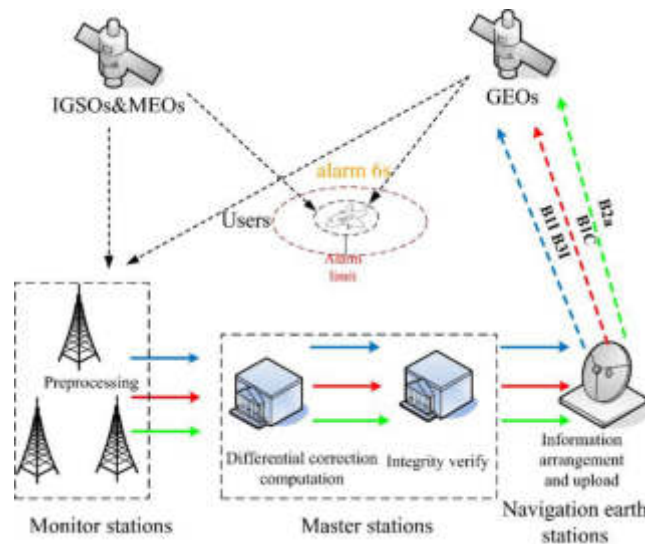


Fig. 2.3: Structure of the Beidou communication system.

3. Data processing methods. When the wire is affected by ice or wind force, the wire will have a sizeable self-excited vibration at low frequency. A small swing from the equilibrium position will eventually produce an increasingly sizeable elliptical orbit dominated by lateral and longitudinal orbits. When the wire does not shake the phenomenon, the system will use the position, acceleration and other related data statistics to determine whether the system is balanced. The "Beidou" system and the sensing system are used to collect the position information of each measuring point in real-time and convert it into discrete values. By comparing the change values of Beidou high-precision coordinates in continuous periods, the change difference values of each observation point on the corresponding axis are obtained [14]. Plot the sway trajectories of the monitoring points. The sloshing rate of the wire can be obtained by differentiating the displacement data by the first

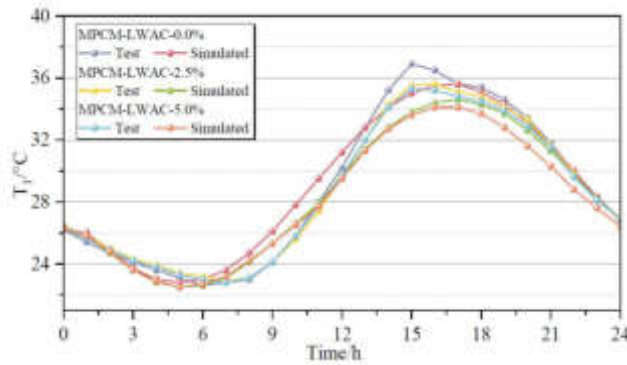


Fig. 4.1: ENU coordinate change curve of simulation test.

difference method. Further, the pendulum acceleration of each measuring point is obtained. Set the horizontal movement direction of the wire to the x direction and the vertical movement direction to the y direction [15]. Then, based on the change of the coordinates before and after each direction, the displacement along the direction is obtained, and the calculation method of the velocity and acceleration in each direction is obtained:

$$C_{x(t)} = \frac{s(X_t - X_0)}{dt}$$

$$\gamma_{x(t)} = \frac{\frac{s(X_t - X_0)}{dt} - C_{x_0}}{dt}$$

$$C_{y(t)} = \frac{s(Y_t - Y_0)}{dt}$$

$$\gamma_{y(t)} = \frac{\frac{s(Y_t - Y_0)}{dt} - C_{y_0}}{dt}$$

where X_0, Y_0 is the coordinates of each monitoring point in the x direction and Y direction at the previous time. X_t, Y_t is the coordinate of the measuring point in the X and Y coordinates after time t . Where C_{x_0}, C_{y_0} is the velocity of the time before the monitoring point. Under the initial conditions, its speed is zero. $\gamma_{x(t)}, \gamma_{y(t)}$ is the value of the acceleration of the measuring point in both directions X and Y. Then, based on the high-precision differential positioning of the Beidou satellite, this paper obtains parameters such as the trajectory, amplitude, vertical amplitude and horizontal frequency of sway through the simulation analysis of sway.

4. Application testing. A 500 kV power line tower simulation test system is established. The research results show that using Beidou navigation technology for transmission line sway detection has high accuracy and practicability. Install the swing monitoring device on the top of the tower. The observation data of Beidou have been accepted by the shaking detector in the field test site, and the condition is good. In this project, the initial observation data of the wobbly monitoring equipment were established on board, 1-second sampling was adopted, and Beidou B1 was used as the receiver frequency. In this way, the measured data at every moment are transmitted in real-time to the fixed platform of the Beidou ground-enhanced positioning system. Through the data exchange with the ground data server, the "Beidou" observation data near the reference station of the swing monitoring network are verified. The Beidou satellite navigation positioning and solving platform monitors the real-time ground sway. After dynamic post-processing under the PPK model, each measurement's X, Y and Z coordinates are obtained. The velocity and acceleration curves of each measuring point are obtained after testing. Figures 4.1, 4.2, and 4.3 show the values of the coordinates, velocity, and acceleration differences over different periods.

A total of 357 pieces were tested. After systematic processing, the definite number of dynamic post-processing was obtained for 355 pieces, and the ratio of the definite number of dynamic post-processing to the

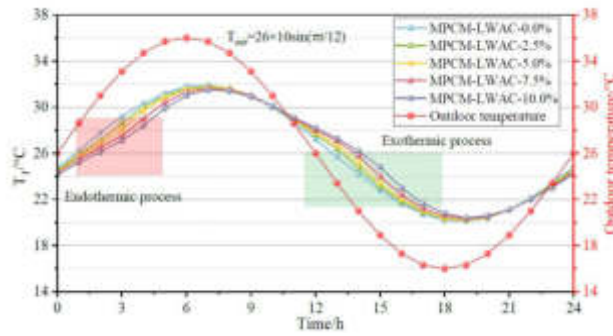


Fig. 4.2: Speed change curve of simulation test.

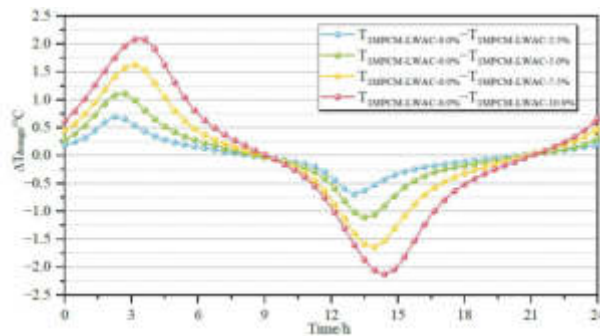


Fig. 4.3: Acceleration change rule in simulation test.

number of outputs was 99.44%. The internal consistency accuracy error (Table 4.1) is calculated for data with coordinate change values, velocity and acceleration $\sigma_i = \sqrt{\frac{\sum \Delta_i^2}{n-1}}$. Where σ_i is the internal consistency of each component, and Δ_i is the calculated value and average difference of each component. n is the actual observed quantity of each component.

At the same time, the system can also quickly obtain high-precision motion speed, acceleration and other parameters to meet the needs of transmission line sloshing detection. The calendar has a total of 2511 elements. By analyzing the system, 2511 definite solution values were obtained after dynamic post-processing, among which the ratio of the number of definite solutions after dynamic post-processing to the total number of iterations reached 99.47%. Several monitoring points were sampled. The MATLAB software was used to analyze the horizontal and vertical displacement data of each measuring point measured at each time, and the corresponding dance traces were fitted. The fitting results show that the vibration amplitude is 0.66 meters.

5. Conclusion. This paper introduces a new idea of real-time monitoring of power line swing by using Beidou satellite positioning technology. This method can improve swing detection and monitoring of transmission lines. The research results of this project will lay the foundation for the realization of real-time monitoring and intelligent analysis of big data of the whole network under space reference. The threshold of the power cable is adjusted individually using the existing history data. The research scheme proposed in this project can adapt to the safety monitoring requirements of power systems for transmission lines. This system has a high reference significance.

REFERENCES

[1] Xu, Y., Wang, W., Yang, X., Deng, K., & He, Z. (2023). Design and research of power system Beidou timing and positioning

Table 4.1: *Results of accuracy calculation following.*

Precision value	Accurate in the X direction	Match accuracy in the Y direction	Accurate in the Z direction
Coordinate accuracy /m	0.0093	0.0078	0.0188
Velocity accuracy /($m \cdot s^{-1}$)	0.0049	0.0053	0.0079
Acceleration accuracy /($m \cdot s^{-2}$)	0.0066	0.0056	0.0072

module based on K-means clustering and gross error processing. IET Cyber-Physical Systems: Theory & Applications, 8(1), 34-42.

- [2] Wu, Z., Zhang, Y., Liu, L., & Yue, M. (2020). TESLA-based authentication for BeiDou civil navigation message. China Communications, 17(11), 194-218.
- [3] Wu, D., Liu, S., Sun, H., & Zhang, L. (2022). Short-message communication Lossy data compression algorithm for BeiDou-3 satellite information transmission. Multimedia Tools and Applications, 81(9), 12833-12855.
- [4] Harahap, C. N. M., & Rakhmadi, R. (2023). The Development of China's Beidou Navigation Satellite System (BDS) Technology to Counter the United States': Perkembangan Teknologi Beidou Navigation Satellite System (BDS) Tiongkok untuk Melawan Global Positioning System (GPS) Amerika Serikat. Jurnal Terekam Jejak, 1(1), 1-11.
- [5] Zhang, J., Luo, X., Fu, X., Wang, X., Guo, C., & Bai, Y. (2020). Experimental study on the influence of satellite spoofing on power timing synchronization. International Journal of Network Security, 22(6), 954-960.
- [6] Liu, S., Wu, D., Sun, H., & Zhang, L. (2022). A novel BeiDou satellite transmission framework with missing package imputation applied to smart ships. IEEE Sensors Journal, 22(13), 13162-13176.
- [7] Yang, Y., Mao, Y., & Sun, B. (2020). Basic performance and future developments of BeiDou global navigation satellite system. Satellite Navigation, 1(1), 1-8.
- [8] Ao, Z., Li, F., Ma, Q., & He, G. (2020). Voice and Position Simultaneous Communication System Based on Beidou Navigation Constellation. Xibe Gongye Daxue Xuebao/Journal of Northwestern Polytechnical University, 38(5), 1010-1017.
- [9] Zhao, R., Lu, J., Guo, W., Zheng, W., & Li, S. (2022). BP-LMS-based BDS-3 power system positioning method. Global Energy Interconnection, 5(6), 666-674.
- [10] Aoyama, R. (2022). China's dichotomous BeiDou strategy: led by the party for national deployment, driven by the market for global reach. Journal of Contemporary East Asia Studies, 11(2), 282-299.
- [11] Liu, P., Zhang, S., Zhou, Z., Lv, L., Huang, L., & Liu, J. (2023). Multiple satellite and ground clock sources-based high-precision time synchronization and lossless switching for distribution power system. IET Communications, 17(18), 2041-2052.
- [12] Yang, T., Liu, Y., & Li, W. (2022). Attack and defence methods in cyber-physical power system. IET Energy Systems Integration, 4(2), 159-170.
- [13] Dou, Z., Zhang, C., Wang, W., Wang, D., Zhang, Q., Cai, Y., & Fan, R. (2022). Review on key technologies and typical applications of multi-station integrated energy systems. Global Energy Interconnection, 5(3), 309-327.
- [14] Liu, S., Guo, X., Lai, J., & Yang, J. (2022). Distributed Timekeeping in BeiDou Inter-satellite Link Network. IEEE Communications Letters, 26(12), 3014-3018.
- [15] Yang, Z., Lin, L., Xu, L., Huo, Y., & He, M. (2023). Research on milk powder production and sales management system based on Beidou positioning. Academic Journal of Management and Social Sciences, 2(3), 1-5.

Edited by: Zhigao Zheng

Special issue on: Graph Powered Big Aerospace Data Processing

Received: Jan 16, 2024

Accepted: Mar 5, 2024



STATE MONITORING AND ANOMALY DETECTION ALGORITHMS FOR ELECTRICITY METERS BASED ON IOT TECHNOLOGY

CHUNGUANG WANG*, TIANFU HUANG†, ZHIWU WU‡, YING ZHANG§ AND HANBIN HUANG¶

Abstract. In response to the practical application of the electricity consumption information collection system in the online monitoring business of measuring equipment, the author introduces a method for analyzing the abnormal flying away of electricity meters based on the IoT technology LOF local anomaly detection algorithm. This method can effectively determine whether the abnormal energy representation value belongs to accidental or trend anomalies by calculating the abnormal factor of the energy representation value. After excluding the influence of accidental data, perform a secondary judgment on the abnormal flight of the energy meter. The experimental results show that when calculating the LOF factor of the electricity meter, it can be found that the LOF curve data range is mainly concentrated in the range of 0.8 to 1.3, and there is no significant change in the LOF factor near the mutation point. This proves that this method can effectively improve the accuracy of anomaly detection, avoid misjudgment of faults, and improve the efficiency of on-site fault handling.

Key words: Smart energy meters, Electricity information collection system, LOF, Outliers

1. Introduction. The electricity settlement of power enterprises is mainly completed through energy metering devices. Electricity metering management is an important link in the production and operation management of power enterprises and the safe operation of the power grid. Its technology and management level not only affect the development and corporate image of power enterprises, but also affect the accuracy and fairness of trade settlement, involving the interests of a large number of power customers [1]. Therefore, in order to ensure the accuracy and reliability of the energy meter, it is necessary to reduce the error of the energy meter and make direct payment. fair and reasonable. In order to ensure the accuracy and reliability of the electricity meter, it is necessary to ensure the accuracy and reliability of the standard electricity measurement equipment first [2]. Electrical measuring instruments are distributed only to various state electrical testing centers and power plant states, and currently, the measurements of measuring instruments in various modes are monitored in the laboratory by directly connected computers by the measuring staff in electronic measuring instruments. Since the power meter cannot be monitored remotely, if any abnormality or problem is detected during the measurement, the calibration staff can only remind the management to go to the laboratory for maintenance immediately, which will delay the detection. problems, long-term solutions, low performance, high cost, and difficulty adapting to the growing demand for energy metering. For example, a three-phase energy meter standard device in the metering center of a certain power supply bureau has a total of 16 calibration meter positions [3]. From the surface analysis, it can be judged that the crimping of positions 3, 4, and 11 is damaged and cannot properly calibrate three-phase energy meters, the remaining meter positions are working normally, and the calibration data also meets the requirements of the electric energy meter calibration regulations. There is a special case where the calibration data of the 16 meter positions is better than that of other normally working meter positions. It took three working days for maintenance personnel to check and find that the compression joint of the 16 meter positions was severely damaged, resulting in abnormal data. Due to the large workload of calibration, calibration personnel can only pay attention to whether the calibration data is out of tolerance and whether the voltage connection of the energy meter is normal for most of the time during the calibration

*Marketing Service Center of State Grid Fujian Electric Power Co., Ltd., Fuzhou, Fujian, 350013, China (Corresponding author, lunwen_zy@163.com)

†Marketing Service Center of State Grid Fujian Electric Power Co., Ltd., Fuzhou, Fujian, 350013, China

‡Marketing Service Center of State Grid Fujian Electric Power Co., Ltd., Fuzhou, Fujian, 350013, China

§Marketing Service Center of State Grid Fujian Electric Power Co., Ltd., Fuzhou, Fujian, 350013, China

¶Marketing Service Center of State Grid Fujian Electric Power Co., Ltd., Fuzhou, Fujian, 350013, China

process. Such problems require a long period of time to be discovered, which can have a long-term impact on the accuracy of energy measurement.

As the core infrastructure equipment of ubiquitous power Internet of Things, the demand for smart energy meters is constantly increasing. In terms of electricity metering, the accuracy of smart energy meters is the core element of fairness in electricity trade; In the field of big data research, accurate and qualified smart energy meters are one of the foundations for building ubiquitous power Internet of Things [4]. As of the end of 2020, more than 300 million electricity meters have been put into operation in China, and this number will continue to grow in the future. Traditional manual calibration methods cannot meet the growing demand for calibration, and the transformation from manual calibration to automated systems is imperative. As a result, intelligent electricity meter automated calibration assembly lines have emerged. During the long-term operation of automated calibration assembly lines, frequent connection of smart energy meters to meter positions can cause deformation of the mechanical crimping terminals at the meter positions; Long term live operation can accelerate the oxidation rate of the surface material of mechanical crimping terminals, leading to terminal corrosion [5]. The deformation and corrosion of the mechanical crimping process of the meter will directly affect the reliability of the error test results, thereby affecting the calibration quality of the smart energy meter. At present, the provincial metrology centers under the State Grid Corporation of China generally adopt the method of regular verification to inspect and repair the meter positions on the calibration assembly line. This method cannot detect meter position faults in a timely manner and relies on manual troubleshooting. Its reliability is insufficient and labor costs are high. Therefore, achieving online anomaly detection of calibration assembly line meter positions is of great significance [6].

The abnormal occurrence of the meter flying away is relatively accidental, and generally it will not occur repeatedly on the same meter; Due to external factors such as communication interference, there is a certain amount of data noise in the electricity consumption data. Therefore, using the general analysis method of threshold judgment can cause a large number of misjudgments of runaway anomalies, which affects the discovery and handling of actual meter runaway faults. The LOF algorithm is a classic density based time series anomaly diagnosis algorithm. The author used this algorithm to establish an intelligent diagnostic analysis method for the flying away of electric energy meters, which can effectively remove the influence of outliers, improve the accuracy and timeliness of detecting and judging the flying away anomalies of electric energy meters.

2. Construction of an accuracy analysis platform for electricity metering.

2.1. System Design. The author uses Hadoop distributed technology to build a massive data storage and high-performance parallel computing cluster that covers various power supply bureaus in the province to cope with large-scale data parallel processing. The author also uses Nigera load balancing technology and Redis as a caching component to improve throughput and system availability, meeting the needs of high concurrency requests [7]. Due to the need to process real-time collection of massive power grid measurement data, the system adopts a distributed architecture based on Hadoop and applies the HDFS distributed file system, from the monitoring terminal of measurement standard devices, laboratory calibration control system, measurement automation system, marketing system, on-site inspection business system, semi-structured or unstructured data, then it is stored in the non relational database HbaSe, and some of the data is cleaned and analyzed offline through the Hive data warehouse. The overall architecture of the accuracy analysis platform for electricity metering is shown in Figure 2.1.

2.2. System Software Architecture. In order to meet the requirements of high concurrency, high reliability, and system scalability for data collection, the system is planned to be constructed in MVC mode. The data collection part adopts Nginx+Tomcat cluster to achieve high concurrency and high reliability of web servers, and the backend uses queues+process pools to achieve multi-channel processing of various protocols and concurrent links; The system adopts a B/S architecture based on J2EE for construction. The core technology framework of the platform adopts JAVA as the development language, based on mainstream open-source J2EE frameworks, including Struts, Spring, Hibernate, JQuery, JBOSSOA, JBPM, Druid, and other frameworks. Capable of supporting multiple heterogeneous databases and compatible with mainstream web containers. The technical structure is divided into four layers: basic environment, DAO layer, logic layer, and presentation layer [8].

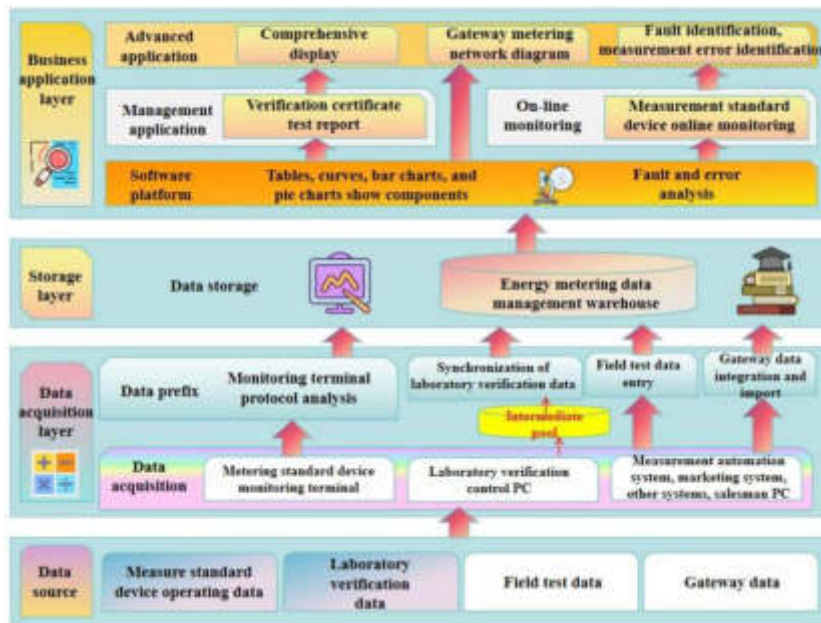


Fig. 2.1: Overall architecture of the accuracy analysis platform for electric energy metering

2.3. System Hardware Architecture. Establish a dedicated network for the metering center on the internal power network and install encrypted communication. VPN gateway achieves remote access by encrypting data packets and converting the target address of the data packets. Implement VPN proxy servers through various methods such as servers, hardware, and software. By using virtual private networks, the measurement center network is made more humanized, software oriented, and intelligent, providing secure, controllable, and flexible resource scheduling capabilities to meet the dynamic restructuring needs of power metering data communication. Addressing network security and monitoring, MAC address tracking, and security vulnerabilities in virtual machine management programs [9].

3. LOF anomaly detection algorithm.

3.1. Abnormal point detection. Outlier detection (also known as outlier detection) is an important part of data mining and refers to the process of identifying objects whose behavior is significantly different from what is expected. These items are called outliers. Vulnerability detection is critical to many applications, including healthcare, public safety, fault detection, image processing, sensor/video networks, and testing see. Misdiagnosis can be divided into controlled and uncontrolled [10]. If the analyst can find registered original and unusual items, they can be used to create an abnormality detection sample before using control methods to detect them. In some applications, unsupervised learning techniques are used when there are no objects labeled as "normal" or "abnormal". Undoubtedly detection assumes that the product contains some kind of impurity.

3.2. Method Comparison. Due to the different electricity consumption patterns of different users, it is not possible to use a unified method for state labeling of energy representation values, and it is not possible to provide a learning set of supervised methods. Therefore, unsupervised anomaly detection algorithms are more suitable for the analysis of energy representation value anomalies. In unsupervised methods, statistical, distance, and density based methods are currently the main methods for anomaly detection [11].

The statistical anomaly detection method mainly analyzes the dispersion of data, analyzes the distribution of data, extracts the variation indicators of data, commonly used include standard deviation, interquartile spacing, etc., and extracts outliers through the variation indicators. This type of algorithm requires prior knowledge of the distribution characteristics of the data, as well as parameter determination of mutation

indicators. The algorithm has poor universality and is mainly used in scientific research fields.

The distance based anomaly detection method considers the neighborhood of an object with a given radius. If it is an anomaly, there are not enough other points in its neighborhood. Compared to statistical algorithms, it does not require users to have any domain knowledge and is more intuitive in concept. However, due to the algorithm detecting outliers from a global perspective by calculating the distance between objects, the detection effect is not good when there are multiple distributions or subsets with different densities in the dataset [12]. Density based detection methods mainly examine the density of the object and its neighbors. If its density is much lower than its neighbors, it is considered an outlier. In this type of algorithm, each point will calculate an outlier degree, overcoming detection errors caused by mixing different density subsets, and the detection accuracy is relatively high. The author mainly adopts density based anomaly detection methods.

3.3. LOF algorithm. The key of density detection is to compare the density around an object with the surrounding density. There is a significant difference between the density of non-outlier objects and the surrounding neighborhood, and there is a significant difference between them and their surroundings. In this paper, a Local Outlier Factor (LOF) is proposed based on the combination of density and anomaly detection. The algorithm is defined as follows:

For a given set of objects D , the objects contained in D are denoted as o . The k -distance of object o is denoted as $dist_k(o)$, which is the distance $dist(o, p)$ between o and another object $p \in D$, such that:

*At least k objects $o \in D - \{o\}$, such that $dist(o, o) \geq dist(o, p)$

*At least $k-1$ objects $o \in D - \{o\}$, such that $dist(o, o) < dist(o, p)$

$dist_k(o)$ is the distance between o and its k -th nearest neighbor. Therefore, the k -distance neighborhood of o includes all objects whose distance to o is not greater than $dist_k(o)$, denoted as:

$$N_o = \{o' | dist(o, o') \leq dist_k(o)\} \quad (3.1)$$

If the average distance between objects in $N_k(o)$ and o is used as the local density measure of o , a problem arises. When a very close neighbor o' is encountered in o , causing $dist(o, o')$ to be very small, the statistical fluctuation of the distance measure will be unexpectedly high. In order to solve this problem, a smoothing effect can be added to convert it into the following reachable distance [13].

$$reach-dist(o, o') = MAX\{k - distance(o), d(o, o')\} \quad (3.2)$$

The local reachable density of object o is the reciprocal of the average reachable distance between object o and its MinPts neighborhood.

$$lrd_{MinPts}(o) = \frac{1}{\frac{\sum_{o' \in N_{MinPts}(o)} reach-dist(o, o')}{|N_{MinPts}(o)|}} \quad (3.3)$$

The local anomaly factor of object o is defined as:

$$LOF_{MinPts}(o) = \frac{\sum_{o' \in N_{MinPts}(o)} lrd_{MinPts}(o')}{|N_{MinPts}(o)|} \quad (3.4)$$

The degree of anomaly in object o can be evaluated through its local anomaly factors; The anomaly factor is directly proportional to the degree of anomaly of the object. If the factor value is larger, the likelihood of anomaly increases; If the factor value is smaller, the likelihood of anomalies decreases, and the LOF factor level of normal values is generally within 1.

4. Example Analysis.

4.1. Calculation steps. In the analysis of electricity meter runaway, we can use the LOF algorithm to filter the original data, effectively identify abnormal points in electricity meter readings, and eliminate the impact of abnormal points in electricity meter readings on the judgment of electricity meter runaway. At the same time, due to the introduction of table readings before and after abnormal points in the judgment, it is necessary to improve the threshold judgment. The author mainly judges based on the daily maximum electricity consumption and daily average electricity consumption [14].

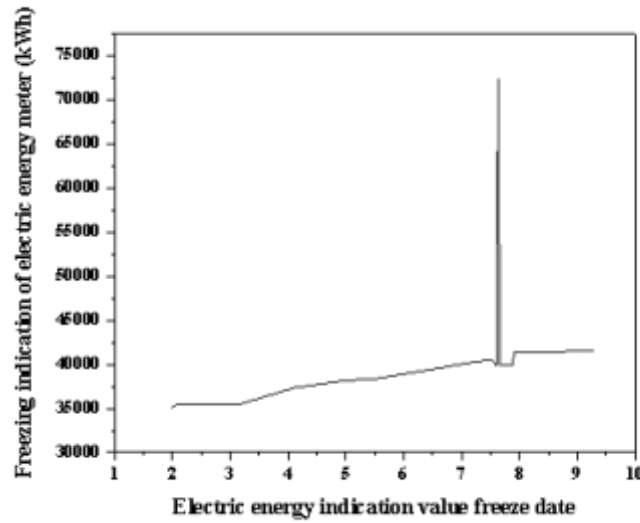


Fig. 4.1: Daily freezing curve of electric energy representation (accidental mutation)

Table 4.1: Daily freezing readings and corresponding LOF factors (accidental mutation) for the three days before & after the electricity meter trip

Date	July 18th	July 20th	July 21st	July 22nd	July 23rd	July 25th	July 26th
Indicating value	40647	39908	39908	72300	39908	39908	39908
LOF	1.501447	0.984361	0.984361	6041.262	0.984361	0.984361	0.984361

- (1) Preliminary judgment of abnormal flight according to the flight formula of the electric energy meter;
- (2) Extract suspected daily data of the flight, generally including at least 7 days before and after the flight date;
- (3) Extract daily data from one of the meters, organize the data and sort the dates;
- (4) Calculate the LOF factor and count the number of abnormal factors;
- (5) Remove the daily data corresponding to LOF abnormal factors;
- (6) Calculate daily electricity consumption;
- (7) Calculate the improvement N value.

$$N = \frac{\text{Maximum daily electricity consumption}}{\text{Average daily electricity consumption}} \tag{4.1}$$

4.2. Calculation results. Perform LOF factor calculation on the energy meter in Figure 4.1, and the calculation result is shown in Figure 4.2. It can be observed that the position of the mutation point on the electric energy representation curve is the same as that on the LOF curve [15,16]. We extracted the daily data of the mutation point and the three days before and after, as shown in Table 4.1. It can be found that the LOF of the mutation point is 6041.262, while the LOF of the other points is basically around 1. The outlier can be filtered out. After elimination, calculate the daily electricity consumption and calculate the N value for flight judgment.

The LOF factor calculation was performed on the electricity meter, and the results are shown in Figure 4.3. It can be found that the LOF curve data range is mainly concentrated in the range of 0.8 to 1.3, and there is no significant change in the LOF factor near the mutation point, as shown in Table 4.2 [17,18,19,20].

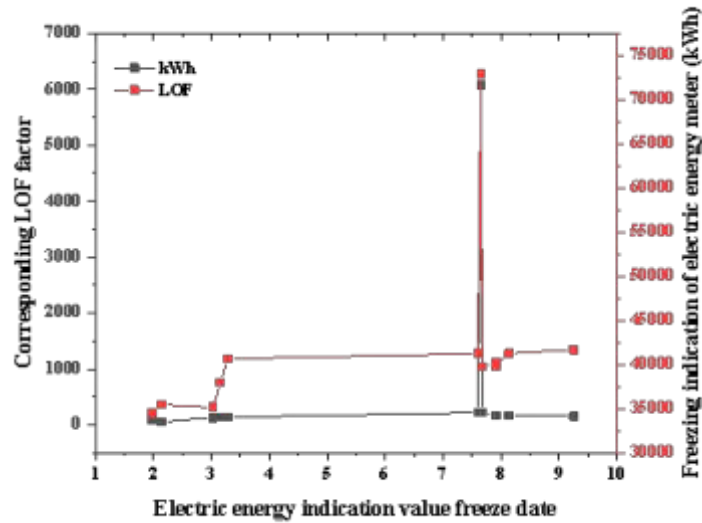


Fig. 4.2: Comparison of Daily Freezing Curve and LOF Curve for Energy Representation (Accidental Sudden Change)

Table 4.2: Daily freezing indication and corresponding LOF factor (indication flying away) for the three days before and after the jump of the electric energy meter

date	June 9th	June 10th	June 11th	June 12th	June 13th	June 14th	June 15th
indicating value	1913	1919	119614	119619	119627	119632	119638
LOF	1.0772	1.1393	1.1476	1.1077	1.0639	1.0495	1.0411

5. Conclusion. Through the above algorithms, it is possible to achieve a secondary judgment of the abnormal flight of the electric energy meter detected by the monitoring of the electricity consumption information collection system, which greatly improves the quality and efficiency of the judgment of the electric energy meter flight. It can timely detect the abnormal flight of the electric energy meter, effectively avoid fault misjudgment, and improve the efficiency of on-site fault handling; This method can be further extended to the analysis of other topics related to electricity consumption information collection, improving the accuracy of diagnosis and analysis of the operating conditions of measuring equipment, and further supporting the marketing business decision-making and implementation of power supply enterprises.

Meanwhile, the above is only a practical method for handling data outliers. With the deepening application of electricity information collection system data, how to handle anomalies in data will be the first problem that needs to be solved in future business applications. The rapid development of technologies such as statistical analysis, data mining, and machine learning has laid a solid theoretical foundation for solving the above problems. We need to further increase the introduction, understanding, and practice of data science in the power business.

REFERENCES

- [1] Douiba, M., Benkirane, S., Guezzaz, A., & Azrou, M. (2022). Anomaly detection model based on gradient boosting and decision tree for iot environments security. *Journal of Reliable Intelligent Environments*, 85(7),1-12.

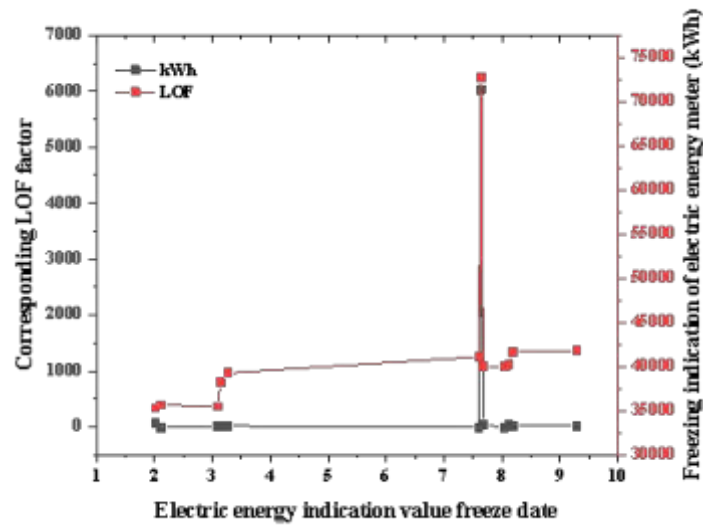


Fig. 4.3: Daily Freezing Curve of Electric Energy Representation Value (Display Value Flying)

- [2] Shu, X., Zhang, S., Li, Y., & Chen, M. (2022). An anomaly detection method based on random convolutional kernel and isolation forest for equipment state monitoring. *Maintenance and Reliability*, 11(2), 72.
- [3] Petrillo, A., Murino, T., Piccirillo, G., Santini, S., & Caiazzo, B. (2023). An iot-based and cloud-assisted ai-driven monitoring platform for smart manufacturing: design architecture and experimental validation. *Journal of Manufacturing Technology Management*, 34(4), 507-534.
- [4] Lee, K. H., Lee, J. E., Seo, J. Y., & Kim, T. H. (2023). Health monitoring and anomaly detection of motor-driven equipment based on the tft model and vibration signal decomposition. *The Journal of Korean Institute of Information Technology*, 16(7), 203.
- [5] Jang, Y. M. (2022). Real-time energy data acquisition, anomaly detection, and monitoring system: implementation of a secured, robust, and integrated global iiot infrastructure with edge and cloud ai. *Sensors*, 46(2), 1864-1881.
- [6] Tang, M., Chen, W., & Yang, W. (2022). Anomaly detection of industrial state quantity time-series data based on correlation and long short-term memory. *Connection Science*, 34(1), 2048-2065.
- [7] Kumar, Y. R. S., & Champa, H. (2022). Anomaly detection framework for efficient sensing in healthcare iot systems. *2022 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, 65(7), 81-85.
- [8] Nayak, J., Naik, B., Dash, P. B., Vimal, S., & Kadry, S. (2022). Hybrid bayesian optimization hypertuned catboost approach for malicious access and anomaly detection in iot nomalyframework. *Sustainable computing: Informatics and systems*, 46(4), 4362-4379.
- [9] Hu, K., Cai, Y., Cai, Z., Li, X., Cen, B., & Chen, Y. (2022). Fault location method based on structure-preserving state estimation for distribution networks. *IET generation, transmission & distribution*, 74(15), 16.
- [10] Omonov, A., Fitriyah, A., Kato, T., & Kawabata, Y. (2022). Improving monitoring technologies for the salinity of irrigated lands based on gis/rs and satellite imagery for the conditions of the syrdarya region, uzbekistan. *Journal of Arid Land Studies*, 32(3), 96-96.
- [11] Zhou, Y., Liu, C., Yu, X., Liu, B., & Quan, Y. (2022). Tool wear mechanism, monitoring and remaining useful life (rul) technology based on big data: a review. *SN Applied Sciences*, 96(8), 4.
- [12] Kong, J., Jiang, W., Tian, Q., Jiang, M., & Liu, T. (2023). Anomaly detection based on joint spatio-temporal learning for building electricity consumption. *Applied energy*, 46(3), 2418-2437.
- [13] Ayalew, L. G., Mattihalli, C., & Asmare, F. M. (2022). Wirelessly controlled plant health monitoring and medicate system based on iot technology. *Communications in Computer and Information Science*.
- [14] Liu, W., Lei, S., Peng, L., Feng, J., Pan, S., & Gao, M. (2022). Active anomaly detection technology based on ensemble learning. *Springer, Singapore*, 20(1), 25.
- [15] Jérémy Renaud, Karam, R., Salomon, M., & Couturier, R. (2023). Deep learning and gradient boosting for urban environmental noise monitoring in smart cities. *Expert Systems with Applications*, 218(32), 119568.
- [16] Wang, N., Zhang, G., Ren, L., Pang, W., & Li, Y. (2022). Correction to: novel monitoring method for belt wear state based on machine vision and image processing under grinding parameter variation. *The International Journal of Advanced Manufacturing Technology*, 122(1), 103-104.
- [17] Martins, I., Resende, J. S., Sousa, P. R., Silva, S., Antunes, L., & Gama, J. (2022). Host-based ids: a review and open issues

- of an anomaly detection system in iot. Future generations computer systems: FGCS,96(133), 133.
- [18] Mukherjee, I., Sahu, N. K., & Sahana, S. K. (2023). Simulation and modeling for anomaly detection in iot network using machine learning. International journal of wireless information networks, 47(3), 557-564.
- [19] Hou, Y., Fu, Y., Guo, J., Xu, J., Liu, R., & Xiang, X. (2022). Hybrid intrusion detection model based on a designed autoencoder. Journal of Ambient Intelligence and Humanized Computing, 14(8), 10799-10809.
- [20] Mukherjee, I., Sahu, N. K., & Sahana, S. K. (2022). Simulation and modeling for anomaly detection in iot network using machine learning. International Journal of Wireless Information Networks, 30(2), 173-189.

Edited by: Zhigao Zheng

Special issue on: Graph Powered Big Aerospace Data Processing

Received: Jan 17, 2024

Accepted: Mar 5, 2024



OPTIMIZATION OF LOGISTICS DISTRIBUTION NETWORK BASED ON ANT COLONY OPTIMIZATION NEURAL NETWORK ALGORITHM

JING YANG*

Abstract. In order to improve the timeliness of logistics distribution, based on the theory of road network smoothness and reliability, the author conducted a study on the optimization of urban logistics distribution and transportation networks based on smoothness and reliability. The concept of logistics distribution and transportation network smoothness and reliability was proposed, and a logistics distribution and transportation network optimization model was established. The solving process of ant colony algorithm was given, and finally, a comparative analysis of a case was conducted. The results showed that: With a 6% increase in total delivery distance, the reliability of the delivery network has increased by 30%. This indicates that when using the model built by the author for distribution network optimization, effective optimization of network smoothness and reliability can be achieved, while only increasing the distance by a small amount. The optimal reliability of a smooth distribution network means that the probability of delivery delays is minimized, which is the most powerful guarantee for the effective accessibility of delivery. Verified the practicality of the constructed model. The proposed logistics distribution network optimization model has practical significance in guiding decision-making for optimizing urban logistics distribution transportation networks and reducing uncertainty in the process of urban logistics distribution.

Key words: Logistics distribution, Network optimization, Smooth reliability, Ant colony

1. Introduction. In the process of globalization, the level of modernization of logistics has become an important indicator of a country's modernization and strength. It is the fundamental driving force for the sustained development of the national economy at a high starting point and an effective way for enterprises to seize competitive advantages in fierce market competition [1]. China is increasingly valuing the development and application of logistics technology to reduce logistics costs and bring considerable economic benefits to enterprises and society. The logistics distribution path optimization problem is a typical combinatorial optimization problem that involves multiple disciplines. Many practical problems can be classified into this category, with broad application prospects. Therefore, it has always been a hot research topic in the fields of operations research and combinatorial optimization. In recent years, significant achievements have been made in the research of vehicle optimization scheduling problems, which have been widely used in various aspects of production and life, such as newspaper or cargo delivery, taxi scheduling, and parcel express delivery.

The logistics network is specifically composed of various logistics nodes and routes connecting each point. Among them, nodes include distribution centers, warehouses, etc., while routes refer to the routes and routes for logistics distribution operations in accordance with regulations. The purpose of optimizing the distribution network path is to optimize the movement (transportation) path of goods during the spatial transfer process from the production area to the consumption area. The purpose of studying it is to find the most reasonable vehicle transportation path, overcome spatial barriers, and achieve cost savings. Due to the fact that solving the distribution path problem is an NP hard problem with high complexity. The traditional methods for solving such problems include precise algorithms and heuristic algorithms. The optimal solution to a problem can be obtained using precise algorithms, but its solving time often increases exponentially with the size of the problem. When there are a large number of nodes to be processed, it takes a considerable amount of time to solve, so its practical application scope is limited. Heuristic methods usually simplify problems into several small problems or modules based on their characteristics, and solve them in a more intuitive way. Therefore, most researchers are committed to the development and improvement of heuristic algorithms. Research on previous research has found that although heuristic algorithms can obtain satisfactory solutions to VRP problems, it is difficult to

* School of Finance and Economics, Hainan Vocational University of Science and Technology, Haikou, Hainan, 570000, China (yjkytg2023@126.com)

obtain the optimal solution within a reasonable computational time using existing algorithms as the problem size increases.

At present, logistics distribution logistics in a general sense refers to the establishment of distribution centers according to the frequency and size of goods order the needs of different customers, as well as the use of 7R (for products, quality, time, time, location, conditions, customers, and costs). Because of the complexity of the transportation system in distribution, especially in urban distribution system, not only the multi-points transportation system, multi-products, and the traffic network, but the distribution of the transportation content in the transportation service area is also incompatible. Therefore, reasonable and efficient road design to reduce the number of vehicles and transportation costs has become an important and practical problem. The goal of logistics distribution center will be how to make good use of the vehicles, determine the cheapest driving way, to deliver the products to the customers in the shortest possible time. Specifically, it means using multiple vehicles from the distribution center to deliver goods to multiple demand points (users), with a fixed location and demand volume for each demand point, a fixed load capacity for each vehicle, and a reasonable arrangement of vehicle routes to achieve predetermined goals (such as shortest distance, least cost, least time, and least number of vehicles used). The interesting situation during this process is that the total number of points needed on each delivery path is within the carriage; The length of each road shall not exceed the maximum distance that a car can deliver at one time; All requests must and only be sent by one vehicle. This is a special topic of vehicle problem detection (VRP).

According to statistics, the annual logistics costs in the US industry reach \$400 billion. As long as it is reduced by 10%, it can save 40 billion US dollars in a year, so the logistics industry is vividly likened to "a gold mine worth 40 billion US dollars yet to be mined." In 2004, the demand for logistics in society continued to grow rapidly, with a total logistics volume of 3.84 billion yuan, a year-on-year increase of 29.9%, an increase of 2.9 percentage points compared to the same period last year. The logistics industry showed a rapid development trend. However, as a rapidly developing industry, the logistics industry has become a bottleneck that must be effectively managed in the process of national economic development. The proportion of logistics costs to GDP continues to be high, far higher than that of developed countries. At present, the logistics cost in developed countries accounts for about 10% of GDP, while in moderately developed countries (such as South Korea) it accounts for about 16%, while in China it is as high as around 20%. Transportation is the key to logistics decision-making, and in general, transportation costs account for a higher proportion than other logistics costs, except for procurement costs. Therefore, optimizing transportation logistics and reducing transportation costs is one of the effective ways to enhance the core competitiveness of logistics enterprises and reduce logistics costs.

Meanwhile, with the accelerated development of urban economy in recent years, various problems in cities have become increasingly prominent. Transportation and distribution have not only become important factors affecting the cost of goods, but also the level of urban development. The problems caused by distribution transportation, such as traffic congestion, traffic accidents, and exhaust pollution from transportation vehicles, have become major problems that hinder urban economic development and affect the normal life of residents. The traditional delivery method of goods involves retailers receiving goods directly from factories, forming a spider web cross transportation route, resulting in chaotic transportation and low efficiency; With the rapid increase of private vehicles in cities, especially in large cities, although cities have built three-dimensional transportation networks on the ground, above ground, and underground, they still cannot meet the transportation needs of population and vehicles, and urban traffic is becoming increasingly congested; In addition, due to the severe reduction of petroleum energy, fuel prices have risen rapidly, posing challenges to improving vehicle operating efficiency and reducing transportation costs. Therefore, as the volume of logistics and distribution business in cities gradually increases, if the previous logistics methods are still used to organize distribution, a series of problems will arise, such as a decline in service quality and inability to meet customer requirements; The emergence of a large number of unreasonable distribution plans will make it difficult to control logistics costs; Unreasonable vehicle path planning can increase the number of trips and routes for logistics delivery vehicles, leading to an increase in urban transportation burden. In order to solve the above problems, the urban logistics distribution network should be optimized to achieve goals such as on-time delivery, lowest total cost, and shortest total driving path. The advantage of distribution is that it eliminates cross transportation,

saves transportation costs, increases the loading rate of transportation vehicles, reduces empty driving rates, timely delivery, improves service quality, and eliminates the need for users to place orders everywhere, while also reducing consumer inventory. With the development of urban transportation and economy, the demand in the market is gradually diversifying, and in such an environment, a new distribution center - a distribution center - has emerged. In this logistics distribution and transportation method, the distribution center organizes delivery according to the high-frequency and small batch ordering requirements of different customers, generating activities such as physical distribution, commercial flow, information flow, and capital flow. Physical distribution refers to the process of delivering products to customers during the production and sales process. In this process, transportation costs often account for the majority of the total transportation and sales costs. Therefore, in recent years, many enterprises have been committed to reducing transportation costs to improve their core competitiveness. In terms of transportation costs, the cost of moving goods in the sales channel is the highest. Therefore, it is of great practical significance to study how to allocate vehicles and generate distribution routes based on the determined amount of goods, in order to reduce the total transportation cost.

With the rapid development of the economy and the continuous acceleration of urbanization, urban spatial structure, transportation layout, and infrastructure construction are constantly undergoing changes [2]. However, it is not optimistic that the problems that may arise during the urbanization process are increasingly exposed, among which urban traffic congestion is particularly prominent, which directly affects the efficiency of logistics distribution based on urban road networks. Therefore, in order to meet market needs, the logistics distribution and transportation network system should be optimized in a timely and reasonable manner [3,4]. From the perspective of traffic flow, the smoothness of logistics distribution networks is closely related to the random changes in road network traffic status. Therefore, while improving the operational efficiency of logistics distribution systems themselves, we must consider the impact of road network traffic status on logistics distribution networks. In reality, there are many random factors that affect the effectiveness of road network traffic, such as natural disasters, traffic accidents, or frequent traffic congestion. The reliability index is an important indicator for measuring the performance of road networks under the influence of random factors. The existing research results on road network traffic reliability can be divided into two categories: One focuses on the physical stability of the road network, only studying its reliability from the perspective of network topology structure, and basically not considering traffic flow; Another type focuses on the comprehensive performance of network functions, rather than just the evaluation of the road network structure itself [5]. In practical situations, the road network often maintains connectivity, but due to random changes in traffic flow, it may not always be smooth, which may cause logistics delivery vehicles to be unable to arrive at customer points in a timely manner. Based on the characteristics of urban logistics delivery itself, the author analyzes from the perspective of smoothness and reliability, and makes reasonable choices for distribution routes to improve logistics delivery efficiency and achieve the goal of optimizing the logistics distribution transportation network [6].

2. Methods.

2.1. Logistics Network Optimization Issues.

(1) *Logistics Network Definition.* From the micro perspective of enterprises, the definition of logistics network can be summarized as: the circulation channels through which goods move from the place of supply to the place of sale. By abstracting the social logistics system, a network consisting of logistics nodes and transportation routes has been formed, as shown in Figure 2.1 [7]. The transportation routes in the network represent the movement routes of goods between different inventory storage points, while logistics nodes refer to the storage or demand points of goods, such as warehouses, distribution centers, logistics centers, factories, suppliers, etc. There may be multiple transportation routes connected between any pair of logistics nodes, representing different transportation modes, routes, or products. Logistics nodes also represent temporary stopping points during the inventory flow process, such as logistics centers.

(2) *Mathematical model of logistics distribution.* The general distribution vehicle routing problem can be described as follows: Linear programming can effectively solve balanced transportation problems, as it can be summarized into the following linear programming model. Assuming the unit freight rate from place of origin i to place of sale j is c_{ij} , $i=1, 2, \dots, m$; $J=1, 2, \dots, n$. X is the optimal shipping volume from origin i to destination j ,

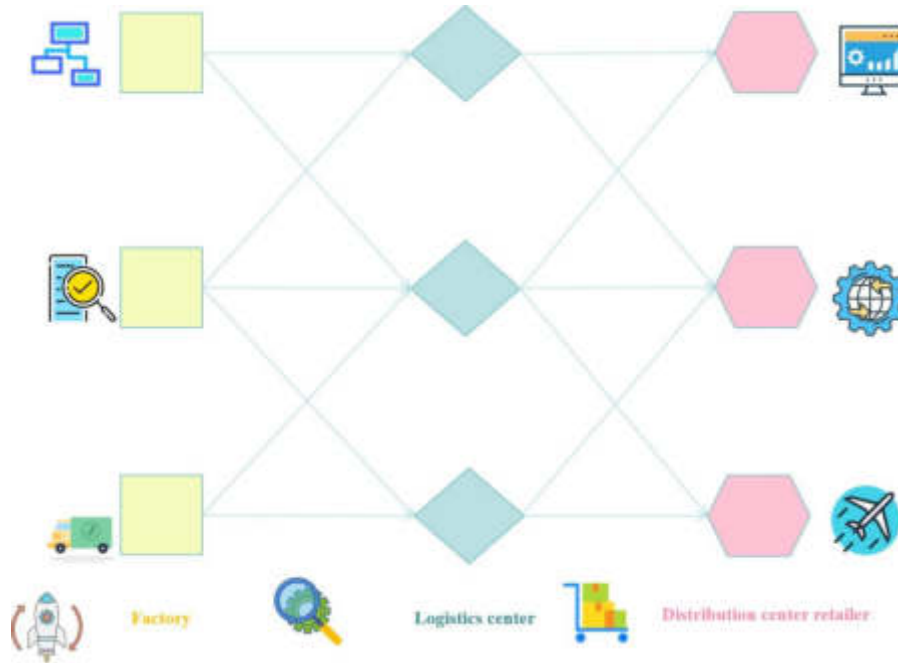


Fig. 2.1: Logistics Network Structure

and z is the optimal total shipping cost, the mathematical model is:

$$\min z = \sum_{i=1}^m \sum_{j=1}^m c_{ij} x_{ij} \tag{2.1}$$

Constraints:

$$\sum_{i=1}^m x_{i,j} = b_j, j = 1, \dots, n \tag{2.2}$$

$$\sum_{i=1}^n x_{i,j} = a_j, i = 1, \dots, n, x_{i,j} \geq 0$$

Therefore, equation 2.1 and its constraints precisely constitute the mathematical model of the linear programming problem. By using the linear programming method to solve it, the optimal solution can be obtained, which is to minimize the total freight cost while satisfying the possible supply and demand of the place of origin [8].

2.2. Optimization modeling of logistics distribution and transportation network based on smooth reliability. The main purpose of optimizing logistics distribution and transportation network is to improve the efficiency of reliability and accessibility of distribution network. Therefore, the purpose of the modeling is to achieve the reliability and reliability of shipping and transportation systems. According to the characteristics of multi-city, multi-variety, small-scale, multi-package, and short cycle logistics distribution, the following parameters are mainly determined:

1. The total amount of goods on each distribution route shall not exceed the limits of vehicle capacity and load capacity [9,10];
2. Within the allowable range of existing transportation capacity in the logistics center;
3. During the delivery process, each delivery point can only be accessed once and must be accessed once;

4. Each vehicle can only serve one route, and each delivery vehicle must depart from the distribution center and finally return to the distribution center;
5. Delivery costs should be controlled at a certain level.

Following the above constraints, establish an optimization model as follows:

$$\begin{aligned}
 \max Z &= \max \Psi = \sum_{n=1}^N \xi_n \Psi_n \\
 \text{s.t.} & \begin{cases} \Psi_n = \prod \Psi_{ij} \\ q_n \leq Q_n \\ N \leq M \\ \sum_{i=0}^n x_{ij} = 1 (j = 1, 2, \dots, n) \\ \sum_{i=0}^n x_{ij} = 1 (i = 1, 2, \dots, n) \\ \sum_{i=0}^n x_{i0} = \sum_{i=0}^n x_{i0} = N \\ \sum_{i=0}^n \sum_{j=0}^n C_{ij} X_{ij} \leq A \end{cases} \quad (2.3)
 \end{aligned}$$

In the formula, Ψ_n represents the smoothness and reliability of the nth distribution line in the network; Q_n is the load capacity of the vehicles arranged on the nth route; Q_n is the total amount of goods at all distribution points on the nth route [11]; N is the overall number of distribution routes; M is the number of delivery vehicles in the logistics center; C_{ij} is the cost from node i to node j , $x_{ij} = \begin{cases} 1, & \text{The vehicle line passes the arc}(i,j) \\ 0, & \text{The vehicle line does not pass through the arc}(i,j) \end{cases}$ A is the cost quota at a certain level of certainty, which is a constant and can be determined based on empirical values.

2.3. Implementation Algorithm of Logistics Distribution and Transportation Network Optimization Model Based on Smooth Reliability. Due to the large number of units in large-scale logistics distribution network systems, traditional algorithms for solving nonlinear programming may encounter "curse of dimensionality" or combinatorial explosion problems; Meanwhile, the design or improvement of the actual distribution network is often selected from several feasible solutions. Therefore, based on the specific situation of the logistics distribution network system, the author chooses ant colony algorithm for system optimization and solution.

(1) *Introduction to Ant Colony Algorithm.* The principle of the ant colony algorithm is to first create an ant colony with a certain number of ants, and let each ant create a solution or part of the solution [12]. Then, each ant starts from the first node of the problem and selects the next node to switch according to the concentration of pheromones along the path until the solution is reached. Each ant releases pheromones that are proportional to the quality of the solution in the state it passes through based on the quality of the solution found; Afterwards, each ant begins a new solving process until a satisfactory solution is found. Ant colony algorithm needs to follow the following rules in computing loops:

State transition rule: The probability calculation method for the h-th ant at location r to choose to transition to location s is:

$$P_{rs}^k = \begin{cases} \frac{[\tau(r,s)] \cdot [\eta(r,s)]^\beta}{\sum_{u \in J_k} [\tau(r,u)] \cdot [\eta(r,u)]^\beta} & s \in J_k \\ 0, & \text{other} \end{cases} \quad (2.4)$$

In the formula, P_{rs}^k represents the probability of ant k transferring from the current location r to the location, $\eta(r, s) = 1/d(r, s)$, $d(r, s)$ represents the distance between location r and location s , the expected degree of $\eta(r, s)$ from location r to location s [13]. $\tau(r, s)$ represents the residual information between location r and location s , J_k represents a location that the k -th ant has not yet visited, parameter β is a parameter used to adjust the relationship between $\tau(r, s)$ and $\eta(r, s)$, it represents the different roles played by the information accumulated by ants during their movement and heuristic factors in ant path selection. Specifically, it represents

the concentration of pheromones on a path and the reciprocal of its length, which person has the greater importance in probability calculation.

Ant's next path selection rule:

$$s = \begin{cases} \arg \max_{u \in J_k(r)} [\tau(r, s) \cdot \eta(r, s)]^\beta \\ S, q \leq q_0(\textit{exploitation}), \textit{otherwise}(\textit{exploitation}) \end{cases} \quad (2.5)$$

In the formula, q is a random variable uniformly distributed on $[0, 1]$, q_0 is a parameter on $[0, 1]$, and S is a probability distribution calculated based on the state transition rule for selection.

Local update rule: During the process of establishing a solution, each ant also undergoes a local update of pheromones. The information intensity of each arc on its path is adjusted according to the following formula:

$$\tau(r, s) = (1 - p) \cdot \tau(r, s) + p \cdot \Delta\tau(r, s) \quad (2.6)$$

In the formula, p represents the volatility factor of local pheromones, and $1-p$ represents the retention rate of pheromones. In order to prevent infinite accumulation of information, the value range of p is limited to $(0, 1)$. $\Delta\tau(r, s)$ is the amount of information increase, which refers to the concentration of pheromones that Ant k increases along the path from City r to City s during the time period t to $(t+n)$.

Global update rule: After all ants have completed one cycle, only the ants that have generated the global optimal solution (i.e. the ants that have constructed the shortest path from the beginning to the present) have the opportunity to perform global updates. The pheromones on all paths passed by the optimal ant are globally updated according to the following formula. The pheromones on paths that do not belong to the optimal ant are updated to 0.

$$\tau(r, s) = \begin{cases} (1 - \alpha) \cdot \tau(r, s) + \alpha \cdot \Delta\tau \\ 0 \end{cases} \quad (2.7)$$

In the formula, α represents the volatility factor of global pheromones, $1-\alpha$ represents the retention rate of global pheromones, and the value range of α is between $(0, 1)$. $\Delta\tau = 1/L_b$, L_b is the shortest path length calculated by the algorithm (the shortest path here only refers to the shortest path obtained in this cycle).

(2) *Model Ant Colony Algorithm Solving*. Using ants instead of vehicles, the next delivery point to be served will cause the total carrying capacity to exceed the vehicle's capacity, and then return to the distribution center, indicating that the vehicle has completed the transportation. Then change to another car and continue to serve the other delivery points until all delivery points receive a service, which represents the completion of an ant tour. When all ants patrol once, it is recorded as a cycle. After one cycle, calculate the pheromone increment and update the pheromones on the relevant paths based on the quality of each ant's patrol history (objective function value). The specific implementation steps are as follows:

Step 1: Initialize each basic parameter and set $nc=Q$;

Step 2: Calculate the transition probability;

Step 3: Based on the calculated transition probability and randomly generated q value, select the next path for each ant to move;

Step 4: When each ant passes through an edge and reaches the next delivery point, a local update of pheromones is performed on this edge according to local update rules;

Step 5: Repeat steps 2 to 4 for each ant in step 5 until all distribution points have ants passing by and the needs of the distribution points are met [14];

Step 6: Find the path that matches the optimal goal of the model among all generated paths, and the ant that passes through that path is the optimal ant;

Step 7: Perform a global pheromone update on each edge passed by the optimal ant according to the global update rule for this path;

Step 8: Repeat steps 2 to 7 until the nc reaches the specified maximum number of iterations or no better solution appears for several consecutive generations.

In this article, Matlab programming is used to implement the ant colony algorithm and solve practical problems.

Table 3.1: Customer demand at delivery points

Delivery point	Demand/t	Delivery point	Demand/t
P1	0.8	P6	1.4
P2	1.4	P7	1.5
P3	1.1	P8	0.7
P4	1.3	P9	0.4
P5	1.3		

Table 3.2: Distance between Distribution Centers and Distribution Points

distance	P0	P1	P2	P3	P4	P5	P6	P7	P8	P9
P0	1	11	9	8.5	10	8	8.5	5	14	10
P1	11	1	6	-	-	-	-	-	18	9
P2	9	6	1	7	-	-	-	-	-	-
P3	8.5	-	7	1	7	12	-	-	-	-
P4	9	-	-	7	1	8	-	-	-	-
P5	8	-	-	12	8	1	7	11	-	-
P6	8.5	-	-	-	9	1	6	-	-	-
P7	5	-	-	-	-	11	6	1	11	16
P8	14	18	-	-	-	-	-	11	1	8
P9	10	9	-	-	-	-	-	16	8	1

Table 3.3: Reliability of unobstructed road units between distribution center P0 and distribution point P₉

Road unit	Unblocked Reliability	Road unit	Unblocked Reliability
①	0.91	2	0.86
②	0.96	3	0.91
③	0.96	4	0.8
④	0.86	5	0.91
1	0.96		

3. Results and Analysis. A certain distribution center P₀ needs to deliver goods to 9 distribution points P_j (j=1, 2, 3,..., 9). The demand for each distribution point is q_j, and the specific values are shown in Table 3.1. There are 2, 4, and 6 ton trucks available for allocation in the distribution center. The distances between distribution centers and distribution points, as well as between distribution points, are shown in Table 3.2. The unit distance cost is 2Q, and the cost quota A is 2000. Given the smoothness and reliability of the road sections and intersections between various distribution nodes, we will try to develop a reasonable distribution route.

3.1. Smooth reliability between distribution centers and distribution points, as well as between distribution points. Based on the smoothness and reliability of the road sections and intersections between various distribution nodes, the OD formula for calculating the smoothness and reliability of the inter transportation system is used to calculate the smoothness and reliability between distribution centers and distribution points, as well as between distribution points and distribution points. Taking the road network between distribution center P₀ and distribution point P₉ as an example for calculation, the smoothness reliability and road network of the road units between the nodes are shown in Table 3.3 [15,16].

According to the OD formula for calculating the smoothness reliability of the transportation system, the smoothness reliability between distribution center P and distribution point P is 0.79. Similarly, calculate the smoothness reliability between other distribution nodes, as shown in Table 3.4 and Figure 3.1 [17].

3.2. Develop delivery routes. Apply optimization models and use ant colony algorithm for solving. When initializing the data, the maximum number of ants taken is 5Q, and the maximum number of executions

Table 3.4: Reliability of smoothness between distribution points

Delivery point pairs	Unblocked Reliability	Delivery point pairs	Unblocked Reliability
P01	0.76	P19	0.66
P02	0.62	P23	0.8
P03	0.72	P34	0.63
P04	0.7	P35	0.69
P05	0.66	P5	0.79
P06	0.7	P56	0.76
P07	0.77	P7	0.7
P08	0.72	P67	0.79
P09	0.8	P78	0.76
P12	0.73	P79	0.68
P18	0.76	P89	0.73

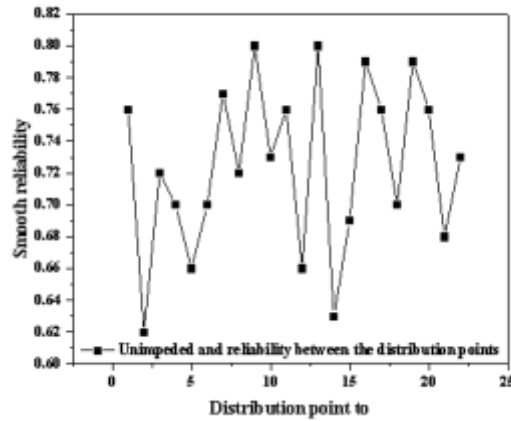


Fig. 3.1: Reliability of smooth delivery between different distribution points

is 200Q. After iterative calculation, the following three reasonable delivery routes are finally obtained.

Route I: P0-P3-P2-P1-P0. Using a 4-ton vehicle with a delivery distance of 28.5 [18]

Route II: P0-P4-P5-P6-P7-P0 uses a 6t vehicle with a delivery distance of 30;

Route III: P0-p8-P9-P0. Using a 2T vehicle with a delivery distance of 29, the optimal smoothness and reliability of the logistics distribution network under various constraints reached 0.297.

3.3. Comparative analysis. Using the mileage saving method, solve this example and obtain three delivery routes. Route i: P0-P2-P1-P9, P8-P0 uses a 4t vehicle with a delivery distance of 41; Route ii: P0-P3-P4-P0 uses a 2-ton vehicle with a delivery distance of 21.5; Route iii: P0-P5-P6-P7-P0 uses a 6t vehicle with a delivery distance of 20, and the reliability of the delivery network under this delivery plan is 0.229. Through calculation and analysis, it can be concluded that with a 6% increase in total delivery distance, the reliability of the delivery network has increased by 30%. This indicates that when using the model built by the author for distribution network optimization, effective optimization of network smoothness and reliability can be achieved, while only increasing the distance by a small amount. The optimal reliability of a smooth distribution network means that the probability of delivery delays is minimized, which is the most powerful guarantee for the effective accessibility of delivery. In today's fast-paced society, where the value of time is increasingly valued by people, the distribution model built by the author undoubtedly has certain practical significance [19,20].

4. Conclusion. Due to the fact that the optimization of logistics distribution and transportation networks only fully utilizes the existing road network and cannot transform it, further analysis of the reliability of logistics distribution and transportation networks can only be conducted based on the analysis of the existing road network's smoothness and reliability. The research conducted by the author aims to maximize the smoothness and reliability of logistics distribution networks at a certain cost level, thereby ensuring the effective accessibility of logistics distribution. However, in practical applications, if the improvement of the smoothness and reliability of logistics distribution networks requires a significant cost, it is necessary to consider whether the distribution center is in a reasonable transportation location or whether the division of distribution areas is reasonable. This is a direction that needs further research in the future.

5. Acknowledgements. 1.2022 Hainan Provincial Natural Science Foundation High-level Talent Project, Research on the Development Potential of Hainan Free Trade Port Industry Chain Based on Dynamic Network Planning Model, 722RC728, 2022-2025.

2.Study on the Strategy of Extending and Complementing the Shipping Industry Chain in Hainan Free Trade Port (HNSK(YB)23-17),2023-2025.

3.Key Laboratory of Philosophy and social Science in Hainan Province of Hainan Free Trade Port International Shipping Development and Property Digitization, Hainan Vocational University of Science and Technology Hainan Social Science [2022] No. 26

REFERENCES

- [1] Zhao, X. Y., & Sun, X. (2022). Optimization design of multi-vehicle urban logistics distribution based on ant colony optimization. Springer, Singapore.3456(547),888
- [2] Sun, B., Hu, Z., Liu, X., Xu, Z. D., & Xu, D. (2022). A physical model-free ant colony optimization network algorithm and full scale experimental investigation on ceiling temperature distribution in the utility tunnel fire. *International Journal of Thermal Sciences*, 174(566), 107436-.
- [3] Cao, M. B. J., emailprotected, Emailprotected, E., Cao, J., Juexian Cao * Juexian CaoDepartment of Physics & Hunan Institute of Advanced Sensing and Information Technology, Xiangtan University, Xiangtan, PR China*Email: emailprotectedMore by Juexian Cao, & Xu, M. B. W., et al. (2023). Optimization of the mixed gas detection method based on neural network algorithm.46(7679),135
- [4] Ashour, M., Elshaer, R., & Nawara, G. (2022). Ant colony approach for optimizing a multi-stage closed-loop supply chain with a fixed transportation charge. *Journal of Advanced Manufacturing Systems*.635(12),78
- [5] Ba, D., Yang, Y., Zhang, Y., & Song, H. (2022). Multi-objective optimal configuration analysis of energy storage system in distribution network based on improved lion colony algorithm. *Journal of Physics: Conference Series*, 2401(1), 012092-.
- [6] Yang, T., & Wang, W. (2022). Logistics network distribution optimization based on vehicle sharing. *Sustainability*, 14(w35),325.
- [7] Wang, T., Qi, Q., Zhang, W., & Zhan, D. (2023). Research on optimization of profile parameters in screw compressor based on bp neural network and genetic algorithm. *Energies*.75774(36),072
- [8] Ahmadi, P., & Rastegar, H. (2022). Chp systems optimal allocation in the interconnected heat and electricity distribution network based on minimizing electrical and heat transfer losses. *IET generation, transmission & distribution*769(13), 16.
- [9] Qian, C., & Wang, A. (2022). Distribution network reconfiguration based on improved differential evolution ant colony algorithm. 2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering 457(ICBAIE), 234-240.
- [10] Ehsan Azad-Farsani, Hamed Zeinoddini-Meymand, & Jafari, H. (2023). Distribution network reconfiguration for minimizing impact of wind power curtailment on the network losses: a two-stage stochastic optimization algorithm. *Energy Science And Engineering*, 11(2), 849-859.
- [11] Chen, S., Hu, W., Du, Y., Wang, S., Zhang, C., & Chen, Z. (2022). Three-stage relaxation-weightsum-correction based probabilistic reactive power optimization in the distribution network with multiple wind generators. *International Journal of Electrical Power & Energy Systems*, 141(379), 108146-.
- [12] Sobouti, M. A., Bigdeli, M., & Azizian, D. (2023). Rooftop photovoltaic system allocation to improve the distribution transformers life span using golden ratio optimization algorithm. *World Journal of Engineering*, 20(3), 458-471.
- [13] Liu, C., Tang, C., & Li, C. (2023). Research on delivery problem based on two-stage multi-objective optimization for takeout riders. *Journal of Industrial and Management Optimization*, 19(11), 7881-7919.
- [14] Yang, L., & Qin, Y. (2023). Research on rolling co-optimization of fault repair and service restoration in distribution network based on combined drive methodology. *Electric Power Systems Research*679(Jul.), 220.
- [15] Sun, B., Hu, Z., Liu, X., Z.-D., X., & Xu, D. (2022). A physical model-free ant colony optimization network algorithm and full scale experimental investigation on ceiling temperature distribution in the utility tunnel fire. *International Journal of Thermal Sciences*467(174-), 174.
- [16] Liu, D., Hu, X., & Jiang, Q. (2023). Design and optimization of logistics distribution route based on improved ant colony algorithm. *Optik*, 273(12), 170405-.

- [17] Tang, D., & Gong, S. (2023). Trajectory optimization of rocket recovery based on neural network and genetic algorithm. *Advances in Space Research: The Official Journal of the Committee on Space Research*7768(COSPAR)(8), 72.
- [18] Hu, R., Wang, W., Wu, X., Chen, Z., & Ma, W. (2022). Interval optimization based coordinated control for distribution networks with energy storage integrated soft open points. *International Journal of Electrical Power & Energy Systems*, 136(67), 107725-.
- [19] Ni, Z., Cai, S., & Ni, C. (2023). Research on energy-saving strategy of wireless sensor network based on improved ant colony algorithm. *Sensors and materials: An International Journal on Sensor Technology*77(6 Pt.1), 35.
- [20] Ji, Y., Chen, X., Wang, T., He, P., Jin, N., & Li, C., et al. (2022). Dynamic reactive power optimization of distribution network with distributed generation based on fuzzy time clustering. *IET generation, transmission & distribution*56(7), 16.

Edited by: Zhigao Zheng

Special issue on: Graph Powered Big Aerospace Data Processing

Received: Jan 22, 2024

Accepted: Mar 4, 2024



REVIEW OF AUTOMATED TEST CASE GENERATION, OPTIMIZATION, AND PRIORITIZATION USING UML DIAGRAMS: TRENDS, LIMITATIONS, AND FUTURE DIRECTIONS

SRINIVASA RAO KONGARANA * A ANANDA RAO † AND P RADHIKA RAJU ‡

Abstract. This systematic literature review examines the effectiveness of automated test case generation, optimization, and prioritization methods based on Unified Modeling Language (UML) diagrams. The review summarizes the methods, main contributions, and limitations, and suggests areas for future research. This paper examines various optimization algorithms, model-based testing methods, and UML diagram validation methods to determine how well they perform. The review highlights some issues with the current situation, such as the fact that it only examines a few types of UML diagrams and does not go into great detail about how they work or compare to other diagrams. However, it also suggests ways in which these issues could be addressed in future research. Some of the suggested directions include researching different modeling languages and devising solutions to handle the complexity of system models. Model-based testing should also be combined with optimization and prioritization methods to increase the flexibility and usefulness of research in this field. This article makes no direct comparisons to UML diagrams, but it does provide a thorough discussion of the current state of the art and a list of strategic priorities to advance the field of automated test case generation, optimization, and prioritization. These reviews are useful for both researchers and practitioners because they demonstrate how things are currently done and how they should be done in the future.

Key words: UML, Optimization, Prioritization, Test Case Generation, Sequence Diagram, SLR

1. Introduction. Software testing finds and fixes system bugs, which is a critical part of the software development lifecycle. Creating inputs and expected outputs to assess the software’s functionality and behavior is known as test case generation, and it is an essential part of software testing [1]. The ability of automated test case generation techniques to increase software testing effectiveness, decrease human labor, and increase efficiency has drawn attention in recent years [1]. Test case generation using Unified Modeling Language (UML) diagrams shows promise for automated testing. Software systems can be visually represented with UML diagrams, which depict interactions, behavior, and structure [2]. These diagrams are appropriate for automated test case generation because they standardize and simplify software system modeling. From 2010 to 2022, a thorough overview of research on creating test case diagrams from UML diagrams is intended to be provided by this systematic literature review. This systematic review outlined the field’s present methods, results, research trends, shortcomings, and potential future directions. The use of UML diagrams in test case generation has several advantages. Between test cases and system requirements or design specifications, UML diagrams provide a clear mapping [3]. Making it simpler to track test coverage and ensure system behavior matches design, traceability enhances understanding and documentation of the testing process.

Because UML diagrams automatically generate test cases and model system behavior, they also enhance model-based testing [4]. Automated test case generation is minimized, dynamic behavior is managed, and system complexity is captured through model-based capturing. Efficient system functionality coverage is another advantage of UML diagrams. A comprehensive view of the system is provided by UML diagrams, which enable testers to pinpoint crucial paths, significant interactions, and test scenarios [5]. With the aid of this data, automated test case generation techniques are able to produce an exhaustive collection of test cases that encompass every facet of the system, guaranteeing a comprehensive assessment of its functionality.]

UML diagrams have limitations when it comes to test case generation. One drawback is the emphasis on state machine, activity, and sequence UML diagrams. These diagrams show important system behavior,

*CSE Department, JNTUA College of Engineering (Corresponding author, srinivas.cst4@gmail.com)

†CSE Department, JNTUA College of Engineering (akepogu@gmail.com).

‡CSE Department, JNTUA College Of Engineering (radhikaraju.p@gmail.com).

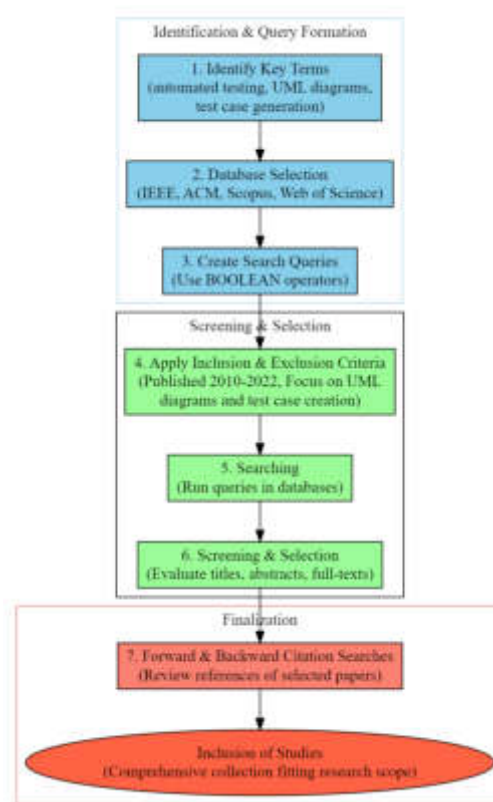


Fig. 2.1: Architecture of the systematic literature review (SLR)

but they may not cover all software details, resulting in test coverage gaps. Evaluate whether UML-based test case generation is applicable to complex systems as well. UML diagrams might not adequately depict the complex behavior, interdependencies, and edge cases of complex software systems. Assessing the effectiveness and scalability of these techniques is essential before applying them to large, complex systems.

Empirical evaluations and comparative analyses of the suggested approaches are crucial to assess their effectiveness, efficiency, and practicality. The lack of thorough empirical evaluations in many of the current studies limits their understanding of adoption and performance in practice. Consolidating knowledge and identifying gaps and limitations in automated test case generation using UML diagrams are the goals of this systematic literature review. Insights into new trends will be provided by this review, which will also help to clarify the current research landscape. By addressing limitations, investigating new avenues for research and development, and promoting the adoption of successful and efficient test case generation methods in software testing practices, the objective is to advance the field of automated test case generation.

2. Search Strategy. Figure 2.1 depicts the systematic literature review (SLR), which used a thorough and methodical search strategy to uncover relevant literature on creating test cases from UML diagrams. The method was carefully planned to ensure that a comprehensive set of research papers were assembled that met the specific criteria for inclusion and covered the entire scope of the research question. With this focused and organized approach, the review sought to compile a body of work that demonstrates the current state and recent progress in creating test cases from UML diagrams.

The following steps were used to create an efficient search strategy:

1. Finding the key terms and concepts associated with the research topic: "automated testing," "UML diagrams," "test case generation," and other related terms.

2. Database selection: Digital libraries and pertinent scholarly databases were looked through. In computer science, software engineering, and testing, common databases include IEEE Xplore, ACM Digital Library, Scopus, and Web of Science.
3. Creating search queries: The key terms and concepts were combined to create search queries. The search space was widened and terms were connected using BOOLEAN operators (AND, OR). UML diagrams were used in the search queries to locate studies on test case generation.
4. Application of inclusion and exclusion criteria: Clearly defined inclusion and exclusion criteria were created in order to weed out studies that did not fit the intended research scope. Published in English between 2010 and 2022, with a focus on UML diagrams and test case creation.
5. Searching: Queries were run in a few databases to get the desired results. Consistency and accuracy were sought after by several researchers working independently.
6. Methods of screening and selection: The titles and abstracts of the retrieved studies were evaluated for their applicability to the research question. The full-text eligibility of the chosen studies was subsequently assessed for systematic literature reviews.
7. Forward and backward citation searches: To increase the search's rigor, we looked through the reference lists of a few chosen papers and carried out forward and backward citation searches to locate more pertinent studies that the first search had overlooked.

In order to reduce bias, incorporate pertinent studies, and offer a thorough and representative sample of UML diagram test case literature, this systematic search approach was employed. The method found and chose studies that supported the systematic literature review and the research goals.

2.1. Search Queries. Compatible search queries are terms or keywords that align with the objectives and topic of a study. In databases and other sources, these queries discover relevant and compatible literature.

Creating compatible search queries requires researchers to take into account the key concepts, terms, and relationships related to their research topic. While excluding irrelevant studies, the queries should cover a wide range of relevant research. Additionally, they should consider variations in terminology or synonyms used in the literature.

Here are some search queries related to perform suggested review:

1. Automated test case generation techniques using UML diagrams
2. Trends in test case generation, optimization, and prioritization with UML diagrams
3. Limitations of automated test case generation with UML diagrams
4. Future directions in test case generation, optimization, and prioritization using UML diagrams
5. Model-based testing and UML diagrams for automated test case generation
6. Optimization algorithms for test case generation with UML diagrams
7. Validation techniques for automated test case generation using UML diagrams
8. Comparative analysis of automated test case generation approaches with UML diagrams
9. Handling complexities in system models for test case generation with UML diagrams
10. Integration of optimization methods, prioritization techniques, and model-based testing for test case generation with UML diagrams

These search queries can assist researchers and practitioners in exploring pertinent literature, trends, limitations, and future directions in the field of automated test case generation, optimization, and prioritization using UML diagrams.

2.2. Search Statistics. Table 2.1, Figure 2.2 categorizes scholarly articles based on the frequency of keywords related to test case development techniques using Unified Modeling Language (UML) diagrams. It shows that "UML" is the most prevalent keyword, appearing in 36 articles, which constitutes 69% of the literature. This is followed by "Test Case Generation" with 30 articles (58%), suggesting a strong academic focus on these areas. Less prevalent, though still significant, are articles on "Sequence Diagram" and "Activity Diagram," with 8 (15%) and 14 (27%) articles respectively, indicating a moderate research interest. "Test Case Optimization" and "Test Case Prioritization" are covered in 11 (21%) and 9 (17%) articles, pointing towards an interest in enhancing the efficacy of testing procedures.

Finally, "UML Diagrams" as a collective category feature in 23 articles, making up 44% of the distribution, highlighting the central importance of UML in the discourse of test case methodologies. This tabulation

Table 2.1: Distribution of Articles According to Keywords

S.No	Number of Articles	Keywords	Ratio of Articles
1	36	UML	69.00%
2	30	Test Case Generation	58.00%
3	8	Sequence Diagram	15.00%
4	14	Activity Diagram	27.00%
5	11	Test case optimization	21.00%
6	9	Test Case Prioritization	17.00%
7	23	UML Diagrams	44.00%

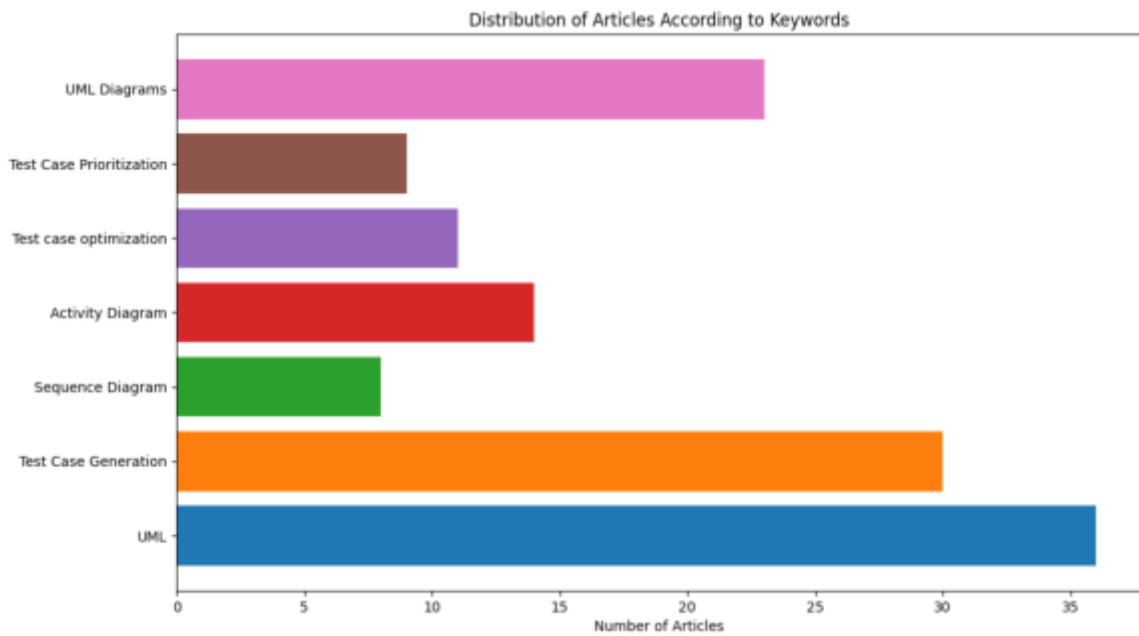


Fig. 2.2: The distribution of articles according to keywords

underscores the varied levels of academic engagement across different facets of test case methodologies within the realm of UML.

A filter was also applied to the publication type of the articles, resulting in the removal of 5 articles. The selected articles' type distribution is displayed in Table 2.2, Figure 2.3. The publisher reviewed the 52 remaining articles, discarding none. By publisher, Table 2.2 lists the articles.

The Figure 2.4, Table 2.3 resulting in the removal of two articles, articles were filtered in the final stage based on their year of publication. After applying qualitative synthesis factors to the remaining 52 articles, two more were discarded. Table 2.3 displays the articles' publication years.

2.3. Research Questions.

Research Question 1: What is the most effective and efficient approach for automated test case generation in object-oriented software development, considering the use of UML behavioral models and diagrams?

Research Question 2: What are the most effective and efficient approaches for automated test case generation in software development, considering the utilization of optimization techniques and UML diagrams?

Research Question 3: How can automated test case generation techniques using UML diagrams be effectively

Table 2.2: Distribution of Articles by Publisher

S.No	Publisher	Ratio of articles
1	Elsevier	8%
2	Springer	6%
3	IEEE	2%
4	Conference	23%
5	Other Journals	62%



Fig. 2.3: The distribution of articles by publisher

utilized to improve software development processes, considering optimization methods, soft computing techniques, and the automation of test case generation?

Research Question 4: How can automated test case generation techniques utilizing UML diagrams be improved to address limitations such as applicability to specific diagram types, complexity limitations, and lack of comparative analysis, while considering factors like model-based testing, optimization, and validation approaches?

Research Question 5: How can the limitations of existing approaches for test case generation and optimization in model-driven software development using UML diagrams be overcome to improve their applicability, effectiveness, and efficiency?

Research Question 6: How can test case prioritization techniques be enhanced to address the limitations identified in the reviewed papers?

3. Literature Review. This review on "Test Case Generation and Optimization" presented in section 3.1 provides valuable insights into the state of software testing today by examining the intricate processes and state-of-the-art methodologies. The architecture diagram shown in figure 3.1 that serves as a visual guide for the plethora of studies, methodologies, and findings that make up this field is used to describe our systematic literature review strategy in detail. These images not only demonstrate the meticulous manner in which data was gathered and examined, but they also demonstrate the connections between various research topics,

Table 2.3: Fraction of articles by year of publication

Article Count	Publish Year
15	2019
7	2020
12	2021
5	2022
13	Others

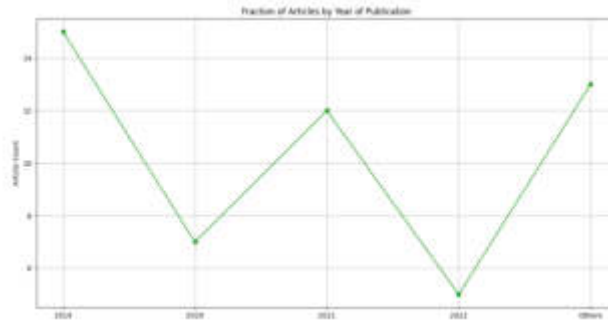


Fig. 2.4: Fraction of articles by year of publications

ranging from the most recent optimization algorithms to the use of UML diagrams to create test cases. By assembling an extensive array of contemporary models and methodologies, this review seeks to provide readers with a comprehensive understanding of test case generation and optimization. It also emphasizes how crucial systematic reviews are to expanding the realm of software testing possibilities.

The review presented in section 3.2 underscores the significance of prioritizing test cases to boost the effectiveness of software testing procedures. The central component of our study is a meticulously constructed architecture diagram that shown in figure 3.2. It demonstrates the sequential and iterative processes involved in prioritizing test cases, from locating them and grouping them according to their importance to applying various prioritization criteria and algorithms and ultimately executing the tests in an orderly fashion. The iterative refinement inherent in test case prioritization criteria is highlighted by the feedback loop for modifying test case prioritization based on empirical results. We aim to provide you with a comprehensive understanding of the approaches, issues, and advancements in test case prioritization by based our review on this architectural framework. Researchers and practitioners who are attempting to enhance software testing outcomes will benefit from this.

3.1. Test case Generation and Optimization. The use of UML diagrams to streamline test cases in software development was covered by Tiwari, R. G., et al. [1]. Contribution includes recommending the usage of UML diagrams (activity, state, and sequence) and outlining their benefits for streamlining test cases. The technique lacks any actual data or experiments and is based solely on a survey of the literature. The goal is to educate software engineers about the usage of UML diagrams for test case simplification. The absence of empirical data, unresolved difficulties or constraints, and the narrow focus on just three categories of UML diagrams are all drawbacks. To establish efficacy and resolve any difficulties, more study is required.

J. Cvetkovi and M. Cvetkovi., [3] provided a research on the creation of test cases with UML diagrams that concentrated on modelling depression brought on by internet addiction. By recommending a technique for creating test cases using UML diagrams and offering a case study to illustrate its use, the paper makes a contribution to the subject. The process comprises categorising test case creation based on unit diagrams or combinations of UML diagrams and employing a variety of UML diagrams. The objective is to advance the area of software testing by offering a fresh method for creating test cases. The case study's specificity

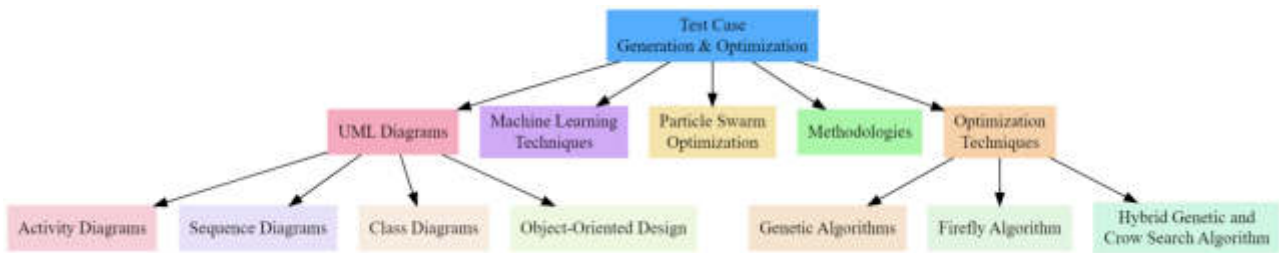


Fig. 3.1: The process of Test Case Generation and Optimization

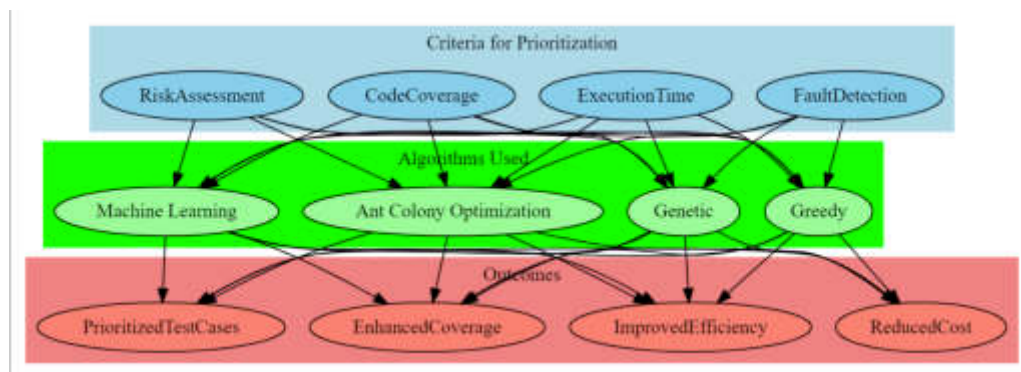


Fig. 3.2: Process model of the Test case Prioritization

and the absence of comparison with other approaches are also drawbacks. An approach for creating test cases from UML sequence diagrams combining MDE (“model-driven engineering”) and MBT (“model-based testing”) methodologies was given by Rocha, M., et al. in [5]. Modifying transformation rules, using real software models in a case study, and using the ModelJUnit and JUnit libraries are all part of the process. The contribution consists of a methodical process for automatically creating test cases from UML models, which raises the productivity and calibre of software development. The inability to use nested combined fragments in sequence diagrams, the manual creation of stubs, and the limited usage of one case study for demonstration are some limitations.

B. N. Biswal. Using UML behavioural models, [6] suggested a technique for creating and refining test cases of object-oriented software. It offers a technique for creating test cases that uses UML activity diagrams, sequence diagrams, as well as class diagrams. An error minimization method for test case optimisation is also presented in the study. Its application to particular categories of object-oriented software as well as reliance on the excellence of UML behavioural models are limitations.

Meena, D.K. [7] A technique for creating test cases using UML diagrams, notably Interaction Overview diagrams as well as Sequence diagrams, was presented by [7]. The difficulty of test case selection in object-oriented programmes is what this aims to address. Building UML diagrams, creating intermediate graphs, and creating test cases that reflect various scenarios are all part of the technique. The paper’s contribution is the message route coverage. Limitations include the fact that it only applies to object-oriented programmes and the absence of actual data contrasting the suggested strategy with other ones.

Jiang, L., and Li, Y. A technique for creating test cases using UML sequence diagrams was covered in [8]. By examining the sequence diagram as well as assessing the message sequence, it gives instructions for creating test cases. The idea is to provide a useful method for testing object-oriented applications. The study does not, however, present a prototype system for autonomous test case generation, and it does not discuss the drawbacks or viability of the suggested approach in practical contexts.

A. Herrmann, M. Felderer, and M. [9] looked at mistakes occurring while manually deriving test cases from state machines and activity diagrams using UML. The goal is to offer instructions for methodically generating test cases while avoiding mistakes. Participants in the methodology's controlled student experiment create test cases from state machines and UML activity diagrams. The study offers a taxonomy of faults, evidence that activity diagrams are more error-prone than state machines, and recommendations for minimising errors. The utilisation of student participation and the emphasis on UML activity diagrams as well as state machines are limitations.

An approach for creating test cases using UML sequence diagrams as well as interaction overview diagrams was proposed by Jena, A. K., et al. in [10]. As part of the technique, produced graphs from various UML components are combined to create test cases and find errors early in the design process. The goal of the study is to identify various error types and raise software quality. The emphasis on a single case study and the lack of a comparison with alternative techniques for creating test cases are both drawbacks.

A. Herrmann, M. Felderer, and M. [11] findings of a controlled experiment [11] examining mistakes caused during manual test case derivation from state machines and UML activity diagrams were provided. The study offers a taxonomy of faults, evidence that activity diagrams are more error-prone than state machines, and recommendations for minimizing errors. The concentration on certain diagram kinds without comparison to other types or automated test case derivation techniques, as well as the sample size's relative smallness, are limitations.

A technique for creating test cases using state chart diagrams and UML use case diagrams was put out by Jagtap, S., et al. [12]. Utilizing testing criteria that address the diagrams' state and transition is part of the technique. The goal is to aid software engineers in the early phases of software development in the creation of efficient test cases. Limitations include the emphasis on use case diagrams as well as state chart diagrams in UML 2.0 and the potential need for manual testing to ensure thorough coverage.

An automated test case creation method for unstructured SysML "Activity Diagrams" (Ads) was provided by Yin, Y., et al. [13]. The process entails converting the ADs into an IBM ("Intermediate Black Box Model"), which is used to generate test cases. The objective is to offer industry practitioners employing SysML ADs an efficient testing methodology. The concentration on the AD model alone, without taking into account other SysML models, is the restriction.

In order to create test cases from UML state chart diagrams, Salman, Y. D., et al. [14] concentrated on finding the most efficient combination of coverage requirements, notably in handling loops. The study suggests appropriate coverage criteria and offers formulas to calculate coverage percentages. The goal is to increase test data generation's efficiency during software testing. The emphasis on UML state chart diagrams and the requirement for context adaption are limitations.

A method for automated test case development utilising UML Statechart as well as Sequence diagrams was proposed by Efendi, N. B. M., et al. [15]. The approach includes using UML diagrams to create test cases automatically. The aim is to keep the price of software development as low as possible. The drawback is the emphasis on UML Statechart and Sequence diagrams at the expense of other UML diagram types.

PSO ("Particle Swarm Optimization") was described by Prakash, V. C., et al. [16] as a tool for automated test case development in combinatorial testing. In order to address the optimisation issue in test case generation, the study offers a comprehensive assessment of PSO and its variations. The goal is to demonstrate how PSO-based algorithms may be used to reduce test suite sizes and boost programme dependability. The emphasis on PSO and its comparison to other optimisation methods is one limitation.

Using a dynamic programming technique and tester specification, Kamonsantiroj, S., et al. [17] suggested a memorising strategy for producing test cases in concurrent UML activity diagrams. The route explosion challenge in testing concurrent systems is the key contribution. The goal is to prevent the development of all feasible concurrent activity routes in order to reduce the explosion of test cases. The application to other diagram types or concurrent systems, as well as the emphasis on concurrency testing in activity diagrams, are limitations.

Mokhati, F.; Dehimi, N. E. H. A novel method for evaluating multi-agent systems using AUML sequence diagrams was put out by [18]. The method addresses interactions between actors and potential scenarios in a parallel, exclusive, or inclusive manner. To make sure that faults found in one interaction or scenario are not

connected to others that are occurring simultaneously or in parallel, it provides plugs utilising OCL restrictions. The goal is to offer a testing strategy that guarantees mistake independence. Only applying to holonic agents and the absence of empirical proof of the efficacy of the suggested strategy are drawbacks.

Dawood, Y. S., and Hashim, N. L. [19] evaluated the prior research on the creation of test cases using UML statechart diagrams. It investigates the domain's algorithms, coverage standards, and assessment techniques. The goal is to highlight discrepancies and give a better knowledge of the test case generating procedures. A content analysis of 24 books in the topic is part of the technique. The emphasis on UML statechart diagrams and the little treatment of prior works are limitations.

A method to produce test cases from software requirements at the use case description level was put out by Alrawashed, T. A., et al. [20]. The method fine-tunes use cases, transforms them into control flow diagrams, and then use a tool to produce test cases. Regression testing is made more effective while test complexity is reduced and test coverage is improved. Limitations include the absence of comparison with current methods and a thorough evaluation of the case study findings.

Using machine learning, Khalifa, E. M., et al. [21] demonstrated a method for creating test cases from use case diagrams. The technique tries to automate the difficult and laborious process of test case generation. In order to increase precision and effectiveness, it employs a metaheuristic method. Information extraction, preprocessing, the use of the metaheuristic approach, and performance evaluation make up the methodology. The application to certain software systems and the reliance on the calibre of use case diagrams are limitations.

A method for creating optimised test cases from UML sequence diagrams using the Firefly algorithm was put forth by Runal, G., et al. [22]. The process of creating test cases will be automated, and the quality and dependability of software systems will be increased. Model-based testing and the structural test selection concept are used in the technique. The emphasis on UML sequence diagrams and possible drawbacks in big and complicated systems are constraints.

A technique for automatically creating executable test scripts for an IoT system using UML state machine diagrams was provided by Swain, R., et al. [23]. The goal is to make testing simpler and require less manual labour. The process entails the creation of mappings between actions and assertions as well as between transition events and functions. The algorithm creates transition pathways together with the corresponding assertions and actions. The contribution is a unique strategy that is shown by a case study on a diabetic monitoring and control system. Limitations include the absence of transition guard evaluation and the potential for method improvement through symbolic evaluation.

ALI, H. M. B. M. [24] addressed looping and iteration issues in order to improve the production of test cases from UML sequence diagrams. The contribution is an enhanced method that gives software testers a quick and easy approach to create test cases. A case study is used in the approach to show how the suggested strategy is used and how it compares to other strategies. The objective is to address the problem of producing test cases from UML sequence diagrams. The emphasis on looping and iteration issues as well as the need for knowledge of UML sequence diagrams are limitations. In their discussion of the automation of software development processes in telecom carrier networks, Kikuma, K., et al. [25] emphasised cost savings while upholding dependability and safety. The method's creation, which uses mathematical principles to boost learning effectiveness in software testing, is the contribution. The methodology makes use of the preparation of test cases and the use of mathematical techniques by qualified engineers. The report emphasises the significance of using the preparation procedure to its full potential. Data creation from already-existing design documentation is one limitation.

A methodical process for creating test cases from a UML model, especially from UML Sequence Diagrams, was provided by Rocha, M., et al. [26]. The contributions include the usage of the ModelJUnit and JUnit libraries for automated test case generation as well as the design of transformation rules using ATL. The goal is to provide UML Sequence Diagrams a defined meaning and make them appropriate for automated testing. ModelJUnit and JUnit libraries are used in the process, which entails translating UML Sequence Diagrams into Extended Finite State Machines. The emphasis on UML Sequence Diagrams and the efficacy relying on the calibre of the UML model are limitations.

Tiwari, H. Swathi, B. [27] investigates soft computing approaches like genetic algorithms and artificial bee colonies while focusing on test case creation in software testing. The contributions also explore soft computing approaches, provide an assessment criterion, and examine the value of code coverage in addition to underlining

the significance of test case production. The methodology uses soft computing techniques to generate test cases and analyse code coverage to determine efficacy. To give a thorough grasp of test case generation and its importance is the goal. The emphasis on soft computing methods and the empirical study's constrained scope are both drawbacks.

Soni, D. Jain, P. [28] gave a survey on several methods for prioritising and generating test cases using UML diagrams. A thorough analysis of relevant work, a comparison of methods, and the identification of knowledge gaps are all included in the contributions. The process entails gathering data and comparing algorithms according to their efficacy. The goal is to outline test case creation techniques using UML diagrams and identify potential directions for further study. The emphasis on UML diagrams as well as the absence of a thorough analysis and useful implementation are limitations.

Using UML interaction diagrams, Minhas, N. M., et al. [29] developed a methodical mapping of test case generating approaches. The objective is to contrast various methods based on their strengths and weaknesses. The study evaluates the original studies' reporting quality and identifies any potential errors. An overview of the possibilities and constraints of test case generating methods utilising UML interaction diagrams is provided in this work. The report demonstrates that the examined studies lacked empirical evaluation in industry contexts and conformity to research principles. It implies that the industry needs stronger tool support for test case creation methods based on UML interaction diagrams.

Using class and activity diagrams, Mburu, J. M., et al. [30] suggested an improved multiview test case generating approach for object-oriented software. The creation and use of the MUTCAS Generator technology constitute the contribution. Utilising UML models, the technique creates test cases automatically that include both structural and behavioural aspects of the system. The aim is to allow early identification of software flaws and to solve the issues of time, money, and effort necessary for manual test case development. Limitations include the fact that they only apply to UML models, the possibility of problems in complicated systems, and reliance on the calibre of the UML models.

Dash and Panda, M. [31] An overview of the model-based and search-based testing strategies used to create test cases and test data for object-oriented programmes can be found in [31]. One aspect of the contribution is the framework for search-based testing that uses hybrid metaheuristics algorithms. The process includes a literature research to compile data on testing procedures and approaches. A framework for search-based testing of object-oriented programmes is proposed together with insights into the various testing methodologies. The suggested paradigm has limitations, such as the lack of a comparative analysis and empirical backing.

S. B. Tatale, V. C. Prakash, & Co. [32] concentrated on leveraging UML Sequence and Activity diagrams to automatically generate test cases. The contribution is a feasibility study on producing test cases focused on combinatorial logic using UML diagrams. The process entails employing dynamic slicing techniques and converting UML diagrams into tree or graph representations. The goal is to present a fresh method for creating test cases automatically from a system's design specification. The only emphasis on UML Sequence and Activity diagrams without taking into account other UML diagrams is one limitation.

K. Jin, K. Lano, and K. A systematic literature review (SLR) on creating test cases from UML diagrams was published in [33]. The contribution consists of identifying methods, results, research trends, gaps, and suggestions for future study. The approach entails running an SLR and using predetermined criteria to choose pertinent documents. The goal is to give a summary of the research on creating test cases from UML diagrams. The concentration on UML diagrams alone without taking other modelling languages into account is the restriction.

For the purpose of load testing mobile apps, Ali, A., et al. [34] suggested an autonomous model-based test case creation technique. An technique that minimises testing time while validating requirements, including both performance and functional elements, is the contribution. The process includes model-based testing, UML model creation, test case generation, and performance evaluation. The goal is to solve time-to-market restrictions and load testing of mobile applications. The absence of a thorough review and the choice of workload are limitations.

A technique for automatically creating test cases for flight control systems using UML state diagrams was developed by Fan, C., and Zou, P. [35]. A real-time extension strategy, a time domain equivalence partition method, and a feasibility check of the approach are the contributions. Modelling flight control systems, ex-

panding UML state diagrams, creating equivalence classes, creating test pathways, and automating test case creation are all part of the technique. The goal is to increase the efficacy and efficiency of flight control system testing. Only being applicable to UML state diagrams, having a cap on complexity, and not taking component interactions into account are some of the restrictions.

Using a Basic Genetic Algorithm (BGA), Sahoo, R. K., et al. [36] suggested a method for improving test data generated from UML activity and state chart diagrams. To produce test data that validates system requirements is the goal. The automated creation of specification-centered test data from UML models is the contribution. The process entails integrating diagrams into intermediate graphical representations, then optimising them using BGA. The paper examines findings and offers a case study. The application to particular diagram kinds, reliance on input quality and BGA settings, and absence of comparison analysis are some limitations.

J. Mburu, J. G. Ndia, & Co. [37] provided a comprehensive mapping research on approaches for creating and optimising test cases based on UML models. Finding trends and gaps in the study is the goal. A comprehensive literature review of publications published between 2010 and 2022 is part of the process. Contributions include an overview of trends and gaps, a list of often used search strategies, a need for greater study on combinational UML diagrams and optimisation, and a list of frequently used validation strategies. The focus on test case optimisation, automation, and combinational UML diagrams is restricted, and the assessment of research is also limited.

Prakash, V. C., Tatala, S. A method for automatically creating combinatorial test cases from UML Activity Diagrams was proposed in [38]. The key contribution is the creation of a programme that takes input parameters from UML Activity Diagrams and use the Particle Swarm Optimisation technique to produce the optimal number of test cases. To show how well the method works in practise, a case study of the Indian Railway Reservation System is presented. The technique uses the Particle Swarm Optimisation algorithm for test case generation, constraint detection, and parameter extraction. To reduce time and effort, test case generation is automated. Application to UML Activity Diagrams only, possible efficacy restrictions in large systems, and appropriateness for certain testing situations are some limits.

A technique for producing scenario-based test cases from UML-ADs (“UML activity diagrams”) was put out by Hettab, A., et al. in [39]. The primary contribution enables early testing in the software development life cycle by automatically generating test cases from UML-ADs. According to the technique, test scenarios are derived from EADG models by creating an EADG (“extended activity dependency graph”) from UML-ADs. By employing graphical simulation to apply the test scenarios to UML-ADs, testers may verify the test scenarios. Through mutation analysis, the method’s capacity to discover faults is assessed. Application to UML-ADs alone, possible drawbacks in big and complicated systems, and reliance on the correctness of UML-ADs are some constraints.

At the beginning of software development, Tamizharasi, A., and Ezhumalai, P. [40] presented a technique for producing optimised test data from UML models using the Hybrid GBCSA (“Genetic and Crow Search Algorithm”). The contribution consists of using GBCSA to streamline the test suite by eliminating unnecessary test data and focusing the search on global optima. Comparing experimental results to conventional crow search and genetic optimisation methods, they show 100% route coverage and time efficiency. To assess the suggested technique, UML models are used, along with experimentation. The goal is to show the method’s efficacy and tackle the problem of producing test data for intricate software systems. The absence of a thorough examination of constraints, comparison with alternative approaches, and scalability for big systems are among the drawbacks.

Based on various coverage requirements, Pradhan, S., et al. [41] suggested methods for creating test cases from a state chart diagram. By addressing an object’s states and transitions, the aim is to spot state-based defects. In order to build test cases based on various coverage requirements, the process entails converting the state chart diagram into a SCIG (“State Chart Intermediate Graph”). In this study, efficient state-based criterion algorithms like RTP (“Round Trip Path”) and ATP (“All Transition Pair”) are introduced. There includes discussion of the two case studies, stack operation and vending machine automation system. The algorithmic suggestions for creating test cases based on coverage requirements constitute the contribution. The restriction is ATP’s inability to provide complete transition coverage.

Hammad, M., and Hamza, Z. A. [42] gave a case study on the creation of test sequences, concentrating on

the usage of a previously suggested method based on use case model analysis. The process entails dissecting the UML use case diagram, transforming it into activity diagrams, simplification, and information extraction. The purpose of the study is to show how well this method works at producing test sequences for software testing. The method's shortcomings, however, include its restricted application to UML use case models, reliance on a single case study, and lack of a comparison with other methodologies already in use.

Sumalatha, V., & Raju, G. S. V. P. [43] used UML activity diagrams to solve the problem of test case creation in model-driven software development. For the purposes of test case creation, reduction, and prioritisation, the authors suggest using Evolutionary and Greedy Heuristic algorithms. The authors of the study compare their methodology to current practises and use UML activity diagrams in an effort to increase the efficacy of software testing. The examination of a single case study, the absence of in-depth comparisons with other methodologies, and algorithm constraints are some of the shortcomings, though.

An innovative method for creating and refining test cases from UML design diagrams was put forth by Khurana, N., et al. [44]. The SYTG ("system graph"), which is explored using a Genetic Algorithm, is created by combining use case, activity, and sequence diagrams. Its limitations include the fact that it only applies to UML design diagrams, how dependent it is on the quality of the input diagrams, and how well suited it is for big and complicated systems.

Using a hybrid bee colony approach, Sahoo, R. K., et al. [45] described a method for creating and refining test cases from combinational UML diagrams. The contribution is an automated test case generation method that aims to make software testing more effective and affordable. Although the technique is model-driven testing-based, it has several drawbacks, such as its emphasis on combinatorial UML diagrams and possible customisation needs for particular software systems.

Chandra and Meiliana, L. C. D. [46] A approach for automatically creating and improving test cases in software testing was put out by [46]. The goal is to use resources more efficiently, especially in the field of mobile technology. Utilising a genetic algorithm, the process entails creating test cases from combinational UML diagrams. Despite offering a beneficial technique, the study has certain drawbacks, such as a small population size and few genetic algorithm operators.

Tiwari, H. Swathi, B. [47] A method for creating test cases for web applications utilising input values and data dependencies was put forth in [47]. Pairwise testing, a genetic algorithm, and a system graph are all components of the process. The contribution of the research is the suggested approach for online applications, which addresses the challenging task of test case creation. The emphasis on functional testing and the absence of other testing kinds, such as security testing, are drawbacks, though.

S. S. Panigrahi. [48] suggested a technique for automatically creating test cases that makes use of a hybrid firefly algorithm and UML Activity diagrams. The strategy focuses on choosing the best test cases in terms of cost and coverage. The case study of an ATM withdrawal strategy is presented in the paper to illustrate the viability of the suggested approach. The objective is to lessen the amount of effort and time needed for software testing. However, drawbacks include the lack of a comparison with alternative approaches and the evaluation's constrained scope, which was only the ATM withdrawal system case study. The contribution. Methodology, merits and limits of these contemporary models have been listed in table 3.1 for quick view.

In table 3.1 there are some research gaps in the use of UML diagrams for test case generation, optimization, and prioritization across multiple domains. One recurring theme is that people tend to focus on specific UML diagrams, such as sequence and activity diagrams, rather than exploring other types of diagrams. An important issue is that there are few empirical studies comparing the proposed methods to existing methods. This lack of comparative data makes it more difficult to demonstrate the effectiveness of new methods and determine how they can be applied in various software development scenarios.

In future research, it may be beneficial to look beyond simple UML diagrams and include modeling languages that are more diverse and complex. This could result in deeper insights and stronger testing frameworks. Furthermore, future research should focus on developing real-world studies that not only compare different test case generation methods, but also assess their effectiveness and scalability. Filling in these gaps can aid in the development of more complex and useful test case strategies, resulting in improvements in automated testing that can keep up with changing software system requirements.

Table 3.1: Summary of contemporary models on test case generation, optimization, and prioritization.

Author	Contribution	Methodology	Merits	Limits
B. N. Biswal [6]	Technique for creating and refining test cases of object-oriented software	UML activity diagrams, sequence diagrams, class diagrams	Error minimization method for test case optimization	Application limited to particular categories of object-oriented software, reliance on the excellence of UML behavioural models
Meena, D.K. [7]	Technique for creating test cases using UML diagrams	Interaction Overview diagrams, Sequence diagrams	Message route coverage	Limited to object-oriented programmes, absence of actual data contrasting the suggested strategy with other ones
Jiang, L., and Li, Y. [8]	Technique for creating test cases using UML sequence diagrams	Examination of sequence diagram, message sequence assessment	Useful method for testing object-oriented applications	No prototype system for autonomous test case generation, no discussion of drawbacks or viability of the suggested approach in practical contexts
A. Herrmann, M. Felderer, and M. [9]	Instructions for methodically generating test cases from state machines and activity diagrams	Student experiment, controlled methodology	Taxonomy of faults, evidence of activity diagrams being more error-prone, error minimization	Utilization of student participation, emphasis on UML activity diagrams and state machines
Jena, A. K., et al. [10]	Approach for creating test cases using UML sequence diagrams and interaction overview diagrams	Combination of produced graphs from UML components, error identification	Early error detection, software quality improvement	Emphasis on a single case study, lack of comparison with alternative techniques for creating test cases
A. Herrmann, M. Felderer, and M. [11]	Controlled experiment examining mistakes in manual test case derivation	Student experiment, taxonomy of faults, evidence of activity diagrams being error-prone	Error minimization, recommendations for minimizing errors	Concentration on certain diagram kinds without comparison to other types or automated test case derivation techniques, relative smallness of sample size
Jagtap, S., et al. [12]	Technique for creating test cases using state chart diagrams and use case diagrams	Testing criteria addressing state and transition in the diagrams	Creation of efficient test cases during early phases of software development	Emphasis on use case diagrams and state chart diagrams in UML 2.0, potential need for manual testing to ensure thorough coverage
Yin, Y., et al. [13]	Automated test case creation method using unstructured SysML ADs	Conversion of ADs into Intermediate Black Box Model, test case generation	Efficient testing methodology for industry practitioners employing SysML Ads	Concentration on the AD model alone, without considering other SysML models
Salman, Y. D., et al. [14]	Efficient combination of coverage requirements in test case generation	Emphasis on UML state chart diagrams, handling loops	Increased efficiency in test data generation during software testing	Emphasis on UML state chart diagrams, requirement for context adaptation
Efendi, N. B. M., et al. [15]	Automated test case development using UML Statechart and Sequence diagrams	Utilization of UML diagrams for automatic test case creation	Cost reduction in software development	Emphasis on UML Statechart and Sequence diagrams at the expense of other
Prakash, V. C., et al. [16]	Tool for automated test case development using Particle Swarm Optimization (PSO)	Comprehensive assessment of PSO and its variations	Reduction of test suite sizes, boost in program dependability	Emphasis on PSO and its comparison to other optimization methods

Continued on next page...

Table 3.1 – continued from previous page.

Author	Contribution	Methodology	Merits	Limits
Kamonsantiroj, S., et al. [17]	Memorizing strategy for producing test cases in concurrent UML activity diagrams	Dynamic programming technique, tester specification	Reduction of explosion of test cases in concurrent systems	Application limited to concurrent UML activity diagrams, emphasis on concurrency testing, limitations on other diagram types or concurrent systems
Mokhati, F.; Dehimi, N. E. H. [18]	Evaluation method for multi-agent systems using AUML sequence diagrams	Parallel, exclusive, or inclusive interactions, OCL restrictions	Testing strategy ensuring mistake independence	Only applies to holonic agents, absence of empirical proof of the efficacy of the suggested strategy
Dawood, Y. S., and Hashim, N. L. [19]	Evaluation of test case creation using UML state-chart diagrams	Content analysis of 24 books, investigation of algorithms and coverage	Identification of discrepancies, improved understanding of test case generation procedures	Emphasis on UML statechart diagrams, limited treatment of prior works
Alrawashed, T. A., et al. [20]	Test case generation from software requirements at the use case description level	Fine-tuning of use cases, transformation into control flow diagrams	More effective regression testing, reduced test complexity and improved coverage	Absence of comparison with current methods, lack of thorough evaluation of case study findings
Khalifa, E. M., et al. [21]	Test case generation from use case diagrams using machine learning	Metaheuristic approach, information extraction, preprocessing	Automation of laborious test case generation process, increased precision and effectiveness	Application limited to certain software systems, reliance on the quality of use case diagrams
Runal, G., et al. [22]	Creation of optimized test cases from UML sequence diagrams using the Firefly algorithm	Model-based testing, structural test selection concept	Automation of test case creation, improved quality and dependability of software systems	Emphasis on UML sequence diagrams, potential limitations in big and complicated systems
Swain, R., et al. [23]	Automated creation of executable test scripts for an IoT system using UML state machine diagrams	Creation of mappings, algorithm for transition pathway creation	Simplification of testing process, reduced manual labor	Absence of transition guard evaluation, potential for method improvement through symbolic evaluation
J. Cvetkovi and M. Cvetkovi. [3]	Creation of test cases with UML diagrams focusing on modelling depression caused by internet addiction	Categorization of test case creation, utilization of various UML diagrams	Advancement in software testing, fresh method for creating test cases	Specificity of the case study, absence of comparison with other approaches
ALI, H. M. B. M. [24]	Improvement of test case generation from UML sequence diagrams	Enhanced method for quick and easy test case creation	Quick and easy approach to create test cases	Emphasis on looping and iteration issues, need for knowledge of UML sequence diagrams
Kikuma, K., et al. [25]	Automation of software development processes in telecom carrier networks	Mathematical principles, preparation of test cases, use of mathematics	Cost savings, dependability, safety	Data creation limited to already-existing design documentation
Rocha, M., et al. [26]	Methodical process for creating test cases from UML model, specifically UML Sequence Diagrams	ModelJUnit, JUnit libraries, transformation rules	Defined meaning for UML Sequence Diagrams, automation of testing process	Emphasis on UML Sequence Diagrams, efficacy relies on the calibre of the UML model
Tiwari, H. Swathi, B. [27]	Soft computing approaches for test case creation in software testing	Genetic algorithms, artificial bee colonies	Assessment criterion, exploration of soft computing approaches	Emphasis on soft computing methods, constrained scope of empirical study

Continued on next page...

Table 3.1 – continued from previous page.

Author	Contribution	Methodology	Merits	Limits
Soni, D. Jain, P. [28]	Survey on methods for prioritizing and generating test cases using UML diagrams	Analysis of relevant work, comparison of methods	Identification of knowledge gaps, overview of test case creation techniques using UML diagrams	Emphasis on UML diagrams, lack of thorough analysis and useful implementation
Minhas, N. M., et al. [29]	Mapping of test case generating approaches using UML interaction diagrams	Evaluation of reporting quality, identification of errors	Comparison of methods based on strengths and weaknesses	Lack of empirical evaluation in industry contexts, lack of conformity to research principles
Mburu, J. M., et al. [30]	Improved multiview test case generating approach for object-oriented software using UML diagrams	MUTCASGenerator technology	Automation of test case creation, inclusion of structural and behavioral aspects of the system	Limited to UML models, potential issues in complicated systems, reliance on the calibre of UML models
Dash and Panda, M. [31]	Overview of model-based and search-based testing strategies for object-oriented programs	Framework for search-based testing, hybrid metaheuristics algorithms	Compilation of data on testing procedures and approaches, insights into various testing methodologies	Lack of comparative analysis, lack of empirical backing
S. B. Tatale, V. C. Prakash, & Co. [32]	Feasibility study on producing test cases focused on combinatorial logic using UML diagrams	Dynamic slicing techniques, conversion of UML diagrams	Automatic generation of test cases from UML Sequence and Activity diagrams	Emphasis only on UML Sequence and Activity diagrams, lack of consideration for other UML diagrams
K. Jin, K. Lano, and K. [33]	Systematic literature review on creating test cases from UML diagrams	Systematic literature review	Identification of methods, results, research trends, gaps, and suggestions for future study	Focus only on UML diagrams without considering other modelling languages
Ali, A., et al. [34]	Autonomous model-based test case creation technique for load testing mobile apps	Model-based testing, UML model creation, performance evaluation	Minimization of testing time, validation of requirements for both performance and functional elements	Lack of thorough review, choice of workload
Fan, C., and Zou, P. [35]	Technique for creating test cases for flight control systems using UML state diagrams	Real-time extension strategy, time domain equivalence partition method	Inclusion of real-time aspects, feasibility check of the approach	Applicable only to UML state diagrams, potential issues in complicated systems, no consideration of component interactions
Sahoo, R. K., et al. [36]	Method for improving test data generated from UML activity and state chart diagrams	BGA (“Basic Genetic Algorithm”)	Test data generation that validates system requirements	Application to particular diagram kinds, reliance on input quality and BGA settings, absence of comparison analysis
J. Mburu, J. G. Ndia, & Co. [37]	Mapping research on approaches for creating and optimizing test cases based on UML models	Comprehensive literature review	Overview of trends and gaps, identification of frequently used strategies	Focus on test case optimization, automation, and combinational UML diagrams, limited assessment of research
Prakash, V. C., Tatale, S. [38]	Automatic creation of combinatorial test cases from UML Activity Diagrams	PSO (“Particle Swarm Optimization”) technique, case study	Quick and efficient test case generation from UML Activity Diagrams	Application limited to UML Activity Diagrams, possible efficacy restrictions in large systems, appropriateness for certain testing situations

Continued on next page...

Table 3.1 – continued from previous page.

Author	Contribution	Methodology	Merits	Limits
Hettab, A., et al. [39]	Technique for producing scenario-based test cases from UML activity diagrams	EADG, graphical simulation	Early testing in software development, automatic generation of test cases from UML activity diagrams	Application limited to UML activity diagrams, possible drawbacks in big and complicated systems, reliance on the correctness of UML activity diagrams
Tamizharasi, A., and Ezhumalai, P. [40]	Producing optimised test data from UML models using the Hybrid (“GBCSA”)	Hybrid GBCSA (“Genetic and Crow Search Algorithm”)	Streamlining the test suite, elimination of unnecessary test data	Absence of thorough examination of constraints, comparison with alternative approaches, scalability for big systems
Pradhan, S., et al. [41]	Methods for creating test cases from a state chart diagram	Conversion of state chart diagram into SCIG (“State Chart Intermediate Graph”)	Spotting state-based defects, algorithmic suggestions for test case creation	Inability of ATP to provide complete transition coverage
Hammad, M., and Hamza, Z. A. [42]	Case study on the creation of test sequences using use case model analysis	UML use case diagram analysis, transformation into activity diagrams	Production of test sequences for software testing	Restricted application to UML use case models, reliance on a single case study, lack of comparison with other methodologies
Sumalatha, V., & Raju, G. S. V. P. [43]	Test case creation, reduction, and prioritisation using UML activity diagrams	Evolutionary and Greedy Heuristic algorithms	Increased efficacy of software testing using UML activity diagrams	Examination limited to a single case study, absence of in-depth comparisons with other methodologies, algorithm constraints
Khurana, N., et al. [44]	Creating and refining test cases from UML design diagrams	Genetic Algorithm, creation of (“SYTG”)	Combination of different UML design diagrams, refinement of test cases	Application limited to UML design diagrams, dependence on input diagram quality, suitability for big and complex systems
Sahoo, R. K., et al. [45]	Creating and refining test cases from combinational UML diagrams	Hybrid bee colony approach, model-driven testing	Automated test case generation, improved software testing	Emphasis on combinational UML diagrams, customization needs for specific software systems
Chandra and Meiliana, L. C. D. [46]	Automatic creation and improvement of test cases in software testing	Genetic Algorithm, creation of test cases from combinational UML diagrams	Efficient resource utilization, application in the mobile technology field	Small population size, limited genetic algorithm operators
Tiwari, H. Swathi, B. [47]	Creation of test cases for web applications using input values and data dependencies	Pairwise testing, genetic algorithm, system graph	Approach for online applications, addressing the challenge of test case creation	Emphasis on functional testing, absence of other testing types
S. S. Panigrahi [48]	Automatic creation of test cases using a hybrid firefly algorithm and UML Activity diagrams	Hybrid firefly algorithm, UML Activity diagrams	Reduction in effort and time for software testing	Lack of comparison with alternative approaches, evaluation limited to a specific case study

3.2. Test Case Prioritization. Regression testing test case prioritisation method employing sequence diagrams and labelled transition systems was suggested by As’ Sahra, N. F., & Komputeran, F. [49]. The method use Bayesian Networks to incorporate source code modifications, software fault-proneness, as well as test coverage data into a single model. However, the research makes an assumption about test case independence that could not hold true in actual circumstances.

A novel method of test case prioritisation for model-based mutation testing in the automotive sector was introduced by Shin, K. W., and Lim, D. J. [50]. They use the UML statechart to create a software model, use

mutation operators to generate mutations, and suggest a TCP technique based on the ("alternating variable method") AVM. The study covers actual research in the automobile sector, however its application outside the sector and the quantity of investigated problems are also limits.

The use of machine learning approaches in "test case prioritization" (TCP) for regression testing was examined by Meçe, E. K., Paci, and Binjaku [51]. They review current research that employs machine learning in TCP and provide details on methods, measurements, data, benefits, and drawbacks. There are drawbacks, such as the potential rejection of important test cases via selection approaches and the requirement for a significant quantity of data for efficient machine learning methods.

An autonomous test route generating method and prioritisation model for software testing were suggested by Fan, L., Wang, Y., and Liu [52]. They build a priority model to order the test pathways, design the activity flow graph, and specify the mapping rules from UML activity diagram to the graph. The article does not, however, compare new algorithms to old ones or evaluate large-scale software systems. It also presupposes that UML activity diagrams are accessible.

For user interface testing, Nguyen, V., & Le, B. [53] introduced a unique test prioritisation technique called RLTCP. The technique uses the coverage graph and reinforcement learning (RL) to prioritise test cases. With an experimental assessment contrasting it with other approaches, the research builds on a past work on prioritising UI automated test cases using RL. The drawback is that it ignores alternative testing methods in favour of a narrow emphasis on user interface testing.

A test route prioritisation approach based on the testers' interests, as well as the altered area in UML activity diagrams, was proposed by Sornkliang, W., and Phetkaew, T. [54]. They use various techniques to give weights to symbols and then order test pathways accordingly. The technique tries to aid testers in swiftly identifying critical issues. The dependence on testers' preferences and the work involved in weight assignment are limitations.

The contribution of the Methodology, merits and limits of these contemporary models have been listed as a table for quick view table 3.2.

Table 3.2 a thorough examination of the most recent models for test case prioritization, several research gaps were identified that could be addressed in future studies. Sahra, N. F., and Komputeran, F. [49] use Bayesian Networks for prioritization, assuming test case independence, which may not be true in practice. Future research could look into models that account for interdependence among test cases. Shin, K. W., and Lim, D. J. [50] use model-based mutation testing exclusively in the automotive industry. This demonstrates the need for mutation testing to be applied in more areas and studied in greater depth. The machine learning approach proposed by Meçe, E. K., Paci, and Binjaku [51] shows promise in terms of efficiency, but it may miss important test cases and need a large amount of data. This suggests that we need more reliable machine learning models that do not require as much data.

Fan, L., Wang, Y., and Liu [52] propose an autonomous test route generation method that needs to be compared to traditional algorithms and tested in large-scale systems. This opens the door for future research to confirm its usefulness and potential for expansion. The RL-based method developed by Nguyen, V., and Le, B. [53] for testing user interfaces is novel and intriguing, but it only tests a few things. This demonstrates that reinforcement learning could be applied in a broader range of testing scenarios. Finally, Sornkliang, W., and Phetkaew, T. [54] interest-based route prioritization heavily relies on tester input to assign weights. This demonstrates the need for more automated, objective prioritization frameworks with less human bias and effort. Filling these identified gaps would significantly improve the development of test case prioritization methods, making regression testing more useful.

4. Observations. The review of the articles [6–15] highlights various techniques and approaches for automated test case generation using UML behavioral models and diagrams in object-oriented software development. Each article focuses on specific UML models or diagrams and has some limitations. The most effective and efficient automated test case generation method that makes use of UML behavioral models and diagrams needs to be determined in order to address these limitations and provide guidance to practitioners. The identification of best practices for automated test case generation in object-oriented software development will be made possible by this research question, which will also allow for a thorough examination of current approaches, comparison of their effectiveness and efficiency.

Table 3.2: Summary of contemporary models on test case prioritization

Author	Contribution	Methodology	Merits	Limits
As' Sahra, N. F., & Komputeran, F. [49]	Regression testing test case prioritisation method employing sequence diagrams and labelled transition systems	Bayesian Networks	Incorporation of source code modifications, fault-proneness, and test coverage	Assumption of test case independence that may not hold true in actual circumstances
Shin, K. W., and Lim, D. J. [50]	Test case prioritisation for model-based mutation testing in the automotive sector	UML statechart, mutation operators, TCP technique based on AVM	Application in the automotive sector, software model creation	Limited application outside the automotive sector, limited investigation of problems
Meçe, E. K., Paci, and Binjaku [51]	Use of machine learning approaches in test case prioritization for regression testing	Review of existing research, analysis of methods, measurements, data	Potential for efficient machine learning methods, insights into benefits	Potential rejection of important test cases, requirement for a significant amount of data
Fan, L., Wang, Y., and Liu [52]	Autonomous test route generating method and prioritisation model for software testing	Priority model, activity flow graph, mapping rules from UML activity diagram	Autonomous test route generation, prioritisation of test pathways	Lack of comparison with old algorithms, lack of evaluation on large-scale software systems, assumption of accessibility of UML activity diagrams
Nguyen, V., & Le, B. [53]	Test prioritisation technique called RLTCP for user interface testing	Coverage graph, RL ("reinforcement learning")	Prioritisation of test cases in user interface testing	Narrow emphasis on user interface testing, ignoring alternative testing methods
Sornklang, W., and Phetkaew, T. [54]	Test route prioritisation approach based on testers' interests and altered area in UML activity diagrams	Weight assignment techniques, ordering of test pathways based on weights	Aid in quickly identifying critical issues, prioritisation based on interests	Dependence on testers' preferences, effort required for weight assignment

The review of the articles [16–24] highlights different approaches for automated test case generation that use optimization techniques and UML diagrams. The article focuses on specific optimization techniques or UML diagram types and has some limitations. In order to address these limitations and offer guidance for practitioners, identify the most effective and efficient approaches for automated test case generation that combine optimization techniques with different kinds of UML diagrams. This research question will enable a thorough examination of current approaches, comparison of their effectiveness, and identification of best practices for automated test case generation in software development.

The review of the articles [25–32], [1], and [5] highlights different approaches and methodologies for automated test case generation using UML diagrams. Each article focuses on specific optimization techniques, soft computing, or UML diagrams and has its own limitations. In order to address these limitations and offer thorough recommendations for software development, it is important to investigate how various methodologies and techniques can be effectively combined to improve test case generation using UML diagrams. In order to enhance the effectiveness and efficiency of test case generation and software development, this research question will enable an exploration of optimization methods, soft computing techniques, and automation approaches.

Reasoning the review of the articles [33–41] reveals limitations and opportunities for automated test case generation using UML diagrams. Limitations include a lack of comparative analysis or thorough evaluation, a focus on specific diagram types, and applicability to complex systems. To overcome these limitations and enhance the effectiveness of test case generation, look into how automated techniques can be improved to address the identified challenges. This research question will examine the exploration of factors such as model-based testing, optimization techniques, and validation approaches to enhance the applicability, efficiency, and effectiveness of automated test case generation using UML diagrams. By addressing these limitations, software developers can enhance test case generation to support more complex systems, stronger comparative analysis, and evaluation.

The review of the articles [42–46] highlights common limitations in current approaches, including a focus on specific diagram types, reliance on input diagram quality, customization demands, and limitations in algorithm design. The applicability and effectiveness of the suggested methods are restricted by these limitations. More research is needed to address these limitations, create better approaches that can be used more broadly, improve software testing effectiveness, and maximize test case generation and prioritization. By looking into and addressing these limitations, researchers can promote model-driven software development and test methodologies.

The review of the papers [47–54] highlights several limitations in the current test case prioritization techniques. These limitations include limited applicability outside of case domains, omission of crucial test cases, lack of comparison with current algorithms, reliance on a specific type of testing, and reliance on testers' interests. It is important to investigate and suggest improvements to current techniques in order to address these limitations and enhance the effectiveness and efficiency of test case prioritization. Considering the limitations of the reviewed papers, this research question allows for the exploration of potential solutions and advancements in test case prioritization.

5. Implications and possible future research. Software testing researchers and practitioners can benefit from a thorough literature review on the topic of creating test cases from UML diagrams in a number of ways. These ramifications point to potential directions for future investigation.

1. **Research Gaps and Deficiencies:** The review points out research gaps and deficiencies in the literature. By highlighting areas that haven't been addressed, these gaps can guide future research efforts. Filling these gaps can be the focus of empirical studies, novel methodologies, and alternative approaches to test case generation from UML diagrams.
2. **Comparative Evaluation:** The review demonstrates the dearth of empirical support and thorough comparative evaluations in many studies. Future research may highlight the necessity of conducting thorough comparative analyses to assess the effectiveness, efficiency, and applicability of different test case generation techniques. Comparative studies assist researchers and practitioners in selecting the most appropriate methodology for their specific requirements.
3. **Generalizability and Applicability:** Identify and discuss the applicability and generalizability limitations of the suggested approaches. Research should focus on creating techniques that can handle various types of UML diagrams rather than just specific diagram types like activity and design diagrams. In different software development contexts, look into ways to scale and make the approaches effective.
4. **Algorithmic Improvements:** Deal with the algorithmic elements' limitations of the suggested approaches. Future research should improve the methods' heuristic and evolutionary algorithms. Investigate advanced optimization techniques, optimize algorithm parameters, or develop hybrid approaches to improve the effectiveness and efficiency of test case generation and optimization.
5. **Quality Assurance of UML Diagrams:** Examine techniques to enhance the accuracy and coherence of UML diagrams used for test case generation and optimization. To ensure that diagrams are accurate and complete, develop automated validation and verification techniques. Investigate techniques to improve UML diagram clarity and interpretability to facilitate test case generation, as well as approaches to handle incomplete or inconsistent diagrams.
6. **Integration of Testing Types:** Go beyond functional tests and incorporate other test types using the recommended approaches. Security, performance, usability, and other testing types should be considered in test case generation and optimization. This ensures that various quality-related issues are addressed during software testing.
7. **Automation and Tool Support:** The review highlights the importance of automation and tool support in test case generation from UML diagrams. Future research can focus on automated frameworks and tools that provide effective test case generation and integrate with UML modeling tools. These tools can facilitate UML-based test practices, reduce manual labor, and enhance productivity.
8. **Scalability and Complexity:** For large and complex software systems, many studies overlook the scalability of test case generation techniques. Future research can address these issues because modern software systems are larger and more complex than ever. In order to address scalability concerns, system complexity, and resource efficiency during test case generation, this includes looking into techniques.

9. **Criteria for Coverage and Optimization:** Review highlights the importance of coverage criteria and optimization techniques in test case generation: In order to produce optimized test cases with high coverage and low redundancy, future research can create sophisticated optimization algorithms, hybrid approaches, and intelligent techniques. Another important area of research is the investigation of novel coverage criteria and the assessment of their effectiveness in detecting different types of faults.
10. **Other Languages for Modeling and Integration:** While other software engineering modeling languages are explored, UML diagrams are the main focus of the review. Future research can examine test case generation techniques for these alternative modeling languages in an effort to enhance the effectiveness and efficiency of test case modeling.
11. **Real-world Evaluation:** Several reviewed studies lacked adequate validation and assessment. In real-world software development projects, test case generation techniques can be applied and evaluated as a research topic. The proposed methods may be assessed for effectiveness and applicability by incorporating realistic software systems, industry partnerships, and case studies from diverse fields.

5.1. Test case Generation and Optimization.

- Utilizing evolutionary and heuristic algorithms, along with UML diagrams, can improve test case generation and prioritization.
- Comprehensive models, like System Graphs, offer a holistic approach to test case generation and optimization.
- Hybrid optimization algorithms and model-driven testing methodologies show promise for automated and cost-effective software testing.
- Evaluating and comparing the performance of different algorithms and approaches for test case generation and prioritization.
- Investigating the applicability of proposed techniques to various types of UML diagrams and software systems.
- Addressing limitations related to input diagram quality, scalability, customization requirements, and the need for flexible solutions.
- Assessing the effectiveness and efficiency of hybrid optimization algorithms in real-world scenarios and comparing them to alternative approaches.
- Exploring the impact of population size, operators, and optimization techniques on test case generation and optimization.
- Extending solutions to address various testing scenarios, assessing their effectiveness across various applications, and managing the growing complexity of software systems.
- For test case prioritization techniques to be effective in practical settings, test case dependencies and relationships must be taken into account.
- Integrating multiple factors, such as source code changes, fault-proneness, and test coverage data, into unified models shows potential for enhancing test case prioritization.
- Machine learning techniques offer improved efficiency but require careful consideration of data availability and the potential exclusion of critical test cases.
- To develop test case prioritization techniques that are capable of handling dependencies and inter-test case relationships, more research is needed.
- Investigate ways to incorporate metrics and factors into prioritization models to increase their thoroughness and accuracy.
- Look into alternative data sources that can offer trustworthy information for prioritization based on machine learning, or investigate ways to lessen the reliance on large amounts of data.
- Evaluate the suggested techniques' generalizability and scalability on complex software systems.
- To develop versatile techniques, broaden your research beyond particular domains or testing types.
- Consider ways to enhance or automate weight assignment for more effective test case prioritization. By addressing these implications and exploring the suggested future research directions, researchers can advance the field of test case generation from UML diagrams, improve the efficiency and effectiveness of software testing practices, and contribute to the development of reliable and high-quality software systems.

6. Conclusion. The current state of research in this field is clarified by the systematic literature review on creating, optimizing, and prioritizing test cases from UML diagrams. Along with implications and future research directions, we review research trends, approaches, limitations, and gaps. Several important areas for additional research are highlighted in the review. The development of automated tools and frameworks for seamless integration with UML modeling tools, scalability and complexity challenges in large and complex software systems, advanced optimization algorithms and coverage criteria to generate optimized test cases, and investigation of test case generation techniques for alternative modeling languages are a few examples. The field of test case generation from UML diagrams can progress by addressing these implications and following recommended future case research directions. The effectiveness and efficiency of software test case generation techniques can be improved, and researchers can offer testers useful solutions. To aid in the development of dependable and superior software systems, they can decrease manual labor, increase automation, and enhance test procedures. The researchers and practitioners in the field of software testing gain from this systematic review of the literature. It discusses current methods, makes recommendations for advancements, and establishes the foundation for future research. The software development industry will benefit from researchers' use of the insights from this review to enhance test case generation from UML diagrams.

REFERENCES

- [1] R. G. TIWARI, ET AL., *Exploiting UML Diagrams for Test Case Generation: A Review*, 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM), IEEE, 2021, pp. 457–460.
- [2] C. MINGSONG, Q. XIAOKANG, AND L. XUANDONG, *Automatic test case generation for UML activity diagrams*, Proceedings of the 2006 International Workshop on Automation of Software Test, 2006, pp. 2–8.
- [3] J. CVETKOVIĆ AND M. CVETKOVIĆ, *Evaluation of UML Diagrams for Test Cases Generation: Case Study on Depression of Internet Addiction*, Physica A: Statistical Mechanics and Its Applications, vol. 525, 2019, pp. 1351–1359.
- [4] K. JIN AND K. LANO, *Generation of test cases from UML diagrams—A systematic literature review*, 14th Innovations in Software Engineering Conference (Formerly Known as India Software Engineering Conference), 2021, pp. 1–10.
- [5] M. ROCHA, ET AL., *Model-Based Test Case Generation from UML Sequence Diagrams Using Extended Finite State Machines*, Software Quality Journal, vol. 29, no. 3, 2021, pp. 597–627.
- [6] B. N. BISWAL, *Test case generation and optimization of object-oriented software using UML behavioral models*, Diss. 2010.
- [7] D. K. MEENA, *Test Case Generation From UML Interaction Overview Diagram and Sequence Diagram*, Diss. 2013.
- [8] Y. LI AND L. JIANG, *The Research on Test Case Generation Technology of UML Sequence Diagram*, 2014 9th International Conference on Computer Science & Education, IEEE, 2014, pp. 1067–1069.
- [9] M. FELDERER AND A. HERRMANN, *Manual Test Case Derivation from UML Activity Diagrams and State Machines: A Controlled Experiment*, Information and Software Technology, vol. 61, 2015, pp. 1–15.
- [10] A. K. JENA, S. K. SWAIN, AND D. P. MOHAPATRA, *Model based test case generation from UML sequence and interaction overview diagrams*, Computational Intelligence in Data Mining-Volume 2: Proceedings of the International Conference on CIDM, 20–21 December 2014. Springer India, 2015, pp. 247–257.
- [11] M. FELDERER AND A. HERRMANN, *Comprehensibility of System Models during Test Design: A Controlled Experiment Comparing UML Activity Diagrams and State Machines*, Software Quality Journal, vol. 27, no. 1, 2019, pp. 125–147.
- [12] S. JAGTAP, ET AL., *Generate Test Cases From UML Use Case and State Chart Diagrams*, International Research Journal of Engineering and Technology (IRJET), vol. 3, no. 10, 2016, pp. 873–881.
- [13] Y. YIN, *An Automated Test Case Generation Approach Based on Activity Diagrams of SysML*, International Journal of Performability Engineering, 2017.
- [14] Y. D. SALMAN, ET AL., *Coverage Criteria for Test Case Generation Using UML State Chart Diagram*, p. 020125. DOI.org (Crossref), <https://doi.org/10.1063/1.5005458>.
- [15] N. B. M. EFENDI AND H. ASMUNI, *Exhaustive Search for Test Case Generation from UML Sequence Diagram and Statechart Diagram*, UTM Computing Proceedings Innovation in Computing Technology and Applications, Vol.2, 2018.
- [16] V. C. PRAKASH, ET AL., *A Critical Review on Automated Test Case Generation for Conducting Combinatorial Testing Using Particle Swarm Optimization*, International Journal of Engineering & Technology, vol. 7, no. 3.8, 2018, p. 22.
- [17] S. KAMONSANTIROJ, ET AL., *A Memorization Approach for Test Case Generation in Concurrent UML Activity Diagram*, Proceedings of the 2019 2nd International Conference on Geoinformatics and Data Analysis, ACM, 2019, pp. 20–25.
- [18] N. E. H. DEHIMI AND F. MOKHATI, *A Novel Test Case Generation Approach Based on A UML Sequence Diagram*, 2019 International Conference on Networking and Advanced Systems (ICNAS), IEEE, 2019, pp. 1–4.
- [19] N. L. HASHIM AND Y. S. DAWOOD, *A Review on Test Case Generation Methods Using UML Statechart*, 2019 4th International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE), IEEE, 2019, pp. 1–5.
- [20] T. A. ALRAWASHED, ET AL., *An Automated Approach to Generate Test Cases From Use Case Description Model*, Computer Modeling in Engineering & Sciences, vol. 119, no. 3, 2019, pp. 409–425.

- [21] E. M. KHALIFA, D. JAWAWI, AND H. A. JAMIL, *An efficient method to generate test cases from UML-use case diagram*, International Journal of Engineering Research and Technology, vol. 12, no. 7, 2019, pp. 1138–1145.
- [22] G. RUNAL, ET AL., *FUNCTIONAL TEST CASE GENERATION AND REDUNDANCY REMOVAL BASED ON MODEL DRIVEN TESTING USING UML ACTIVITY DIAGRAM*, International Journal of Mechanical Engineering and Technology (IJMET), vol. 10, no. 05, May 2019, pp. 318–324.
- [23] R. SWAIN, ET AL., *Automatic Test Case Generation From UML State Chart Diagram*, International Journal of Computer Applications, vol. 42, no. 7, Mar. 2012, pp. 26–36.
- [24] H. M. B. M. ALI, *MODEL-BASED SEMI-AUTOMATED TEST CASE GENERATION APPROACH USING UML DIAGRAMS*, 2019.
- [25] K. KIKUMA, ET AL., *Preparation Method in Automated Test Case Generation Using Machine Learning*, Proceedings of the Tenth International Symposium on Information and Communication Technology - SoICT 2019, ACM Press, 2019, pp. 393–398.
- [26] M. ROCHA, ET AL., *Test Case Generation by EFSM Extracted from UML Sequence Diagrams*, 2019, pp. 135–140.
- [27] B. SWATHI AND H. TIWARI, *Test Case Generation Process Using Soft Computing Techniques*, International Journal of Innovative Technology and Exploring Engineering, vol. 9, no. 1, Nov. 2019, pp. 4824–4831.
- [28] P. JAIN AND D. SONI, *A Survey on Generation of Test Cases Using UML Diagrams*, 2020 International Conference on Emerging Trends in Information Technology and Engineering (Ic-ETITE), IEEE, 2020, pp. 1–6.
- [29] N. M. MINHAS, ET AL., *A Systematic Mapping of Test Case Generation Techniques Using UML Interaction Diagrams*, Journal of Software: Evolution and Process, vol. 32, no. 6, June 2020.
- [30] J. M. MBURU, G. M. MUKETHA, AND A. M. OIRERE, *An Enhanced Multiview Test Case Generation Technique for Object-Oriented Software Using Class and Activity Diagrams*, International Journal of Recent Technology and Engineering (IJRTE), vol. 9, no. 4, Nov. 2020, pp. 185–196.
- [31] M. PANDA AND S. DASH, *Test-case generation for model-based testing of object-oriented programs*, Automated Software Testing: Foundations, Applications and Challenges, 15, 2020, pp. 53–77.
- [32] S. B. TATALE AND V. C. PRAKASH, *A Survey on Test Case Generation using UML Diagrams and Feasibility Study to Generate Combinatorial Logic Oriented Test Cases*, International Journal of Next-Generation Computing, vol. 12, no. 2, 2021, pp. 254–269.
- [33] K. JIN AND K. LANO, *Generation of Test Cases from UML Diagrams - A Systematic Literature Review*, 14th Innovations in Software Engineering Conference (Formerly Known as India Software Engineering Conference), ACM, 2021, pp. 1–10.
- [34] A. ALI, ET AL., *Model-Based Test Case Generation Approach for Mobile Applications Load Testing Using OCL Enhanced Activity Diagrams*, 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS), IEEE, 2021, pp. 493–499.
- [35] C. FAN AND P. ZOU, *Research on Automatic Test Case Generation Method of Flight Control System Based on UML State Diagram*, Journal of Physics: Conference Series, vol. 1961, no. 1, July 2021, p. 012019.
- [36] R. K. SAHOO, ET AL., *Test Case Generation from UML-Diagrams Using Genetic Algorithm*, Computers, Materials & Continua, vol. 67, no. 2, 2021, pp. 2321–2336.
- [37] J. M. MBURU AND J. G. NDIA, *A Systematic Mapping Study on UML Model Based Test Case Generation and Optimization Techniques*, International Journal of Computer Applications, vol. 184, no. 13, May 2022, pp. 26–33.
- [38] S. TATALE AND V. C. PRAKASH, *Automatic Generation and Optimization of Combinatorial Test Cases from UML Activity Diagram Using Particle Swarm Optimization*, Ingénierie Des Systèmes d’Information, vol. 27, no. 1, Feb. 2022, pp. 49–59.
- [39] A. HETTAB, ET AL., *Automatic Scenario-Oriented Test Case Generation from UML Activity Diagrams: A Graph Transformation and Simulation Approach*, International Journal of Computer Aided Engineering and Technology, vol. 16, no. 3, 2022, p. 379.
- [40] A. TAMIZHARASI AND P. EZHUMALAI, *Genetic-based Crow Search Algorithm for Test Case Generation*, International Transaction Journal of Engineering, Management, & Applied Sciences & Technologies, vol. 13, no. 4, 2022, pp. 1–11.
- [41] S. PRADHAN, ET AL., *Transition Coverage Based Test Case Generation from State Chart Diagram*, Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 3, Mar. 2022, pp. 993–1002.
- [42] Z. A. HAMZA AND M. HAMMAD, *Generating Test Sequences from UML Use Case Diagram: A Case Study*, 2020 Second International Sustainability and Resilience Conference: Technology and Innovation in Building Designs, IEEE, 2020, pp. 1–6.
- [43] V. SUMALATHA AND G. S. V. P. RAJU, *Model-based test case optimization of UML activity diagrams using evolutionary algorithms*, Int J Comput Sci Mob Appl, vol. 2, no. 11, 2014, pp. 131–142.
- [44] N. KHURANA, R. S. CHHILLAR, AND U. CHHILLAR, *A Novel Technique for Generation and Optimization of Test Cases Using Use Case, Sequence, Activity Diagram and Genetic Algorithm*, Journal of Software, vol. 11, no. 3, 2016, pp. 242–250.
- [45] R. K. SAHOO, ET AL., *Model Driven Test Case Optimization of UML Combinational Diagrams Using Hybrid Bee Colony Algorithm*, International Journal of Intelligent Systems and Applications, vol. 9, no. 6, June 2017, pp. 43–54.
- [46] L. C. D. MEILIANA AND A. CHANDRA, *Optimization of test case generation from uml Activity diagram and sequence diagram By using genetic algorithm*, vol. 13, no. 07, 2019, pp. 585.
- [47] B. SWATHI AND H. TIWARI, *Integrated Pairwise Testing Based Genetic Algorithm for Test Optimization*, International Journal of Advanced Computer Science and Applications, vol. 12, no. 4, 2021.
- [48] S. S. PANIGRAHI, ET AL., *Model-Driven Automatic Paths Generation and Test Case Optimization Using Hybrid FA-BC*, 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), IEEE, 2021, pp. 263–268.
- [49] N. F. AS’ SAHRA AND F. KOMPUTERAN, *Test case prioritization technique using sequence diagram and labeled transition systems in regression testing*, Diss. Universiti Teknologi Malaysia, 2015.
- [50] K.-W. SHIN AND D.-J. LIM, *Model-Based Test Case Prioritization Using an Alternating Variable Method for Regression*

- Testing of a UML-Based Model*, Applied Sciences, vol. 10, no. 21, Oct. 2020, p. 7537.
- [51] E. K. MECE, ET AL., *The Application Of Machine Learning In Test Case Prioritization - A Review*, European Journal of Electrical Engineering and Computer Science, vol. 4, no. 1, Jan. 2020.
- [52] L. FAN, ET AL., *Automatic Test Path Generation and Prioritization Using UML Activity Diagram*, 2021 8th International Conference on Dependable Systems and Their Applications (DSA), IEEE, 2021, pp. 484–490.
- [53] V. NGUYEN AND B. LE, *RLTCP: A Reinforcement Learning Approach to Prioritizing Automated User Interface Tests*, Information and Software Technology, vol. 136, Aug. 2021, p. 106574.
- [54] W. SORNKLIANG AND T. PHETKAEW, *Target-Based Test Path Prioritization for UML Activity Diagram Using Weight Assignment Methods*, International Journal of Electrical and Computer Engineering (IJECE), vol. 11, no. 1, Feb. 2021, p. 575.

Edited by: Anil Kumar Budati

Special issue on: Soft Computing and Artificial Intelligence for wire/wireless Human-Machine Interface

Received: Dec 11, 2023

Accepted: Apr 26, 2024



HIGH SPEED LOW POWER ANALYSIS OF 12 TRANSISTORS 2×4 LINE DECODER USING 45GPDK TECHNOLOGY

SRUTHI PAVANI JAVVADI* C R S HANUMAN† SIVADURGARAO PARASA‡ AND SANNAJAJI NARAGANENI§

Abstract. This paper proposes the high speed low power analysis of 12 transistors 2×4 Low Power (LP) and Low Power Inverting (LPI) Decoders by using Dual Value Logic (DVL) and Complementary Metal Oxide Semiconductor (CMOS) Logic. A huge challenge faced by this era of developing is power reduction. The LP circuit design is a requesting issue in high performance digital frameworks, for example, microchips, DSPs and other different applications. Power and speed are the main highlights considered while comparing any design. Diminishing chip area is additionally truly impressive factor, designers need to recall when suggesting any novel design. 2×4 LP and LPI Decoders using 12T (Transistor) is used for conversion of binary inputs to associated output bits in a pattern. A novel design (CMOS logic and DVL logic) of 2×4 LP and LPI Decoders using 12T is proposed with area optimization, LP and high speed in this paper. Delay and power is evaluated between the novel design and CMOS logic. The novel design of 12T LP and LPI 2×4 Decoders is 60.72% optimized for power in contrast to CMOS logic design at a typical value of 1.8V. The proposed method has been validated using Cadence 45 GPDK (Generic Product Design Key) Virtuoso Tool.

Key words: CMOS and DVL, LP Decoder, LPI Decoder and 12T.

1. Introduction. The bulk integrated circuits consist primarily of logic gates created utilizing static CMOS circuits [1]. The pullup PMOS and NMOS pulldown network are the two main components of a CMOS circuit. Display resilience in the face of background noise and device fluctuations, consistent performance at low voltages and small transistor sizes are two advantages of CMOS circuitry [2]. Since CMOS circuits can only accept inputs at the transistor's gate terminals, fewer building blocks are available when synthesizing cell-based logic. To compete with CMOS logic, the 1990s saw the development of Pass Transistor Logic (PTL) [3]. When compared to CMOS logic [4], pass transistor logic has advantages in terms of speed, power, and area. No matter which diffusion terminals of the transistors the inputs are tied to the either source/drain or gate, determines The primary design difference between pass transistor circuits. The two most popular techniques for building pass transistor circuits are, at first the PMOS and NMOS transistors are used. while in the second, a transmission gate is used to combine the two types of transistors in parallel [5].

The requirement for miniaturization and voltage scaling are two examples of how advances in VLSI technology have imposed new requirements on the design of swift, space- saving, and low-power logic systems. For high-performance computing devices like microprocessors and digital signal processors, LP design is a significant challenge. In computing, A straightforward combinational circuit called a decoder that transforms series of input signals into another code. Seven-segment displays, address decoding with in arrays of memory, data de-multiplexing and microchip/microcontroller personal based architectures are just a few of the many uses for line Decoders. Address Decoders are critical components because their design heavily influences consumption of power and access time of the SRAM [6] memory cell. For the sake of lowering the electrical bill, delay, and transistor count when constructing Decoders, a novel mixed logic approach is presented in this study [7].

*Department of Electronics and Communication Engineering, Sasi Institute of Technology and Engineering, Tadepalligudem, West Godavari, Andhra Pradesh, India, 534101 (sruthipavani@sasi.ac.in)

†Department of Electronics and Communication Engineering, Sasi Institute of Technology and Engineering, Tadepalligudem, West Godavari, Andhra Pradesh, India, 534101 (crshanuman@sasi.ac.in)

‡Department of Electronics and Communication Engineering, Sasi Institute of Technology and Engineering, Tadepalligudem, West Godavari, Andhra Pradesh, India, 534101 (sivadurgaraoparasa@gmail.com)

§Department of Electronics and Communication Engineering, Sasi Institute of Technology and Engineering, Tadepalligudem, West Godavari, Andhra Pradesh, India, 534101 (sannajaji@sasi.ac.in)

Table 2.1: Non-Inverting 2×4 Decoder Truth table

A	B	D_0	D_1	D_2	D_3
0	0	1	0	0	0
0	1	0	1	0	0
1	0	0	0	1	0
1	1	0	0	0	1

Table 2.2: Inverting 2×4 Decoder Truth table

A	B	I_0	I_1	I_2	I_3
0	0	0	1	1	1
0	1	1	0	1	1
1	0	1	1	0	1
1	1	1	1	1	0

2. Literature. Binary codes in digital systems are used to represent discrete amounts of information. An n-bit binary code can represent up to n distinct pieces of encoded data. Combining many circuits, a decoder converts binary data from n input lines to a maximum of $2 \times n$ different output lines or fewer if the n-bit coded data contains any unused combinations [8]. The circuits were looked at. These decoders are n-to-m line and produce the $m = 2 \times n$ input variable minterms.

A decoder with 2×4 lines produces four outputs from two inputs. Depending on the corresponding input combinations, there will only one active output at any given time. Table 2.1 summarizes the non-inverting 2×4 decoder's truth table [9], inputs A and B produces D_0, D_1, D_2 & D_3 as outputs. The complimentary outputs I_0, I_1, I_2 & I_3 produced by inverted 2×4 Decoder are always set to logic 0 for the selected output and logic 1 for the other three outputs as shown in 2.2[10].

Transmission gates are most frequently utilized in circuits using XOR logic, such as complete adders and multiplexers, as the main switching portion. In any event, as demonstrated in, we consider the possibility of utility in line decoders when AND/OR logic is applied. Figure display the TGL AND/OR gates with two possible inputs. Even if they are completely swinging, not every combination of inputs produces a restoration. The two most prevalent types of circuits in pass-transistor logic are those that employ only NMOS pass transistors, such as CPL, as well as DPL and DVL, which use both PMOS and NMOS pass transistors. With this study, we focus on the DVL approach, in which full swing operation is preserved while DPL is improved fewer transistors. Figure display two input DVL AND/OR gates [11]. Similar to the TGL gates, They can swing but not restore. CMOS NAND/NOR gates require four transistors, but TGL/DVL gates only require three. presuming the presence of complementing inputs. High-fan-out circuits called decoders allow multiple gates to share a limited number of inverters. Using TGL/DVL gates facilitates transistor count reduction [12].

One significant similarity trait of these gates' asymmetry is one of their characteristics, are the reality that their input loads are not balanced. We identified the two The X and Y inputs to the gate are displayed in Figure 2.1 [2]. In TGL gates, input X controls the gate terminals of three transistors. while input Y is sent from the input to the output node through the transmission gate.

However, only one gate terminal is controlled by input Y in DVL gates, and it is routed to the output. The X and Y inputs of the gate will receive both the control signal and the propagation signal. This asymmetric characteristic allows a designer to arrange signals by deciding which input acts as the control and which one acts as the signal propagation of gate in each situation. When the propagating signal is a complimentary input, there is a considerable increase in latency since an inverter needs to be added to the propagation channel. It is greater effectiveness to use the Inverse variable as the control signal when using the inhibition ($A'B$) or implication ($A'+B$) function. The AND (AB) and OR ($A+B$) operations have the same power. Finally, regardless of whether you're working with $A'+B'$ in NAND or $A'B'$ in NOR any picking to make will unavoidably result in a complementing propagation signal [12].

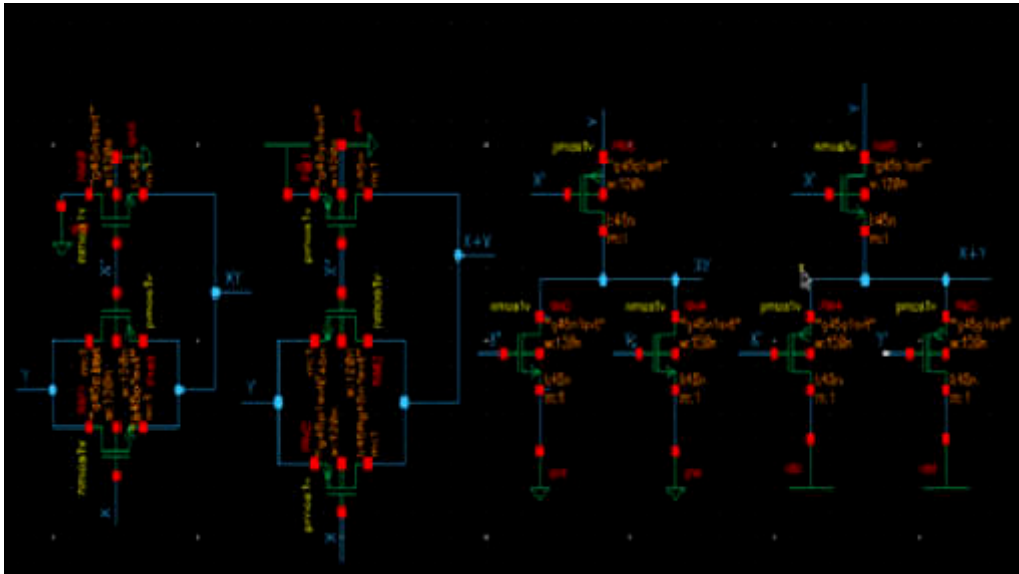


Fig. 2.1: (a) TGL AND gate , (b)TGL OR gate , (c) DVL AND gate ,(d) DVL OR gate

NAND and NOR gates have an advantage over AND and OR in typical CMOS architecture because they can express logic operations more efficiently with 4 transistors as compared to 6 transistors. It is possible to construct a 2×4 decoder using two inverters and four NOR gates. In contrast, an inverting decoder needs four NAND gates and two inverters which together produce “20” transistors.

Similarly the 16 min terms $D_0 D_{15}$ of the four input variables A, B, C, and D are produced by a Decoder with 4×16 lines, while the corresponding min terms $I_0 I_{15}$ are produced by inverted 4×16 line decoder. A predecoding approach, which divides blocks of n address bits into 1-of-2n pre decoded lines [12] that act as inputs to the final stage decoder, can be used to create such circuits. Pre decoding may be used to create such circuits [14]-[15]. As a result, two 2-four inverting decoders and sixteen 2-input NOR gates can be used to build a 4×16 decoder and two 2-four decoders and sixteen 2-input NAND gates can be used to implement an inverting one. These designs require eight inverters and twenty-four 2-input gates in CMOS logic, which adds up to “104” transistors per design [10].

2.1. 14T LP Topology for 2×4 Decoder (Non-Inverting). Developing a plan based on multiple logics A 2×4 decoder would require sixteen transistors altogether, two inverters, four TGL or DVL AND/OR gates, and two inverters. A 14 transistor decoder architecture is created by merging TGL and DVL AND gates in this design, which eliminates one of the two inverters, by carefully utilizing control and propagation signals.

In Figure 2.2 [1], the Decoder has two inputs A and B which produces the 4 min terms D_0, D_1, D_2 & D_3 , with the intention of removing inverter B. DVL AND gates are utilized to put into practice the min terms D_0 means $(A'B')$ and D_2 means (AB') , where Signals A and B are transmitted. The TGL AND gate is employed to realize the min terms [12]. D_1 means $(A'B)$ and D_3 means (AB) , with B functioning as the propagate signal for D_1 means $(A'B)$ and D_3 means (AB) each. The elimination of the B inverter is made possible by the choice of inputs and gates, resulting in a topology with 14T for the Decoder.

Figure 2.3, demonstrates the simulation of 14T LP Topology 2×4 Decoder. When $A=B=0$, D_0 will be enabled [14], when $A=0$ and $B=1$, D_1 will be enabled, when $A=1$ and $B=0$, D_2 will be enabled, when $A=1$ and $B=1$, D_3 will be enabled.

2.2. 14T LPI Topology for 2×4 Decoder (Inverting). Similarly, a 14T architecture with inverting 2 inputs 4 outputs Decoder can be built out of an inverter and four TGL/DVL OR gates. TGL /OR gates are utilized for min terms I_0 and I_2 , while DVL OR gates are utilized to put into practice for min terms I_1 and I_3

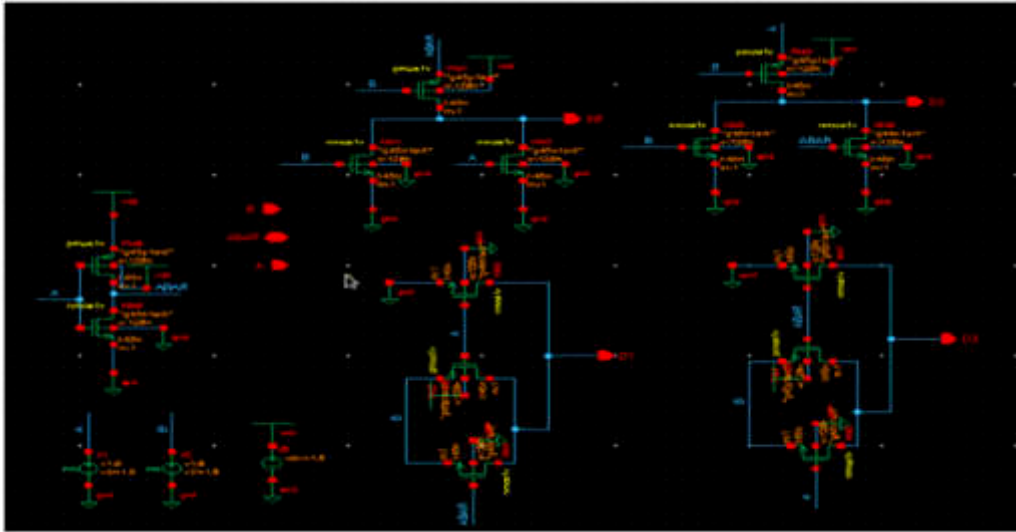


Fig. 2.2: 14T LP Topology for 2×4 Decoder (Non-Inverting)

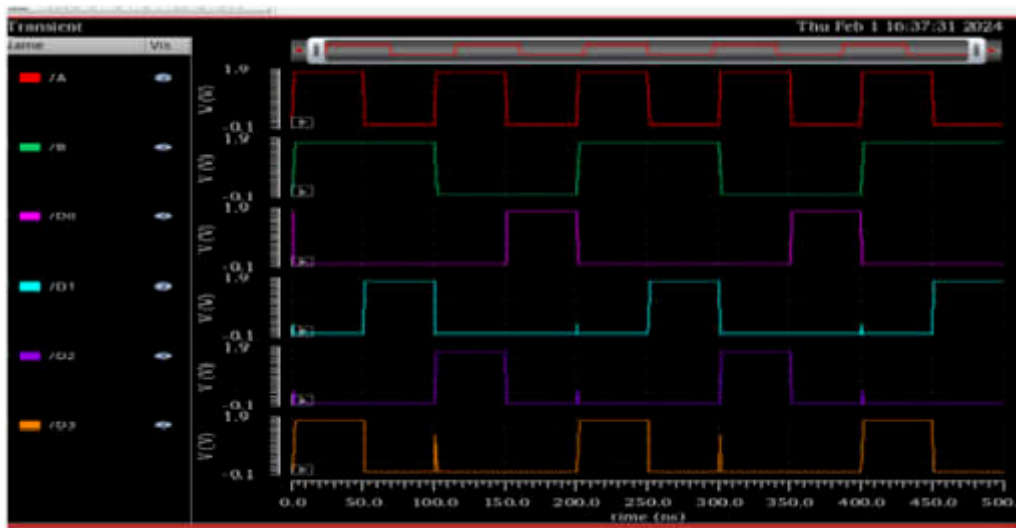


Fig. 2.3: Simulation result of 14T LP 2×4 Decoder (Non- Inverting)

are the output signal [16].The terms "2×4 LP" and "2×4 LPI," which "2×4 LP" stand for "Low power" and "2×4 LPI, stand for" Low Power Inverting" respectively, describe two LP Decoder architectures are used. These are shown Figure 2.2 and Figure 2.4.

Figure 2.5[1], demonstrates the simulation of 14T LPI Topology for 2×4 Decoder (Inverting). When A=B=0, I_0 will be enabled, when A=0 and B=1, I_1 will be enabled, when A=1 and B=0, I_2 will be enabled, when A=1 and B=1, I_3 will be enabled.

2.3. 15T HP Topology for 2×4 Decoder (Non-Inverting). Due to the complimentary propagate signal utilized in minterms D_0 and I_3 , One drawback of the above-discussed 14 transistor low power decoder topologies is their maximum latency. Due to their lack of requirement for complementary inputs, these minterms may be implemented using normal CMOS logic gates, which overcomes this disadvantage. minterm D_0 is

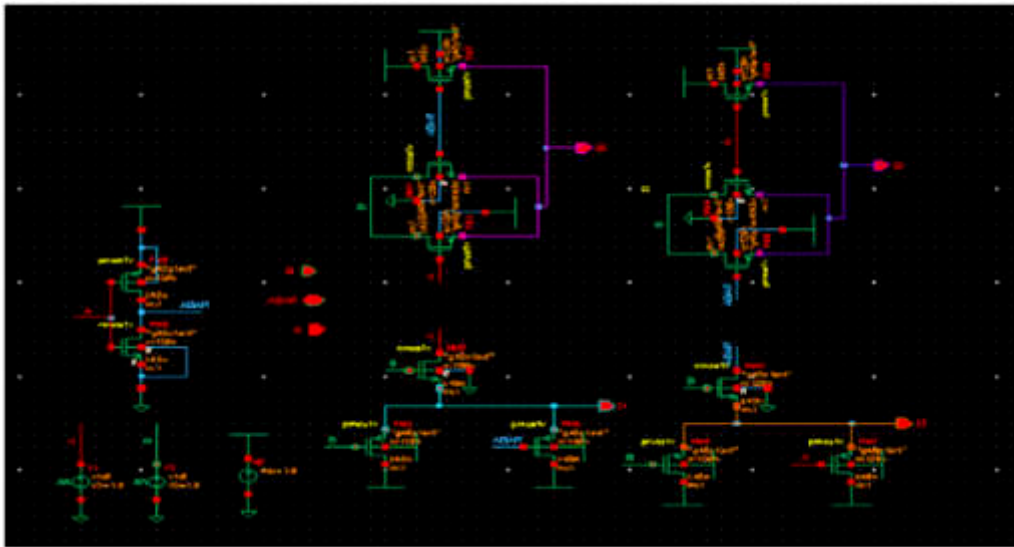


Fig. 2.4: 14T LPI Topology for 2x4 Decoder (Inverting)

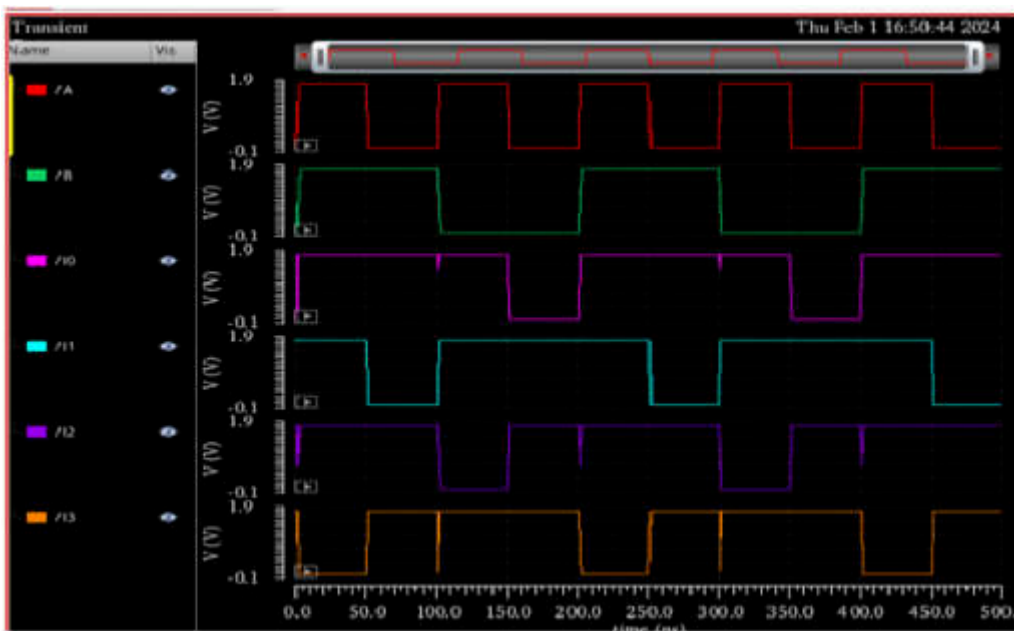


Fig. 2.5: Simulation result of 14T LPI 2x4 Decoder (Inverting)

put into practice using a CMOS NOR gate, while I_3 is constructed using a CMOS NAND gate. One more transistor is added to each structure. The resultant decoder structure, known as High Performance (HP) topology, comprises three distinct logic types in a single circuit (CMOS, TGL, and DVL) [17], improving power and delay performance are both excellent. The schematics of 2x4 HP decoder is as illustrated in Figure 2.6. This modification resulted in a Decoder architecture that has three distinct logic types in a single circuit (CMOS, TGL, and DVL) into a single circuit that improves power and timing efficiency. The HP topology describes

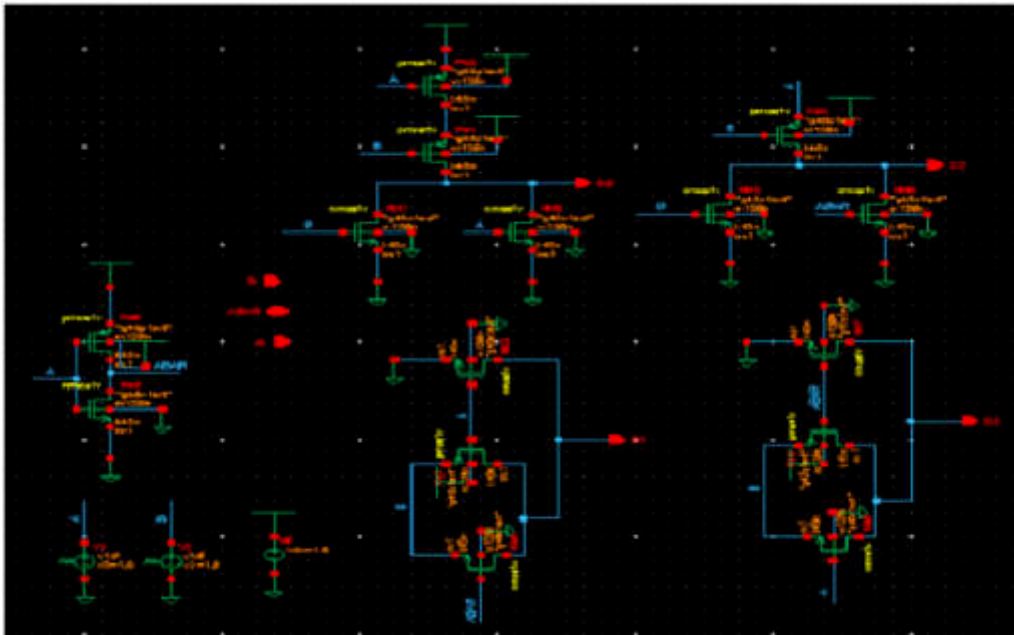


Fig. 2.6: 15T HP Topology for 2×4 Decoder (Non-Inverting)

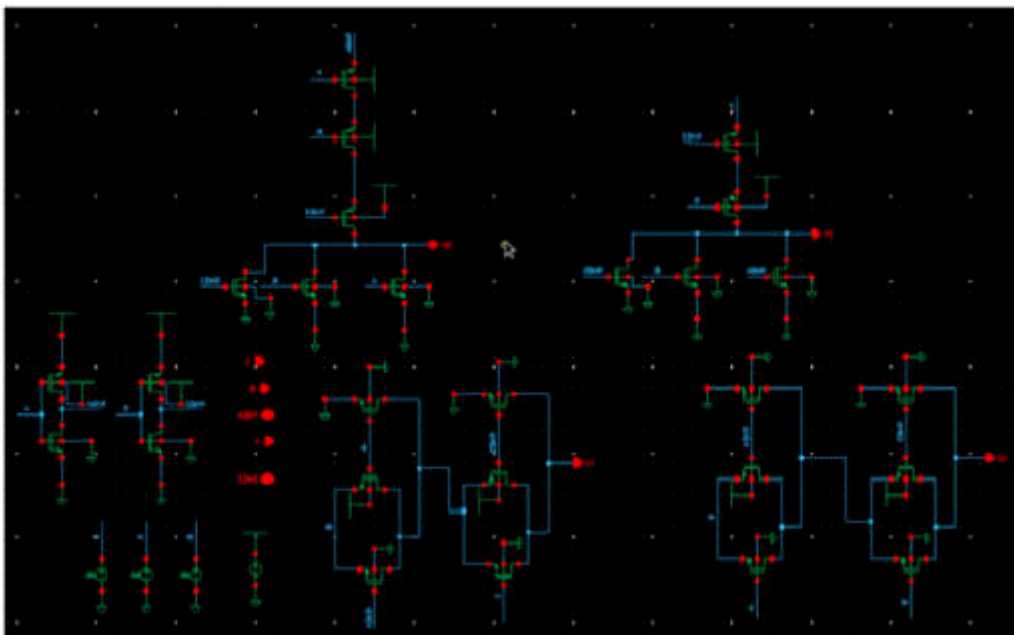


Fig. 2.7: Simulation result of 15T HP 2×4 Decoder (Non-Inverting)

this configuration specifically for its high throughput [9].

Figure 2.7[1], Shows the simulation of 15T HP Topology for 2×4 Decoder (Non-Inverting). When $A=B=0$, D_0 will be enabled, when $A=0$ and $B=1$, D_1 will be enabled, when $A=1$ and $B=0$, D_2 will be enabled, when $A=1$ and $B=1$, D_3 will be enabled.

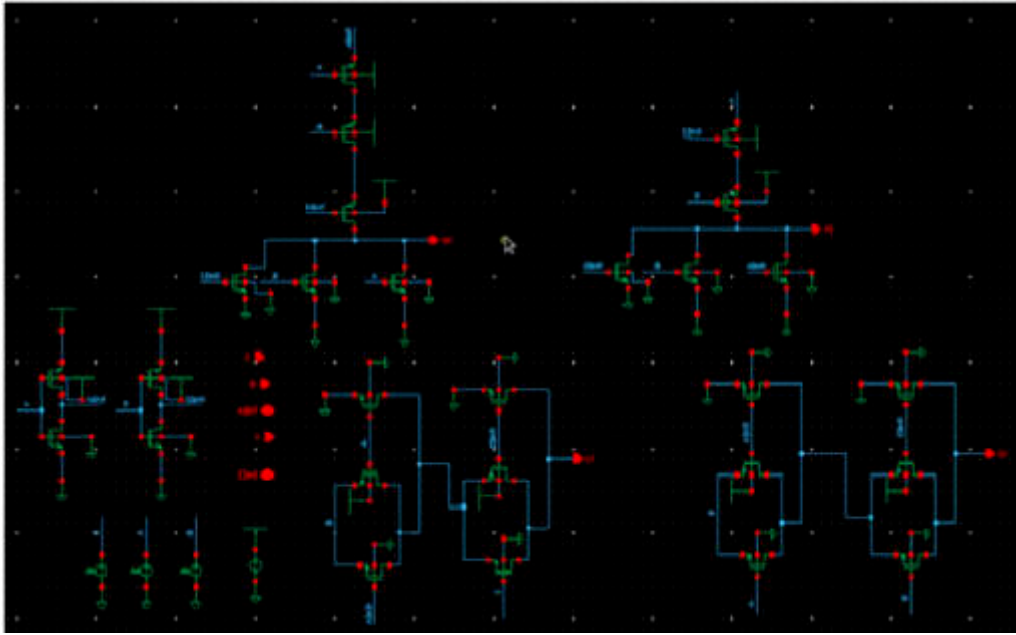


Fig. 2.8: 2×4 HP Mixed Logic Decoder with Enable (Non-Inverting)

2.4. 2×4 HP Mixed Logic Decoder with Enable (Non-Inverting). This section covers the design of decoders with enable input. The use of 2×4 decoder blocks can reduce the number of transistors in a $n: 2n$ decoder. Since they don't need any extra logic gates, 2×4 decoders can be used to build higher stage decoders, which reduces design complexity. There are a number of benefits to logic gates are used in the post-decoder phases [18]– [22], including a reduction in wire delays and cross-talk caused by connections. However, the area overhead increases linearly with the number of logic gates needed. Only decoders without enable input are designed in every previous decoder study utilizing method of mixed logic design. This work uses the method of mixed logic design to construct decoders with enable input. The regulated functioning of decoder circuits is achieved with the addition of an enable input. In other words, only when the enable input is set to "on" does the decoder function, which also lowers dynamic power dissipation. Furthermore, instead of implementing at the transistor or gate level, it is always advised in contemporary techniques for chip design to construct macros or minor circuits that can be reused based on requirements.

A total of 30 transistors are needed for the non-reversing standard CMOS 2×4 decoder which includes enable, which also needs three NOR gates with three inputs and three inverters. DVL AND gates can be used to implement D_0 and D_2 in the 2×4 LP decoder which includes enable. A is used as the propagation signal, while EN and B are the signs of control; EN is the enable input. TGL AND gates are used to implement the minterms D_1 and D_3 , with B serving as the propagation signal and A and EN serving as the signs of control. 2×4 LP decoder which includes enable input requires 26 transistors. Similarly for 2×4 HP decoder which includes enable input requires "27" transistors use a CMOS NOR gate with three inputs to replace D_0 minterm as shown In Figure 2.8

Figure 2.8 [1], Shows the mixed logic line Decoder with enable. Inputs A and B produces D_0, D_1, D_2 & D_3 as outputs. Figure 2.9, Shows the simulation of Mixed Logic 2×4 HP Decoder with Enable (Non- Inverting). When $A=B=0$, D_0 will be enabled, when $A=0$ and $B=1$, D_1 will be enabled, when $A=1$ and $B=0$, D_2 will be enabled, when $A=1$ and $B=1$, D_3 will be enabled.

3. Proposed Methodology. A novel 2×4 Decoder with only 12T Transistor is proposed with area optimization in this research. CMOS logic is additionally used for execution of 2×4 Decoder. Delay and power is used for evaluation between the novel design and CMOS logic. The novel design of 2×4 Decoder [1] is Less

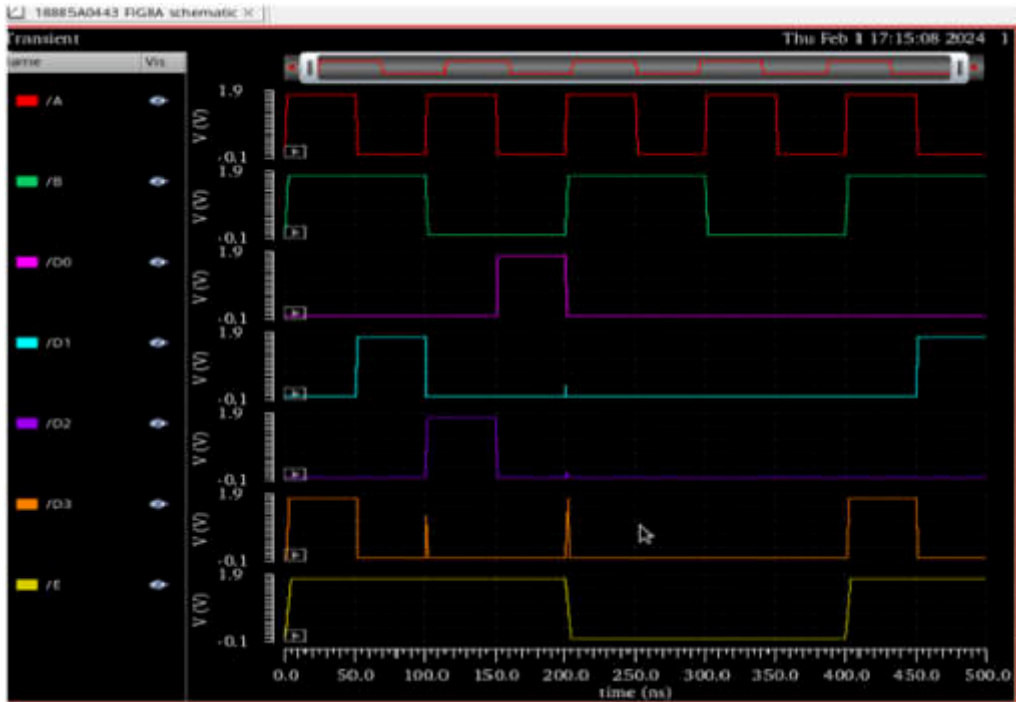


Fig. 2.9: Simulation of New Mixed-Logic 2×4 HP Decoder with Enable (Non-Inverting)

Table 3.1: Truth table for 2×4 LP Decoder

A	B	D_1	D_2	D_3	D_4
0	0	1	0	0	0
0	1	0	1	0	0
1	0	0	0	1	0
1	1	0	0	0	1

optimized for power in contrast to CMOS logic design at a typical value of 1.8V.

The LP circuit designs is a requesting issue in high performance digital frameworks, for example, microchips, DSPs and other different applications. Power and speed are the main high lights considered while comparing any circuit or design. Diminishing chip area is additionally truly impressive factor, creators need to recall when suggesting any novel design. Decoder is used for conversion of binary inputs to associated output bits in a pattern. There are wide range of applications of Decoder such as seven-segment display, data de-multiplexing, etc. Numerous studies using sequential and combinational circuits are currently being conducted different logics.

3.1. 12T LP Topology for 2×4 Decoder (Non-Inverting). Decoders are crucial circuits, for the most part utilized in the hardware involving collections of RAM [19]. In this study, we propose an innovative approach to putting them into practice, which rapidly decreases the amount of transistors in 2×4 Decoder circuit. Power and area efficient Decoders with less number of transistors plays a very significant role in circuit designing and act as elementary units. Figure 3.1 Shows proposed non inverting 2×4 Decoder generates 4 min terms D_1, D_2, D_3 & D_4 with two inputs A and B and its truth table is shown in below Table 3.1.

The proposed circuit for implementation of 12T LP 2×4 Decoder designed using CMOS and Dual Value Technique is shown in Figure 3.6. One of the four outputs is chosen and set to 1 depending on the input

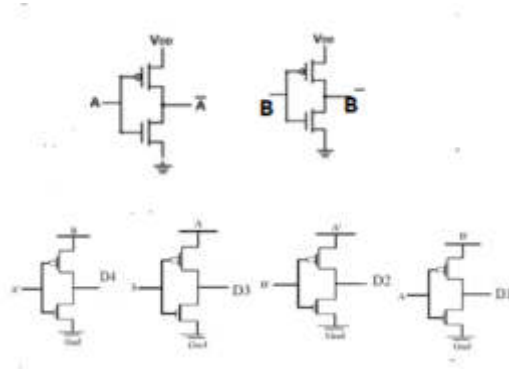


Fig. 3.1: Proposed 2x4 Decoder with 12T

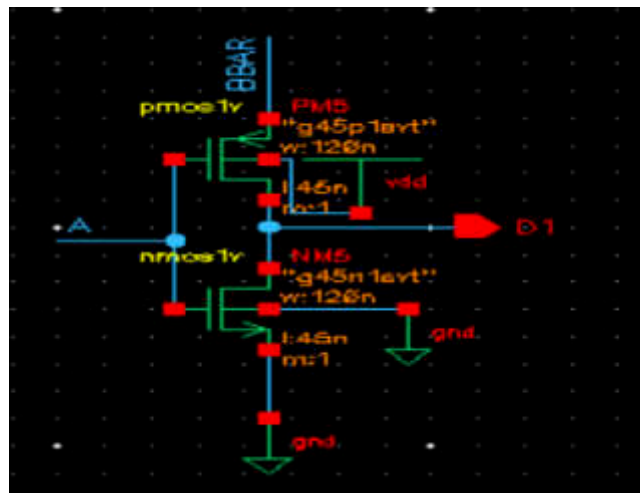


Fig. 3.2: 12T LP Topology for 2x4 Decoder with "D₁" data output

combination, while the other three are set to 0 [7].

3.1.1. 12T LP Topology for 2x4 Decoder with "D₁" data output. The proposed circuit for implementation of 12T LP 2x4 Decoder designed using CMOS and Dual Value Technique is shown in Figure 3.6. First term "D₁" using proposed design as shown in Figure 3.2 is both PMOS and NMOS gates has been connected with input A, PMOS drain has connected with one of the input complement of B. "D₁". will be formed by the combination of the PMOS and NMOS linked to the output

3.1.2. 12T LP Topology for 2x4 Decoder with "D₂" data output. The proposed circuit for implementation of 12T LP 2x4 Decoder designed using CMOS and Dual Value Technique is shown in Figure 3.6. Second term "D₂" using proposed design as shown in Figure 3.3 is both PMOS and NMOS gates has been connected with input complement of B, PMOS drain has connected with one of the input complement of A. PMOS and NMOS combined with their connections to the output will result in a "D₂".

3.1.3. 12T LP Topology for 2x4 Decoder with "D₃" data output. The proposed circuit for implementation of 12T LP 2x4 Decoder designed using CMOS and Dual Value Technique is shown in Figure 3.6. Third term "D₃" using proposed design as shown in Figure 3.4 is both PMOS and NMOS gates has been connected with input B, PMOS drain has connected with one of the input A. A "D₃" will be produced by connecting PMOS and NMOS.

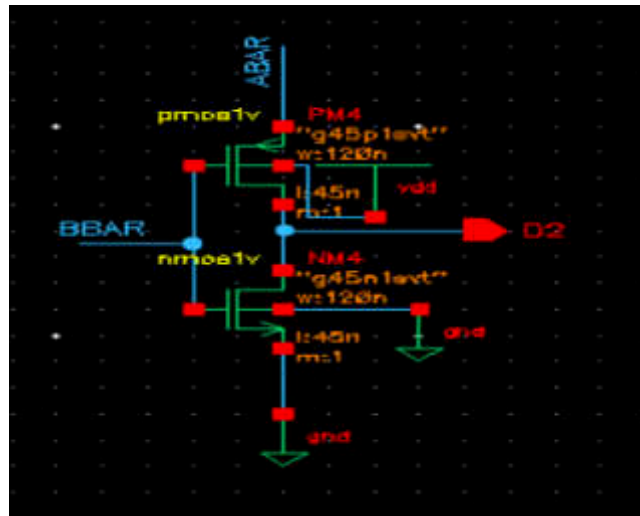


Fig. 3.3: 12T LP Topology for 2x4 Decoder with "D₂" data output

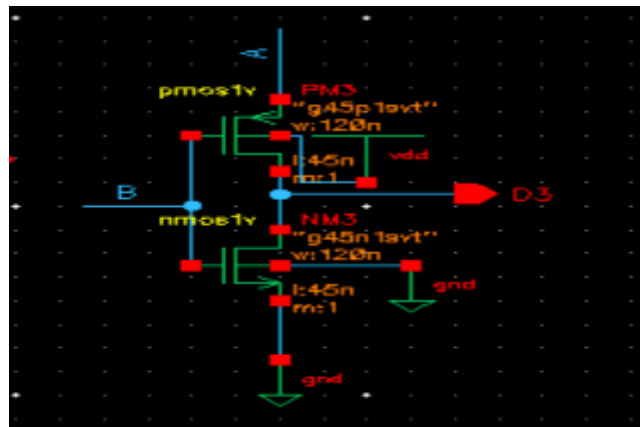


Fig. 3.4: 12T LP Topology for 2x4 Decoder with "D₃" data output

3.1.4. 12T LP Topology for 2x4 Decoder with "D₄" data output. The proposed circuit for implementation of 2x4 Decoder designed using CMOS and Dual Value Technique is shown in Figure 3.6. Fourth term "D₄" using proposed design as shown in Figure 3.5 is both PMOS and NMOS gates has been connected with input complement of A, PMOS drain has connected with one of the input B. The combination of the both PMOS and NMOS linked to the output will produce data "D₄".

The proposed circuit 12T LP Topology for 2x4 Decoder has been validated with Cadence Virtuoso with 45GPDK Technology. The entire technology designed with CMOS and Dual Value Methodology shown in Figure 3.6.

Figure 3.7 Shows the simulation of Non-Inverting 12T LP Topology for 2x4 Decoder with CMOS and DVL topology. When A=B=0, D₁ will be enabled, when A=0 and B=1, D₂ will be enabled, when A=1 and B=0, D₃ will be enabled, when A=1 and B=1, D₄ will be enabled.

3.2. 12T LPI Topology for 2x4 Decoder (Inverting). Figure 3.8 shows 12T 2x4 Decoder in inversion mode. It contains two inverters that will produce complement of A and complement of B. i.e A' and B'. A 2x4

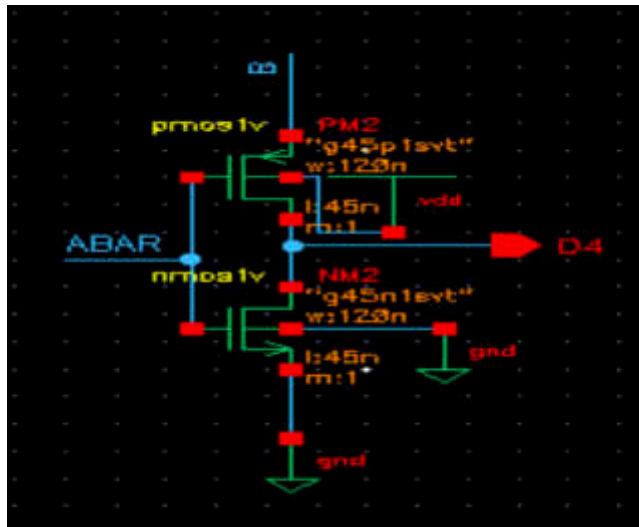


Fig. 3.5: 12T LP Topology for 2x4 Decoder with "D₄" data output

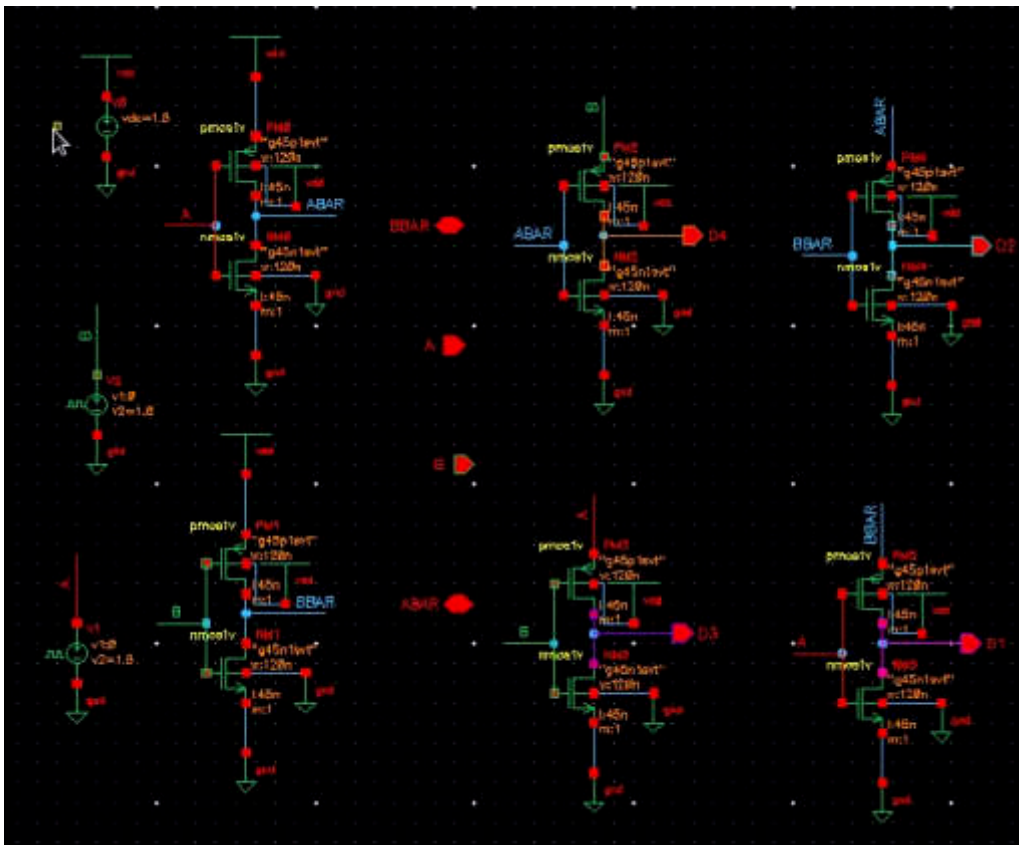


Fig. 3.6: Non-Inverting 12T LP Topology for 2x4 Decoder

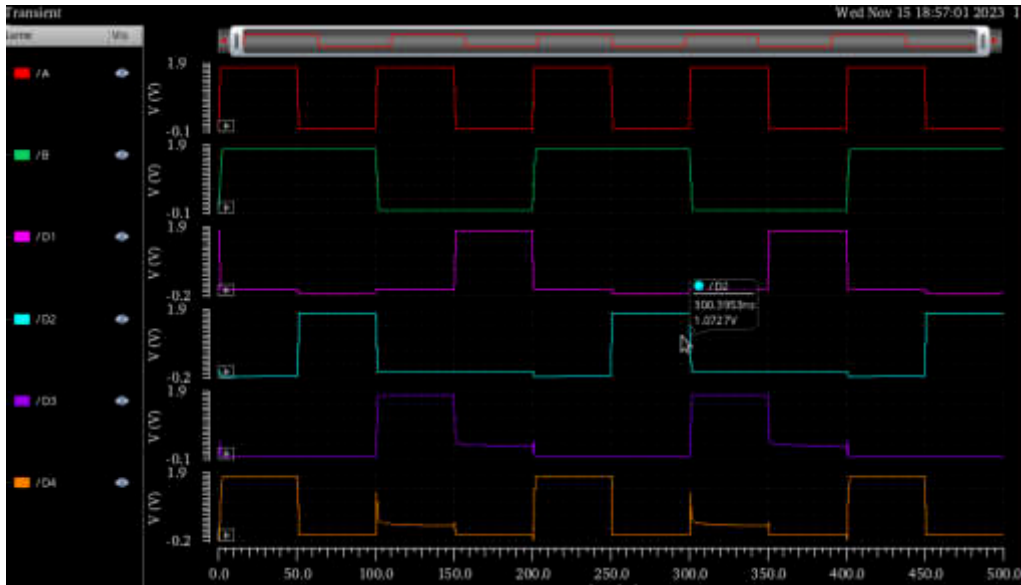


Fig. 3.7: Simulation result of Non-Inverting 12T LP 2×4 Decoder

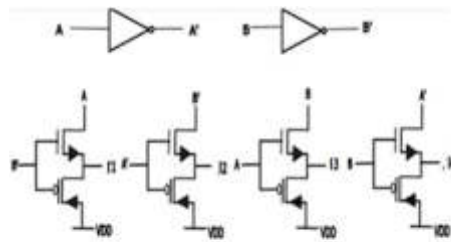


Fig. 3.8: 12T LPI Topology for 2×4 Decoder (Inverting)

Table 3.2: Truth table for 2×4 LPI Decoder (Inverting)

A	B	I_1	I_2	I_3	I_4
0	0	0	1	1	1
0	1	1	0	1	1
1	0	1	1	0	1
1	1	1	1	1	0

LPI Decoder generates the 4 minterms I_1, I_2, I_3 & I_4 of two inputs A and B and its truth table is shown in Table 3.2.

3.2.1. 12T LPI Topology for 2×4 Decoder (Inverting) with "I₁" data output. The proposed circuit for implementation of 12T LPI 2×4 Decoder designed using CMOS and Dual Value Technique as shown in Figure 3.13. First term "I₁" using proposed design as shown in Figure 3.9 is both PMOS and NMOS gates has been connected with complement of B, NMOS drain has connected with one of the input A. PMOS source is connect with V_{DD} . The combination of the both PMOS and NMOS connected to the output will produce

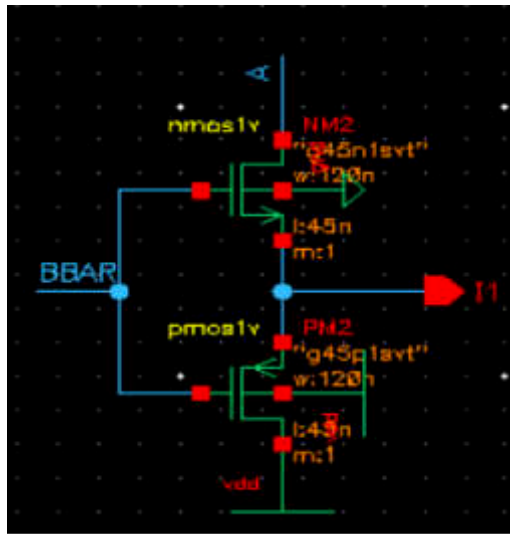


Fig. 3.9: 12T LPI Topology for 2×4 Decoder (Inverting) with "I₁" data output.

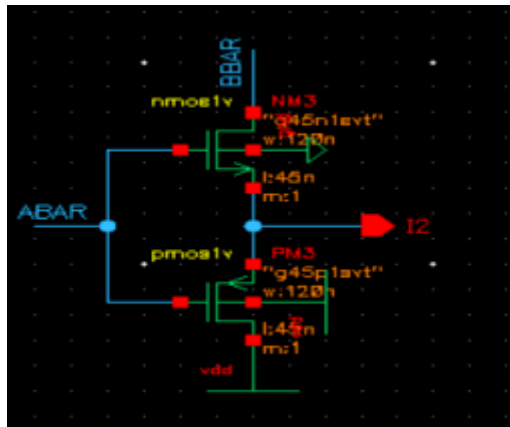


Fig. 3.10: 12T LPI Topology for 2×4 Decoder (Inverting) with "I₂" data output.

data "I₁".

3.2.2. 12T LPI Topology for 2×4 Decoder (Inverting) with "I₂" data output. The proposed circuit for implementation of 12T LPI 2×4 Decoder designed using CMOS and Dual Value Technique as shown in Figure 3.13. Second term "I₂" using proposed design as shown in Figure 3.10 is both PMOS and NMOS gates has been connected with complement of A, NMOS drain has connected with one of the input complement of B. PMOS source is connect with V_{DD} . The combination of the both PMOS and NMOS connected to the output will produce data "I₂".

3.2.3. 12T LPI Topology for 2×4 Decoder (Inverting) with "I₃" data output. The proposed circuit for implementation of 12T LPI 2×4 Decoder designed using CMOS and Dual Value Technique as shown in Figure 3.13. Third term "I₃" using proposed design as shown in Figure 3.11 is both PMOS and NMOS gates has been connected with input A, NMOS drain has connected with one of the input B. PMOS source is connect with V_{DD} . The combination of the both PMOS and NMOS linked to the output will produce data "I₃".

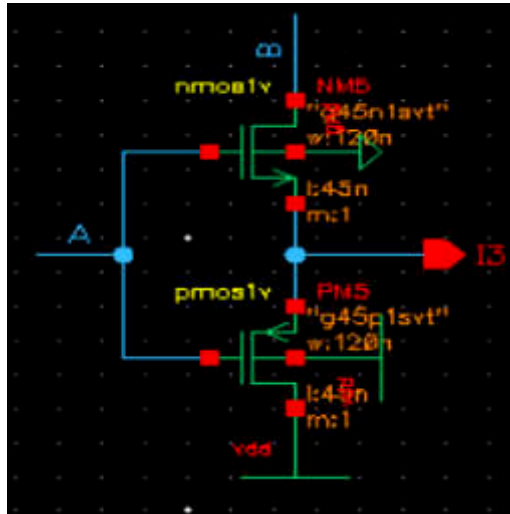


Fig. 3.11: 12T LPI Topology for 2×4 Decoder (Inverting) with "I₃" data output.

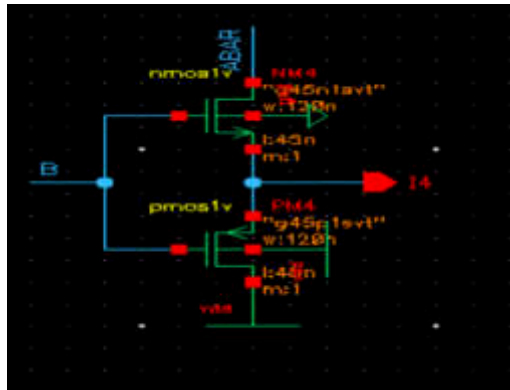


Fig. 3.12: 12T LPI Topology for 2×4 Decoder (Inverting) with "I₄" data output.

3.2.4. 12T LPI Topology for 2×4 Decoder (Inverting) with "I₄" data output. The proposed circuit for implementation of 12T LPI 2×4 Decoder designed using CMOS and Dual Value Technique as shown in Figure 3.13. Fourth term "I₄" using proposed design as shown in Figure 3.12 is both PMOS and NMOS gates has been connected with input B, NMOS drain has connected with one of the input complement of A. PMOS source is connect with V_{DD} . The combination of the both PMOS and NMOS connected to the output will produce data "I₄".

The proposed circuit 12T LPI Topology for 2×4 Decoder has been validated with Cadence Virtuoso with 45GPDK Technology. The entire technology Designed with CMOS and Dual Value Methodology as shown in Figure 3.13.

Figure 3.14, Shows the simulation results of 12T LPI Topology for 2×4 Decoder with CMOS and DVL Topology. When A=B=0, I₁ will enabled, when A=0 and B=1, I₂ will be enabled, when A=1 and B=0, I₃ will be enabled, when A=1 and B=1, I₄ will be enabled.

4. Results and Discussion. Every simulation is run using the 45 GDPK technology Cadence Virtuoso tool. The Table 4.1 shows that when comparison with the proposed 12T 2×4 LP LP and LPI Decoders, with existing 2×4 LP Decoders, Since the suggested approaches increase power with a delay overhead and fewer

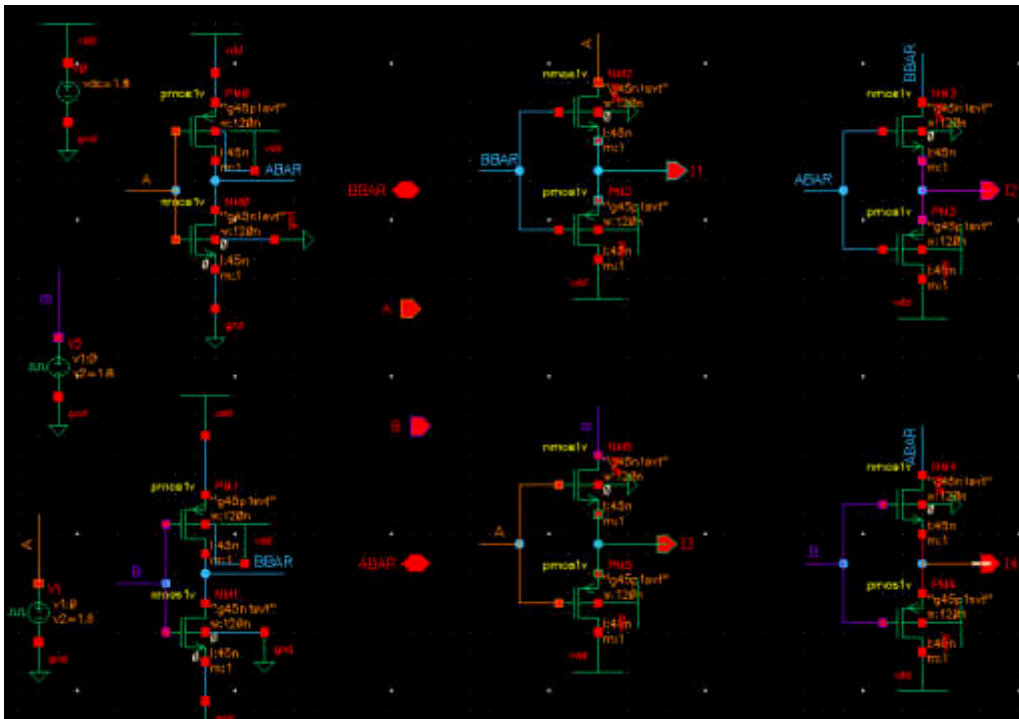


Fig. 3.13: Inverting 12T LPI Topology for 2×4 Decoder

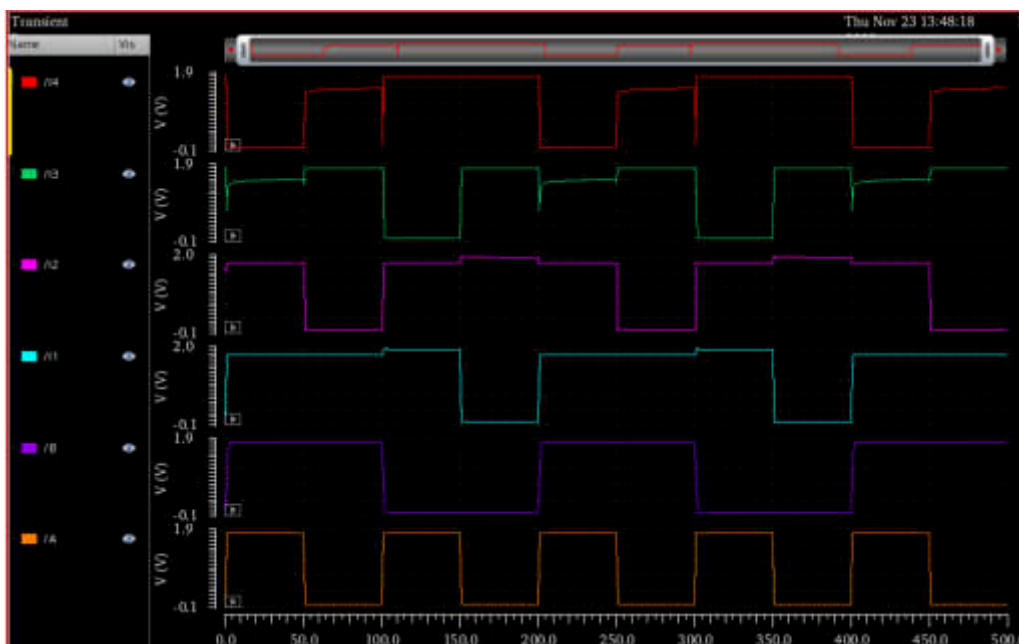


Fig. 3.14: Simulation result of Inverting 12T LPI 2×4 Decoder

Table 4.1: Comparison of Results

S.No	Method Decoder	AVG Power (w)	Static Power (w)	Dynamic Power (w)	Delay (sec)
1	14T LP	509.8E-9	721 p	80.69 μ	100.7E-9
2	14T LPI	530.8E-9	680.47 p	992.41 p	211.8E-12
3	15T HP	572.5E-9	730.5 p	87.143 μ	50.23E-9
4	15T HPI	16.81E-6	680.04 p	54.15 μ	210.6E-12
5	26T LP	726.8E-9	1.07 n	82.97 μ	149.5E-9
6	26T LPI	32.82E-3	1.39 n	66.44 μ	100.3E-9
7	26T HP	724.0E-9	1.12 n	83.91 μ	150.3E-9
8	26T HPI	32.82E-3	1.474n	66.44 μ	100.9E-9
9	12T LP Proposed	1.129E-6	105.28 p	5.36 μ	631.2E-12
10	12T LPI Proposed	651.9E-9	1.053 n	37.51 μ	50.34E-9

transistors, they are suitable for scenarios where size and power loss are the main design issues. The two innovative topologies (CMOS and DVL) used in the design of the proposed decoders are low power and low power inverting, respectively. The propagation delay, the circuit's static and dynamic powers, and the overall average power are computed in each scenario. For simulation, an operating voltage of 1.8V is utilized. When compared to their typical CMOS predecessors, All of the Applied decoders can swing freely and have fewer transistors. The suggested 12T 2×4 LP decoder consumes average power of 1.129 μ w, static power of 105.28 pw, dynamic power of 5.36 μ w and delay of 631.2 ps and 12T 2×4 LPI decoder consumes average power of 651.9 nw, static power of 1.053 nw, dynamic power of 37.51 μ w and delay of 50.34 ns. The results for 12T LP 2×4 and 12T LPI 2×4 Decoders are tabulated in Table 4.1

5. Conclusion. The design in this research was validated using Cadence Virtuoso 45GPDK Technology. An efficient Decoders designed by blending DVL Topology with static CMOS. Simulation results prove that, when comparison with the proposed 12T LP, LPI Decoders with existing 2×4 LP, LPI Decoders provides the upgrade in power with an upward on delay with diminished quantity of transistors and are acceptable for operations in area and power dissipation are the crucial device consideration. The 2×4 Decoders implemented has essentially less number of transistors so it will inhabit less on chip area as compared to CMOS logic which can be observed from 4.1. The novel design is 60.72% optimized for power as compared to CMOS Logic design at a typical value of 1.8V which can also be observed from the Table 4.1. these Implemented Decoders have full swing capacity and diminished quantity of transistors count in contrast to usual CMOS counterparts.

REFERENCES

- [1] SUMANA, N.S., SAHANA, B. AND DESHAPANDE, A.A, "Design and implementation of low power-high performance mixed logic line decoders", 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT).IEEE, pp.529-534, 2019.
- [2] KUMAR, SAVALAM, C.S., PRASANTI, K. AND HARANATH, A.S, "Design and implementation of high performance and low power mixed logic line decoders", International Journal of Innovative Technology and Exploring Engineering (IJITEE),8(6S4), 2019.
- [3] RAUT, K.P.M. AND DESHMUKH, D.R, "Design of Low-Power 2-4 Mixed Logic Line Decoders with Clock Based Technique", International Research Journal of Engineering and Technology (IRJET), pp.2395-0056, 2018.
- [4] KUMAR, A.P. AND MOUNIKA, K, "Design of Low Power, High-Performance 2-4, and 4-16 Mixed-Logic Line Decoders", International Journal of Engineering Trends and Applications (IJETA), 4(5), 2017.
- [5] NIRMALA, M., USHASREE, V. AND PAVANI, K., "Design of Low Power, High Performance 2-4 and 4-16 Decoders by using GDI methodology", IJEECS, 6(12), pp.392-397, 2017.
- [6] TABASSUM, K.F. AND DAKEY, S, "Design of Area and Power Efficient Line Decoders for SRAM", International Journal of Emerging Technologies in Engineering Research (IJETER), 5(11), 2017.
- [7] G. KIRAN AND L. N. REDDY, "Design of area efficient high-performance 2-4 and 4-16 mixed-logic line decoders", International Journal of Professional Engineering Studies, 9(4), pp.326-332, 2017.
- [8] S. RANJITHA AND P.SRIKANTH, "Design of decoders using mixed logic for various applications", International Journal of Engineering and Advanced Technology (IJEAT), pp.43-47, 2017.

- [9] M. S. PRASANNA AND K. SEETHARAM, "Design of low-power high-performance 2-4 and 4-16 mixed-logic line decoders", International Journal of Research in Engineering, Science and Management ,1(1), 2018.
- [10] BALOBAS, D. AND KONOFAOS, N., "Design of low-power high-performance 2-4 and 4-16 mixed-logic line decoders", IEEE Transactions on Circuits and Systems II: Express Briefs, 64(2), pp.176-180, 2016.
- [11] BHATNAGAR, V., ATTRI, C. AND PANDEY, S, "Optimization of row decoder for 128× 128 6T SRAMs", International Conference on VLSI Systems, Architecture, Technology and Applications (VLSI-SATA) (pp. 1-4). IEEE., pp.1-4, 2015.
- [12] N. H. WESTE AND D. HARRIS, "CMOS VLSI design: A circuits and systems perspective", Pearson Education India, 2015.
- [13] MISHRA, A.K., ACHARYA, D.P. AND PATRA, P.K, "Novel design technique of address Decoder for SRAM", IEEE International Conference on Advanced Communications, Control and Computing Technologies, pp.1032-1035, 2014.
- [14] BRZOZOWSKI, I., ZACHARA, Ł. AND KOS, A, "Designing Method of Compact n-to-2n Decoders", International Journal of Electronics and Telecommunications, 59(4),2013.
- [15] LOTZE, N. AND MANOLI, Y, "A 62 mV 0.13μ m CMOS Standard-Cell-Based Design Technique Using Schmitt-Trigger Logic", IEEE journal of solid-state circuits, 47(1), pp.47-70, 2011.
- [16] TURI, M.A. AND DELGADO-FRIAS, J.G, "High-performance low-power selective precharge schemes for address decoders", IEEE transactions on circuits and systems II: express briefs, 55(9), pp.917-921, 2008.
- [17] D.MARKOVIĆ, B. NIKOLIĆ, AND V. OKLOBDŽIJA, "A general method in synthesis of pass-transistor circuits", Microelectronics Journal, 31(11-12), pp.991-998, 2000.
- [18] R. ZIMMERMANN AND W. FICHTNER, "Low-power logic styles: CMOS versus pass-transistor logic", IEEE journal of solid-state circuits, 32(7), pp.1079-1090, 1997.
- [19] V.G. OKLOBDZIJA AND B.DUCHENE, "Pass-transistor dual value logic for low-power CMOS", International Symposium on VLSI Technology, Systems, and Applications. Proceedings of Technical Papers.IEEE, pp.341-344, 1995.
- [20] M.SUZUKI, N.OHKUBO, T.SHINBO, T.YAMANAKA, A.SHIMIZU, K.SASAKI AND Y. NAKAGOME, "A 1.5-ns 32-b CMOS ALU in double pass-transistor logic", IEEE Journal of Solid-State Circuits, 28(11), pp.1145-1151, 1993.
- [21] X. WU, "Theory of transmission switches and its application to design of cmos digital circuits", Systems, and Applications. Proceedings of Technical Papers.IEEE, 20(4), pp.349-1356, 1992.
- [22] K. YANO, T. YAMANAKA, T. NISHIDA, M.SAITO, K.SHIMOHIGASHI, AND A. SHIMIZU, "A 3.8-ns CMOS 16* 16-b multiplier using complementary pass-transistor logic", IEEE journal of solid-state circuits, 125(2), pp.388-395, 1990.

Edited by: Anil Kumar Budati

Special issue on: Soft Computing and Artificial Intelligence for wire/wireless Human-Machine Interface

Received: Dec 30, 2023

Accepted: Mar 8, 2024



MULTI OBJECTIVE DATA TRANSFORMATION IN HYBRID CLOUDS NETWORKS FOR OFFLOADING DATA

V SRIDHAR REDDY*, N. JAYANTHI†, SHARON ROSE VICTOR JUVVANAPUDI‡, SRINIVAS BACHU§ AND MADIPALLI SUMALATHA¶

Abstract. Recently hybrid cloud solutions integrating public and private cloud is proposed to address the privacy and security concerns faced by Enterprises in their data offloading decisions. In these solutions, the transformed data is kept in public cloud while transformation keys are kept in private cloud. The existing works for data transformation used in hybrid clouds does not address multiple objectives of privacy, security, fine grained access control, utility preservation for mining and data retrieval efficiency. This work proposes a multi objective data transformation technique for hybrid cloud to address all these objectives. The proposed solution is built on attribute based hierarchical data access control with hierarchy selection based on joint consideration of security, utility preservation and retrieval efficiency. The proposed solution is able to provide 5% higher security strength, 1.34% higher clustering accuracy over perturbed data and 29.95 % higher data retrieval efficiency over perturbed data compared to existing works.

Key words: Multi objective data transformation, Hybrid cloud, Hierarchical data access control, Generalization control .

1. Introduction. Enterprises are adopting cloud for offloading both storage and computations. The adoption is triggered due to various benefits like CAPEX and OPEX reduction, high availability and mobility etc. With increasing cloud adoption rate, there is also increasing cloud security breaches. The recent security survey by IDC and Ermetic [1] reports that almost 98% of enterprises suffer atleast one security breach. The average total cost of data breach globally is estimated about 4.24 million USD. Data breaches and leakage can create huge financial loss for the Enterprise, lose to competitors and sometimes wipe out from market. Thus ensuring security and privacy of data has become a important requirement for enterprises in their cloud offloading and vendor selection decisions. Though there are various data protection mechanisms, enterprises are adopting multi cloud and hybrid clouds to reduce the risk. Flexera's 2021 State of the Cloud Report [2] found that almost nine out of ten enterprises are adopted multi cloud approach and eight in that nine enterprises are adopting hybrid cloud to reduce security risks. Hybrid cloud solutions are also not full proof. Though they have reduced risks and breach cost compared to public and private cloud, they could not completely eliminate the data breach cost as evident from the IBM and the Ponemon Institute's 2021 Cost of a Data Breach Report Figure 1.1

This data breach cost could be avoided in hybrid cloud with more effective security and privacy enforcement. The data must be prevented from compromise either directly or through inference. Many works have been proposed in category of anonymization, randomization, cryptography, diversification and aggregation to address the security and privacy concerns. In addition to security and privacy, the data transformation techniques must also address other requirements like differential access control, utility preservation and retrieval efficiency. Most solutions as discussed in Section II do not address all these requirements. This work proposes a multi objective data transformation technique which jointly addresses all the five requirements of privacy, security, differential

*Department of Information Technology, Vignana Bharathi Institute of Technology, Hyderabad, India (vsridharreddy19@gmail.com)

†Department of Computer Science Engineering, CMR Institute of Technology, Bengaluru, India (jayanthi.n@cmrit.ac.in)

‡Department of Electronics and Communication Engineering, Pragati Engineering College, Surampalem, AP, India (jsr.victor@gmail.com)

§Department of Electronics and Communication Engineering, Siddhartha Institute of Technology & Science, Hyderabad, Telangana, India (bachusrinivas@gmail.com)

¶Department of Electronics and Communication Engineering, Siddhartha Institute of Technology & Science, Hyderabad, Telangana, India (sumasriee@gmail.com)



Fig. 1.1: Data breach report (Courtesy: IBM and Ponemon Report 2021)

access control, utility preservation and retrieval efficiency. Attribute based hierarchical data transformation with mix of anonymization, aggregation and diversification is adopted in this work with hierarchy selection fine tuned to address the five requirements.

The rest of paper is organized as follows. Section II provides the survey of data transformation techniques used in cloud and their shortcomings. Section III details the proposed multi objective data transformation technique. Section IV provides the results of the solution and its comparison to existing works. Section V provides the concluding remarks and the scope for future work.

2. Related Work. A survey of data transformation techniques for cloud is presented in this section. Yang et al [3] proposed a data transformation technique addressing security and privacy for data in cloud. The data is partitioned vertically and transformed using cryptographic primitives. The keys for transformation are kept at the private cloud and transformed data at public cloud. The transformation is not distance preserving and the method is not able to provide differential privacy to users. Kao et al [4] proposed a reversible privacy contrast mapping (RPCM) algorithm for data transformation.

Data is transformed by replacing two adjacent values by a new value. The mapping between adjacent values to new value is kept separately in private cloud. By grouping the values, anonymity is created among the records. But without consideration for distance preservation in transformation, the utility of data for data mining becomes infeasible. Yun et al [5] proposed a faster data perturbation algorithm using tree travel strategy.

A multi tier tree structure is built, which is able to transform a numeric attribute to another attribute. Though the retrieval efficiency is ensured, the differential privacy is not considered in this work. Zhang et [6] proposed a data transformation technique called as Cocktail.

This technique applied quasi identified partitioning with differential privacy strategy. The data transformation is lossless. But the retrieval efficiency is low in this approach. Zhou et al [7] proposed a data partition strategy. The strategy is application independent. The sensitive attributes are detected based on entropy with the identifier. The sensitive columns are kept in private cloud. The insensitive columns are moved to public cloud. Though retrieval efficiency is high in this approach, it distorts the data mining utilization. Lyu et al [8] transformed the data using repeated Gompertz (RP) followed by random projection (RP). The data is transformed to less dimension space with distance preserving property so that utility is not affected. But the retrieval efficiency is poor and it is not possible to execute any query operations on transformed data. Security is strong this approach. The transformed data is secure against estimation and component analysis attacks. Chen et al [9] proposed geometric data perturbation scheme. The geometric perturbation has three steps of rotation, translation and noise addition.

The mechanism has tighter security and privacy, bit retrieval efficiency is low for higher dimensional datasets. The same author in [10] proposed a random projection perturbation extension to geometric perturbation. The method is able to achieve faster geometric perturbation for multi dimensional datasets. Yuan et al [11] used

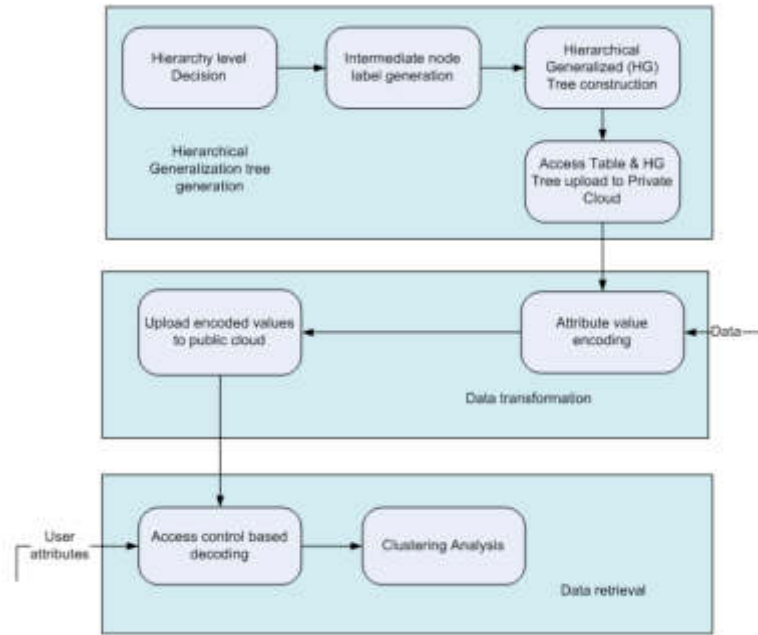


Fig. 2.1: Multi objective data transformation

compressive sensing based data transformation and fast indexing to improve retrieval efficiency. Due to violation of distance preservation, the transformed data becomes unsuitable for data mining operations like clustering and classification. Majeed et al [12] proposed a data transformation technique based on data anonymization. The attribute values in fixed interval are replaced with their averages. This method affects the original data and suitable only for certain data publishing requirements. Li et al [13] proposed two K-anonymity algorithm for data transformation. The data is transformed in way not to affect the classification capability. It is done by checking the entropy after transformation and choosing the level of transformation based on entropy. Begum et al [14] proposed data transformation scheme by removing sensitive items based on support and confidence values. The minimum number of items is removed in such a way to remove sensitivity.

Though the method is able to reduce the security leakages, it reduces the utility of data for data mining operations. Sridhar et al [15] clustered the data and passed the clustered data to geometric data perturbation. The solution considered security, privacy and utility preservation but did not consider retrieval efficiency. The solution did not consider differential privacy and query matching. Kodhai et al [19] proposed a secure fuzzy keyword search technique for data stored in cloud. But the technique cannot be used for the case of differential privacy and fine grained access controlled search problem considered in this paper work. Gheisari et al [20] used four different techniques of data micro aggregation, sampling, swapping and random noise to perturb data. But these schemes do not consider fine grained access control and differential privacy. Jafar et al [21] used public key cryptosystem for providing security to data but the scheme does not support differential privacy, fine grained access control and search over perturbed data. From the survey, it can be seen that most of the data transformation techniques focused on privacy and security, but they have not considered multi objectives of providing differential privacy, fine grained access control over data, retrieval over perturbed data and preservation of utility mining etc. In addition, the existing works have not considered efficiency in data transformation on hybrid cloud platform. This work considers this problem of multi objective data transformation and efficiency in data transformation for hybrid cloud environment.

3. Multi objective data transformation. The architecture of the proposed multi objective data transformation is given in Figure 2.1. As seen from figure, the proposed solution has three important functionalities: generation of hierarchical generalization tree, data transformation and retrieval. Each of the functionalities is

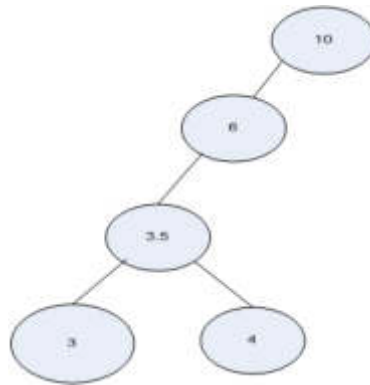


Fig. 2.2: Generalization tree for categorical variable

detailed below.

3.1. Generation of hierarchical generation tree. The data uploaded by data owner is in table format in which each column is an attribute. The attributes are marked as sensitive or in-sensitive by the data owner. For each of the sensitive attributes, a hierarchical generalization tree is constructed. This generalization tree is constructed to transform the data values of the corresponding attribute. The transformation using the generalization tree provides a differential view on the data after retrieval as desired by owner. Generalization is the key to the differential view. A sample generalization tree is shown in Figure 2.2. The attribute values are at the leaf nodes. The intermediate nodes in the tree are the generalization labels. It is not possible to construct a semantically correct generalization label for categorical variables as it requires domain knowledge. Thus data owner must provide the translation for each of the categorical variable in terms of numbers. The number must be provided in such way that if two categorical variables a and b are semantically close by degree d_1 , then the distance between their corresponding numerical variables must be in proportion to d_1 .

$$|N(a) - N(b)| \propto d_1 \tag{3.1}$$

Also if there is order in categorical variable, then if

$$a < b \text{ then } N(a) < N(b) \tag{3.2}$$

The generalization tree for attributes is constructed automatically by normalized the attribute values and repeated binary split till the maximum level allowed by data owner is achieved. The generalization tree is constructed in way to maximize the data mining utilization. The attribute value is first normalized in range of 0 to 1 from their actual value to decide the hierarchy. The normalization is done as

$$NV = \frac{AV}{(MaxV - Minv + 1)} \tag{3.3}$$

where NV is the normalized value, AV is the actual value, MaxV is the maximum value of the attribute and Minv is the minimum value of the attribute.

A Gaussian kernel density function is plotted with the normalized values. The minima of kernel density estimation for normalized values are taken the partitions as shown in Figure 3.2. In the Figure 3.2, minima of kernel is at $\langle 0, 0.4, 0.5, 0.9, 1 \rangle$. Thus 4 clusters need to created with values from (0 to 0.4), (0.4 to 0.5), (0.5 to 0.9), (0.9 to 1).

Once the normalized ranges are identified, they can again brought back to actual value.

$$AV = NV \times (MaxV - Minv + 1) \tag{3.4}$$

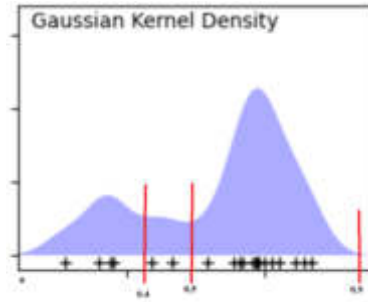


Fig. 3.1: Gaussian Kernel Density

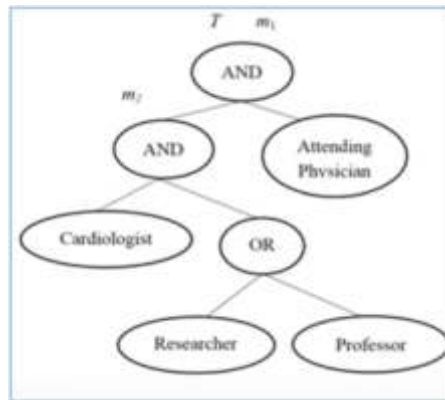


Fig. 3.2: Access tree

The cluster ranges are now in actual values. The mean of the values in those cluster is taken as next level generalization label. The generalization label is calculated for next level in same way till the maximum level specified by the owner is achieved. Once the generalization trees for the attributes are constructed, access control tree is provided by data owner for each level. A sample access tree for each generalization label is given in Figure 4.1.

The access tree is generated for each level based on the user attributes by data owner and access tree is uploaded to private cloud.

3.2. Data transformation. Each attribute value (V) is encoded into all its level generalization as $E_n = V, L_1(V), L_2(V), \dots, L_n(V)$ where $L_x(V)$ is the level x generalization of V and n is the number of levels. A homomorphic encryption (HE) key (k) is generated for each data owner. The encoding E_n is then homomorphically encrypted using the k, the encrypted E_n is given as

$$E(E_n) = HE(V, k), (HE(L_1)(V), k), \dots, (HE(L_n)(V), k) \tag{3.5}$$

The control to the particular level in the $E(E_n)$ is enforced using AHAC (Attribute based Hierarchical Access Control) CP-ABE (Cipher policy Attribute Based Encryption) [16]. The access tree for each level and $E(E_n)$ is passed as input to the AHAC CP-ABE encryption algorithm. The encryption algorithm provides a transformed $E(E_n)$ as output. The algorithm for transformation of attribute value is given in Algorithm 1.

Algorithm 1: Encoding

Input: attribute value (V) , HG tree of Attribute, HE key k, access control tree(T)

1. $E_n = V$

Table 4.1: Clustering accuracy

K	RG+RP [8]	GP [15]	SFAC-SHC [17]	Proposed
2	66.23	65.89	68.52	70.12
3	66.56	71.25	75.62	77.14
4	73.33	74.59	78.85	79.89
5	67.22	79.58	82.34	83.56

2. for $x=1$: num of level(HG)

$$En = En \cup HE(L_x(V), k)$$

3. EV AH ACCP-ABE. *Encrypt*(En,T) [16]

4. upload EV to public cloud

5. upload HG, k,T to private cloud

Each of the attribute value in the data is transformed using Algorithm 1. The transformed data is then uploaded to private cloud.

3.3. Data retrieval. When user requests the data for data mining utilities like clustering, the transformed must be decoded at first stage. The user attributes and transformed attribute value as input and provides the value at index of $E(En)$ corresponding the access provided for the user. Since the value is homomorphically encrypted, distance preservation is maintained and the data is suitable for data mining operations like clustering without any need for decryption. Since the $HE(L_x(V), k)$ is provided only for the level x matching the user access control credentials, it is difficult for user to learn any data characteristics beyond his access rights. The decoding algorithm for de-transformation of each attribute value in transformed data is given as Algorithm 2.

Algorithm 2: Decoding

Input: Query User (U).

Ouput: Decoded Value (DV).

1. $EV \leftarrow \text{downloadfromprivatecloud}$
2. $QV \leftarrow \text{Download user attributes from public cloud}(U)$
3. $DV \leftarrow AHACCP - ABE.Decrypt(EV, QV)$ [16]
4. return DV

Each of the attribute value in transformed data is decoded using Algorithm 2. This de-transformed data is then used for data mining utilities like clustering analysis.

4. Result. The performance of the proposed solution is tested against Arrhythmia dataset [18] in UCI machine learning repository. The performance is measured in terms of: (i) clustering accuracy (ii) data storage overhead (iii) retrieval efficiency and (iv) security against attacks. The performance of proposed solution is compared against geometric data perturbation (GP) [15], RG+RP [8] and searchable fine access control on secure hybrid clouds (SFAC-SHC) [17].

The clustering accuracy is calculated by measuring the differences between clusters of original and perturbed data. K-means algorithm is used for clustering the original and perturbed dataset.

The clustering accuracy is calculated as

$$ACC = \frac{1}{N} \sum_{i=1}^N | - Cluster_i(P) | - | Cluster_i(P') | \quad (4.1)$$

where P is the original data, is the transformed data, k is the number of clusters and N is the number of items in the dataset. The value of ACC is measured for different values of K and result is given in Table 4.1.

The value of ACC in proposed multi objective transformation is 9.3% higher compared to RG+RP, 4.87% higher compared to GP and 1.34% higher compared to SFAC-SHC. With the increase in K value, the ACC

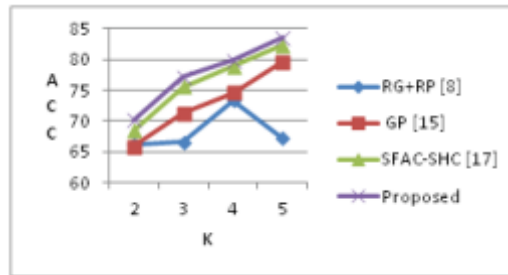


Fig. 4.1: Comparison of clustering accuracy

Table 4.2: Data upload time

Size (MB)	Upload time (sec)			
	Proposed	SFAC-SHC[17]	GP [20]	RG+RP [8]
20	13	14	19	22
40	18	24	35	38
60	24	43	61	64
80	30	78	112	122
Average	21.25	39.75	56.75	61.5

increases in the proposed solution. ACC value increases by 13.82 % for K value increase from 2 to 5. Distance preservation is ensured using Homomorphic encryption in proposed solution. This has increased the accuracy in the proposed solution.

The data storage time is measured as the time taken for transformation of data and uploading of transformed data to cloud. The data storage time is measured for various volumes of data and result is given in Table 4.2.

The storage time in proposed multi objective transformation is atleast 20.36% lower compared to GP and 36.93% lower compared to RG+RP. While storage time increases exponentially with increase in data volume in existing works, it is linear in proposed solution. This is because generalization tree construction and data transformation are linear operation in proposed solution.

The data retrieval efficiency is measured by varying the volume of data and the result is given in Table 4.3. The data retrieval efficiency is on average 29.95% lower compared to GP and 35.36% lower compared to RG+RP. Like storage time, retrieval time also increases exponentially with increase in data volume in existing works, but it is linear in proposed solution. The data de transformation using AHAC CP-ABE is linear in proposed solution and due to this retrieval time increases linearly with increase in data volume.

The security strength is measured in terms of difficulty in predicting the original data from perturbed data, provided the attacker has access to the perturbed data. The difficulty level is estimated in terms of measure called Variance of difference (VoD).

Let X_i be a random variable representing the column i , X'_i be the estimated result of X_i and difference $D_i = X'_i - X_i$. Let mean of D be $E(D_i)$ and variance be $Var(D_i)$. VOD for column i is $Var(D_i)$. VOD is measured for each column and average VOD is given as privacy measure(pm)

$$pm = \sum_{i=1}^N \frac{VOD_i}{N} \quad (4.2)$$

A guess is launched for 5 hours on the perturbed data and the privacy measure (pm) is measured for every 1-hour interval and plotted in Figure 4.4.

Higher the value of VoD, the effort to predict the original data is difficult. VoD in proposed solution is very high in proposed solution. It is almost twice compared to GP and RG+RP and it is on average 5% higher

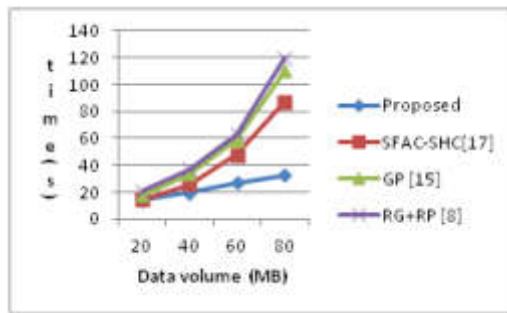


Fig. 4.2: Storage time

Table 4.3: Data retrieval time

Size (MB)	Upload time (sec)			
	Proposed	SFAC-SHC[17]	GP [20]	RG+RP [19]
20	14	15	18	21
40	20	26	34	37
60	27	48	59	63
80	33	87	110	120
Average	23.5	44	55.25	60.25

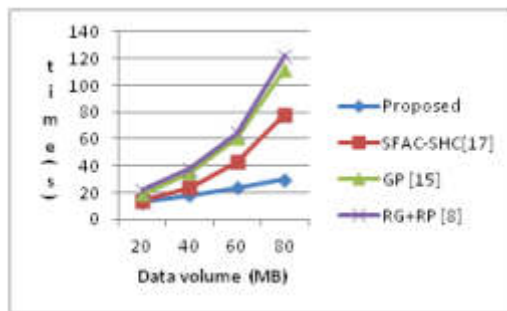


Fig. 4.3: Comparison of retrieval time

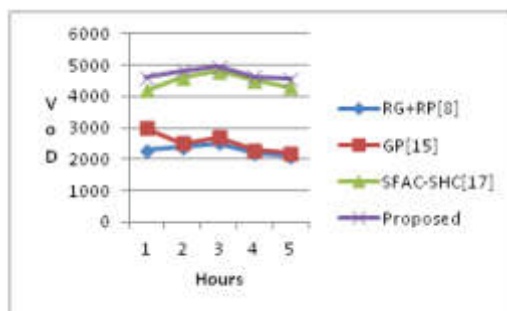


Fig. 4.4: VOD over time

compared to SFAC-SHC. Two level of encryption – first with Homomorphic encryption, followed by AHAC CP-ABE has made it difficult to predict the original data in the proposed solution.

5. Discussion. The data transformation technique proposed in this work addressed multiple objectives of privacy, security, differential access control, utility preservation and retrieval efficiency. The private data were perturbed using homomorphic encryption (HE) with access control over the keys for HE. The privacy and security was measured in terms of VOD and it is found that it is harder to infer any private information by launching brute force attacks. The security is higher by 5% compared to existing works. The higher security is due to differential privacy provided over data. Existing works used same keys without considering differential privacy. The proposed solution preserved distance based statistics in data even after perturbation. This has increased the accuracy for clustering operations on perturbed data by atleast 1.34% compared to existing works. Distance preservation in proposed solution is due to HE based transformation. The data retrieval efficiency is also higher in proposed solution as the transformation is light weight and easy to reverse the data. But existing works based geometric transformation had higher overhead during the stage of data retransformation. The proposed solution had fine grained access control over the data which was not considered in earlier works. The access control was using hierarchical access tree and the tree itself was secured by keeping it in private cloud. Thus the proposed solution performed better compared to existing works in terms of privacy, security, fine grained access control and retrieval efficiency.

6. Conclusion. A multi objective data transformation function for hybrid cloud is proposed in this work. The solution addressed multi objectives of privacy, security, differential access control, utility preservation and retrieval efficiency in data transformation specific to hybrid cloud environment. A generalized hierarchical tree is constructed from the data and data transformation is done based on generalization labels and access control rights for the users in the proposed solution. The proposed solution also provides differential privacy and access control to different users without affecting the utilization of data for data mining operations. The proposed solution provides atleast 5% higher data security, 29.95% higher data retrieval efficiency and 1.34% higher clustering accuracy over perturbed data compared to existing works. Adaption of solution for streaming data is in scope of future work.

REFERENCES

- [1] DAN YACHIN "l.ermetic.com/wp-idx-survey-results", 2021.
- [2] "https://info.flexera.com/CM-REPORT-State-of-the-Cloud'.
- [3] YANG, JI-JIANG, JIAN-QIANG LI, AND YU NIU., "A hybrid solution for privacy preserving medical data sharing in the cloud environment.", Future Generation computer systems 43 (2015): 74-86.
- [4] KAO, YUAN-HUNG, WEI-BIN LEE, TIEN-YU HSU, CHEN-YI LIN, HUI-FANG TSAI, AND TUNG-SHOU CHEN, "Data perturbation method based on contrast mapping for reversible privacy-preserving data mining.", Journal of Medical and Biological Engineering 35 (2015): 789-794.
- [5] YUN, UNIL, AND JIWON KIM., "A fast perturbation algorithm using tree structure for privacy preserving utility mining.", Expert Systems with Applications 42, no. 3 (2015): 1149-1165.
- [6] ZHANG, HONGLI, ZHIGANG ZHOU, LIN YE, AND XIAOJIANG DU. , "Towards privacy preserving publishing of set-valued data on hybrid cloud.", IEEE Transactions on cloud computing 6, no. 2 (2015): 316-329.
- [7] ZHOU, ZHIGANG, HONGLI ZHANG, XIAOJIANG DU, PANPAN LI, AND XIANGZHAN YU. , "Prometheus: Privacy-aware data retrieval on hybrid cloud" , In 2013 Proceedings IEEE INFOCOM, pp. 2643-2651. IEEE, 2013.
- [8] LYU, LINGJUAN, JAMES C. BEZDEK, YEE WEI LAW, XUANLI HE, AND MARIMUTHU PALANISWAMI., "Privacy-preserving collaborative fuzzy clustering.", Data & Knowledge Engineering 116 (2018): 21-41.
- [9] CHEN, KEKE, GORDON SUN, AND LING LIU. , "Towards attack-resilient geometric data perturbation.",In proceedings of the 2007 SIAM international conference on Data mining, pp. 78-89. Society for Industrial and Applied Mathematics, 2007.
- [10] CHEN, KEKE, AND LING LIU., "Geometric data perturbation for privacy preserving outsourced data mining.", Knowledge and information systems 29 (2011): 657-695.
- [11] YUAN, XINGLIANG, XINYU WANG, CONG WANG, JIAN WENG, AND KUI REN, "Enabling secure and fast indexing for privacy-assured healthcare monitoring via compressive sensing.", IEEE Transactions on Multimedia 18, no. 10 (2016): 2002-2014.
- [12] YUAN, XINGLIANG, XINYU WANG, CONG WANG, JIAN WENG, AND KUI REN. , "Enabling secure and fast indexing for privacy-assured healthcare monitoring via compressive sensing.", IEEE Transactions on Multimedia 18, no. 10 (2016): 2002-2014.
- [13] LI, JIUYONG, JIXUE LIU, MUZAMMIL BAIG, AND RAYMOND CHI-WING WONG., "Information based data anonymization for classification utility.", Data & Knowledge Engineering 70, no. 12 (2011): 1030-1045.
- [14] SABIN BEGUM, R., AND R. SUGUMAR., "Novel entropy-based approach for cost-effective privacy preservation of intermediate datasets in cloud." , Cluster Computing 22, no. Suppl 4 (2019): 9581-9588.

- [15] REDDY, VULAPULA SRIDHAR, AND BARIGE THIRUMALA RAO., "A Combined Clustering and Geometric Data Perturbation Approach for Enriching Privacy Preservation of Healthcare Data in Hybrid Clouds." ,International Journal of Intelligent Engineering & Systems 11, no. 1 (2018).
- [16] HE, HENG, LIANG-HAN ZHENG, PENG LI, LI DENG, LI HUANG, AND XIANG CHEN., "An efficient attribute-based hierarchical data access control scheme in cloud computing." ,Human-centric Computing and Information Sciences 10 (2020): 1-19.
- [17] VULAPULA, SRIDHAR REDDY, AND SRINIVAS MALLADI., "Attribute-Based Encryption for Fine-Grained Access Control on Secure Hybrid Clouds." ,International Journal of Advanced Computer Science and Applications 11, no. 10 (2020).
- [18] , "<https://archive.ics.uci.edu/ml/datasets/Arrhythmia>" .
- [19] PACHIPALA, YELLAMMA, AND JAFAR ALZUBI., "Managing the cloud storage using deduplication and secured fuzzy keyword search for multiple." ,International Journal of Pure and Applied Mathematics 118, no. 14 (2018): 563-565.
- [20] GHEISARI, MEHDI, HAMID ESMAEILI NAJAFABADI, JAFAR A. ALZUBI, JIECHAO GAO, GUOJUN WANG, AAQIF AFZAAL ABBASI, AND ANIELLO CASTIGLIONE. , "OBPP: An ontology-based framework for privacy-preserving in IoT-based smart city." ,Future Generation Computer Systems 123 (2021): 1-13.
- [21] ALZUBI, JAFAR A., RAMACHANDRAN MANIKANDAN, OMAR A. ALZUBI, ISSA QIQIEH, ROBBI RAHIM, DEEPAK GUPTA, AND ASHISH KHANNA., "Hashed Needham Schroeder industrial IoT based cost optimized deep secured data transmission in cloud." ,Measurement 150 (2020): 107077.

Edited by: Anil Kumar Budati

Special issue on: Soft Computing and Artificial Intelligence for wire/wireless Human-Machine Interface

Received: Dec 31, 2023

Accepted: Apr 8, 2024



OPTIMIZING TASK SCHEDULING: EXPLORING ADVANCED MACHINE LEARNING IN DEW-POWERED CLOUD ENVIRONMENTS

A. GANESH*, K SREE DIVYA[†], CHINTHAKUNTA SASIKALA[‡], E. POORNIMA[§], NIDAMANURU SRINIVASA RAO,[¶]
A.V.L.N SUJITH^{||} AND G.RAMESH**

Abstract. Research into Dew computing environments has recently emerged as a result of the increasing prevalence and processing power of mobile and IoT devices. In these settings, even low-powered devices can share some of their computational resources with their neighbors. This paper proposes a novel approach to workflow scheduling in dew enabled cloud computing environment, called Deep Q-learning (DQN) + Chronological Geese Migration Optimization (CGMO). DQN is a deep learning-based method for scheduling workflows, while CGMO is a hybrid optimization algorithm that combines the chronological idea and the Wild Geese Migration Optimization (GMO) algorithm. The proposed approach aims to optimize multiple objectives, including predicted energy, Quality of Service (QoS), and resource usage, by scheduling workflows in the cloud. The approach also takes into account the current state of the Virtual Machine (VM) and the job. The assessment measures employed for DCGM include maximum QoS, minimum energy usage, and maximum resource utilization. The results show that DCGM achieved the highest QoS (0.865), lowest energy usage (0.0322), and highest resource utilization (1.000) compared to other approaches.

Key words: Dew computing, deep learning, task scheduling

1. Introduction. In recent years, resource-constrained devices have been under severe computational strain due to the meteoric rise of computationally expensive jobs in mobile applications. Smartphones and Internet of Things devices, unfortunately, typically fall short of such requirements. Dew computing is a solution that offers moving computationally expensive tasks to more capable (nearby) machines [1, 2, 3, 4, 5]. A device is considered nearby if it is on the same local network as another device. The concept is that you may use your phone to delegate tasks to your computer, saving battery life on both devices. However, learning how to efficiently divide tasks across adjacent devices is a major hurdle for Dew computing to be useful in practice. The issue of scheduling tasks in Dew settings is investigated in this paper. The term "Dew environment" refers to a network of interconnected gadgets. Different devices may have different storage capacities, sensor arrays, processing speeds, wireless connectivity options, and battery life. Furthermore, consumers may engage with various gadgets at various moments. All these considerations are necessary for efficient task distribution in a Dew setting. Existing solutions for distributing work in Dew settings adhere to human-designed policies. By adhering to a predetermined set of rules, these policies attempt to distribute the workload evenly across the devices. The Simple Energy-Aware Scheduler (SEAS) [6], the Batch Processing Algorithm (BPA) [7], and the Round Robin (RR) [8] are all examples of such schedulers. Due to their inflexibility, these approaches consistently produce bad decisions and wasteful resources when applied to a Dew context.

In this study, we suggest utilizing RL to figure out how to allocate tasks in a Dew ecosystem [9]. Learning optimal behavior by interaction with an environment is the focus of RL, a branch of AI that investigates how

*Department of CSE, Sri Venkateswara College of Engineering, Tirupati, Andhra Pradesh, India, (achari.ganesh@gmail.com)

[†]Department of Computer Science & Technology, Madanapalle Institute of Technology and Science, Madanapalle, Andhra Pradesh, India. (divya.kpn@gmail.com)

[‡]Department of Computer Science and Engineering, Srinivasa Ramanujan Institute of Technology (Autonomous), Ananthapuram, Andhra Pradesh, India, (sasikalareddy27@gmail.com)

[§]Department of CSE (AI& ML), Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India (poornimacse561@gmail.com)

[¶]Department of CSE, Narsimha Reddy Engineering College, Secunderabad, Telangana State, India (rao75nidamanuru@gmail.com)

^{||}Department of CSE, Narsimha Reddy Engineering College, Secunderabad, Telangana State, India (sujeeth.avln@gmail.com)

**Department of CSE, Gokaraju Rangaraju Institute of Engineering & Technology, Hyderabad, India (ramesh680@gmail.com)

to create such creatures. Every action the agent takes results in some form of reward signal that it then tries to maximize. This is done by the agent learning from its past mistakes and adjusting its present policy (a mapping from observations to actions). Research in fields as diverse as robotics [10], conversational agents [11], and drug discovery [12] have all shown success using RL agents to address complicated decision-making problems powered by deep learning. We propose letting an RL agent figure out how to divide up work in a Dew system through experience. The use of RL for task distribution in Edge and Cloud computing has been investigated in previous research [13, 14, 15, 16]. However, no one has yet implemented RL on a Dew system. The unique difficulties of Dew computing cannot be compared to those of Cloud or Edge computing. Some gadgets in Dew computing, for instance, may experience battery drain or user interaction. Further, it has not been investigated in prior publications if RL agents may acquire rules that generalize well to novel contexts. Our results demonstrate that, when tested in novel scenarios, deep RL agents outperform state-of-the-art heuristic approaches in their ability to learn to offload tasks.

Workflow, a relatively new technology, has been widely used to keep tabs on the best apps' performance. In the context of scientific disciplines, a workflow is the collection of individual actions that are interconnected by data [7, 8]. Workflow applications have expanded in recent years to include domains as diverse as e-commerce, biology, astronomy, and physics. The workflow's tasks, in general, need time- and resource-intensive operations to be carried out effectively [9]. Multi-objective scheduling-based techniques are broken down into two distinct types: QoS-inhibited algorithms and QoS-optimization algorithms [10]. Several approaches rely on QoS constraints to transform this problem into a single-objective optimization problem [2]. In order to influence QoS, extensive procedures are often designed and implemented in widely disseminated large-scale evaluation settings. How to coordinate the jobs has been a focus of the inquiry for a while now [9, 11]. Workflow scheduling solely considered the amount of investigations based on price and timeliness. When asked about how to reduce costs without sacrificing quality of service, they often recommend something called multi-objective scheduling, which is used to derive the pricing. The precaution was included into the workflow scheduling of certain recent studies based on cloud infrastructure. The impact of communications between tasks and virtual machines on cloud security as a whole was not adequately examined [3].

The allocation of jobs to relevant resources in CC [12] may either be insufficient or optimized to meet the needs of users depending on QoS. Cost, time, security, load, and success rate are all integrated into the QoS requirement based on process scheduling [13]. Nonetheless, many studies in this field have only optimized a few or a few QoS factors. A great deal of planning, plans, and improvements went into the framework of cloud security, all of which work together to guarantee the safety of data stored in cloud architectures [12]. The researchers developed many workflow scheduling strategies in which the technical word was associated with CC [5]. Due to the importance of the applications, a large number of studies have been conducted to develop a model for workflow management in clouds in line with scheduling schemes. These studies include Condor Dagman [14], Gridbus toolkit [15], Icenl [16], and Pegasus [17]. By hiding their orchestrations and implementations, the aforementioned structures may be seen as a kind of platform service that aids the network and cloud-based computerization of technical and commercial applications [4]. Workflow scheduling uses data mining and regression based methods. In addition, researchers often try new approaches to workflow scheduling in an attempt to fix the issues that have plagued the ones they've used before. Due to its easy convergence qualities, flexibility, and error-tolerance abilities, the Firefly algorithm stands out as the best of all the swarm-based models used in workflow scheduling [5].

In dew-powered cloud environments, efficiently scheduling microservice-enabled tasks is crucial for optimizing resource utilization and minimizing energy consumption. Traditional scheduling methods often fall short in adapting to the dynamic nature of these environments. To address this challenge, we propose a novel machine learning algorithm tailored for task scheduling. Our algorithm leverages advanced predictive models to dynamically allocate resources based on workload patterns and environmental conditions, thereby improving system performance and energy efficiency. This paper presents a comprehensive discussion on the applicability and benefits of our approach, bridging the gap between theoretical concepts and practical implementation in dew-powered cloud environments. In this paper, we provide a new framework, DQN+CGMO, based on multi-objective and DL for scheduling cloud-based workflows. In this study, we focus on two processes: multi-objective workflow scheduling and workflow scheduling using DL. Predicted energy, QoS, and resource consumption, ac-

tual task running time, bandwidth utilization, memory capacity, makespan equivalent of total cost, and task priority are all taken into account to determine the fitness of workflow scheduling in multi-objective-based workflow scheduling. Workflow scheduling is done with consideration for CGMO, which is derived from the combination of the GMO algorithm with the chronological idea. At the same time, DQN is fed the incoming task in the DL-based workflow scheduling process. Workflow scheduling also takes into account VM and task settings in real time. In the end, the combined product is what we get.

2. Literature Review. The concept of EMO (Evolutionary Multi-objective Optimization) was created by Zhu, Z., et al. [2]. This approach was motivated by the Amazon EC2 on-demand instance kinds. If only we could simply include IaaS's resourcing choices and cost estimate tools into the suggested paradigm! The term "regressive" was coined by the researchers G. Narendrababu Reddy and S. Phani Kumar [5]. RWO optimization (Whales optimization). The RWO model maximized resource efficiency with minimum expenditure of time, money, and energy. The original presentation of the Whale optimization algorithm (WOA) may be found in the work of Strumberger (I.), et al. [18]. In this work, we improve the implementation of the unique whale optimization and explore various approaches to solving CC's resource scheduling problem. Stephanakis, I.M., et al. [19] developed particle swarm optimization (PSO). It was a dividing line in the scheme's design, and it helped many individuals in the population agree on the best course of action.

A new paradigm called as Mobile Edge Computing has arisen to deal with the latency and network traffic problems that plague Mobile Cloud Computing systems [23]. Khan et al. [24], stating that "Edge Computing" refers to "a model that allows a cloud-based computing capacity providing services making use of the infrastructure that is on the edge of the network." Mobile Edge Computing is the practice of deploying more robust applications by using local servers or workstations inside a network to minimize latency and maximize data processing efficiency. This paradigm allows for the implementation of more reliable and efficient applications [25, 26] by enabling Edge servers to collaborate with neighboring nodes or with Cloud services. Despite the fact that Edge computing helps to ease network challenges, the network backbone may not be accessible or available while working with IoT devices in, say, mines and on ships, in deserts, or when traveling.

In dew computing, nearby machines on a network take turns processing data. This suggested design [27] aims to reduce network latency, the energy cost of transferring data over great distances, and the expenses associated with utilising Cloud infrastructure. Dew computing does this by enhancing functionality in two fundamental areas of mobile and IoT device performance. To begin, it treats mobile devices as clients inside the network architecture, offloading their duties to other devices on the same network. Second, in Dew computing, mobile and IoT devices are seen as resources that may be exploited to improve the performance of the existing system. This method enables one device in the network to use the capabilities of another (including mobile and IoT devices) to accomplish a goal [29,30, 31]. Ramesh et al. [32] explored machine learning algorithms with conventional network defense mechanisms offers a proactive and flexible strategy to protect systems from Distributed Denial of Service (DDoS) attacks, ensuring robust security measures that can adapt to evolving threats in real-time.

Jia, Y.H., et al. [9] developed the method now known as Ant Colony Optimization (ACO). This strategy received a flawless score in both the cost and efficiency categories. Time was lost because precise methods were not documented, even though they were used by numerous workflow programs. F. Abazari et al. created a method called multi-objective workflow scheduling (MOWS). The strategy's better security and risk in a wide range of workload characteristics led to an overall improvement in the building's integrity. Additionally, it demonstrated that the developed system successfully predicted attacks in cloud environments. A significant advancement in both cost and energy utilization was achieved by the development of Dynamic Voltage and Frequency Scaling with Multi-objective Discrete Particle Swarm Optimization (DVFS-MODPSO) by Yassa, S., et al. [4]. However, issues of trust and safety were not addressed. Kakkottakath Valappil Thekkepurayil et al. presented Ant Lion optimization (ALO) in [12]. ALO had a great convergence rate and was very searchable. However, there was a lack of multi-cloud collaboration integration in various clouds, therefore just one service was provided.

3. System Model. With the help of CC, the on-demand services are furnished on the basis of the necessities of user where the operation will be conducted. Therefore, numerous applications of workflow in CC may be operated with the available on-demand is the critical task in the scheduling of workflow. An

Table 2.1: Literature Analysis

Optimization Technique	Origin	Motivation/Application	Limitations/Gaps
Evolutionary Multi-objective Optimization (EMO)	Zhu, Z., et al. [2]	Inspired by Amazon EC2 on-demand instance kinds; aims to optimize resourcing choices and cost estimates in cloud environments.	Limited documentation on precise methods used; potential gap in methodological transparency.
Whale Optimization Algorithm (WOA)	Strumberger (I.), et al. [18]	Focuses on maximizing resource efficiency with minimal expenditure of time, money, and energy; applied to solve cloud computing's resource scheduling problem.	Lack of consideration for multi-cloud collaboration integration; limited to single-service provisioning.
Particle Swarm Optimization (PSO)	Stephanakis, I.M., et al. [20]	Facilitates collaboration among individuals in the population to determine the best course of action; used for optimization tasks in cloud environments.	May face challenges in handling complex optimization landscapes; potential for premature convergence.
Ant Colony Optimization (ACO)	Jia, Y.H., et al. [9]	Achieves high scores in cost and efficiency categories; applied to workflow scheduling problems in cloud environments.	Limited documentation on the precise implementation details; potential for replication issues.
Multi-objective Workflow Scheduling (MOWS)	F. Abazari et al.	Improves security and risk management while enhancing building integrity; effectively predicts attacks in cloud environments.	Limited consideration for trust and safety aspects; potential gaps in addressing security vulnerabilities.
Dynamic Voltage and Frequency Scaling with Multi-objective Discrete Particle Swarm Optimization (DVFS-MODPSO)	Yassa, S., et al. [4]	Enhances cost and energy utilization efficiency; addresses challenges in cloud computing resource management.	Does not address trust and safety concerns; potential limitations in real-world scalability and applicability.
Ant Lion Optimization (ALO)	Thekkepurayil et al.	Demonstrates high convergence rate and searchability; primarily focused on optimization tasks with single-cloud services.	Limited integration for multi-cloud collaboration; potential gaps in addressing diverse service requirements.

effectual scheduler should be required for the proper scheduling with the consideration of available resources. In addition to that, in order to process the sensitive data to control the safety and reliability of data sharing is the significant part in the secure infrastructure. Hence, the finest scheduling of workflow in the CC is designed with the advancement of safety measures that is illustrated in figure 3.1. The pre-processing data in CC has physical machines (PM) and VM, which offers the service to the user on the basis of demand. the scheduling of workflow is utilized finely in the presented model on the basis of multi-objective function in accordance with six attributes namely, resource utilization, QoS, bandwidth utilization, makespan equivalent of total cost, predicted energy, and memory capacity.

Consider the PMs employed in CGMO is illustrated by,

$$P = \{P_1, P_2, \dots, P_u, \dots, P_v\} \quad (3.1)$$

The Directed Acyclic Graph (DAG) is employed for the illustration of the scheduling of workflow and it is expressed as $Q = (G, L)$. The factor G signifies the task equivalent to the workflow that is computed as,

$$G = \{G_1, G_2, \dots, G_K, \dots, G_J\} \quad (3.2)$$

4. Proposed DQN+CGMO. The fundamental purpose of this study is to develop DQN+CGMO, a multi-objective and DL-based hybrid optimization for workflow scheduling in CC. In this setup, CGMO sched-

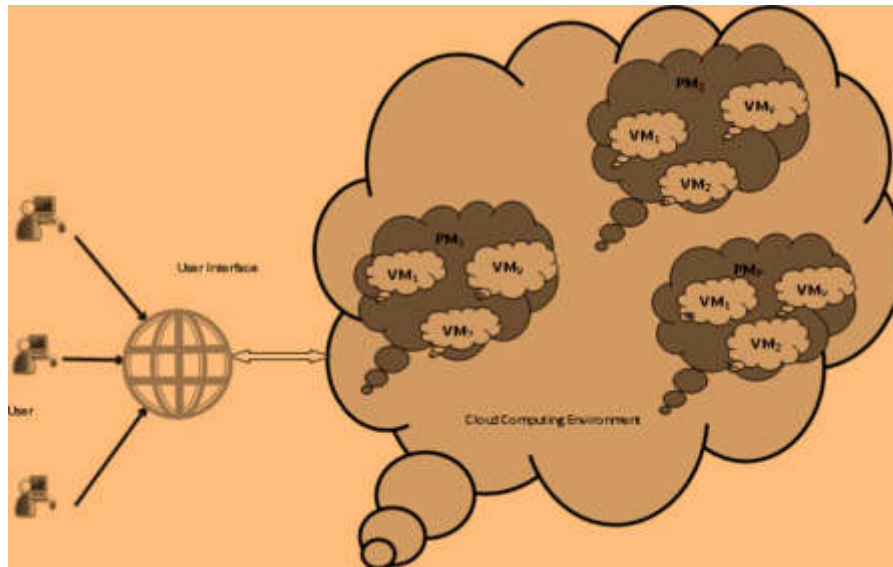


Fig. 3.1: Systematic View of workflow scheduling in cloud within the interval.

ules workflows based on many objectives, whereas DQN schedules workflows using deep learning. Predicted energy, QoS, and resource consumption, actual task running time, bandwidth utilization, memory capacity [20], makespan equivalent of total cost [21], and task priority are all used to determine the fitness of workflow scheduling in multi-objective-based workflow scheduling. CGMO, a hybrid of the GMO algorithm and the chronological idea, is used to schedule workflows [22]. The incoming job is also supplied to DQN [23] in the DL-based workflow scheduling process. Workflow scheduling also takes into account real-time VM and task characteristics. Here, both the tasks' parameters, such as average computation cost, earliest start time, earliest finish time, duration, and priority, and the VMs' parameters, including bandwidth utilization, memory utilization, capacity, and central processing unit (CPU), are taken into account. Last but not least, the total production is what matters most. Schematic representation of CGMO for DQN-based multi-objective workflow scheduling is shown in Figure 2. Combining Deep Q-learning (DQN) with Chronological Geese Migration Optimization (CGMO) could yield an innovative approach for optimizing task scheduling in dew-powered cloud environments. Here are the steps involved in this hybrid algorithm:

Initialization

- Set up the environment with historical data on task execution, resource utilization, and environmental conditions.
- Initialize the DQN and CGMO components of the algorithm.

Data Preprocessing

- Normalize and preprocess the input data to ensure consistency and remove noise.
- Feature engineering may be applied to extract relevant information, such as workload patterns and environmental factors.

Deep Q-learning (DQN) Exploration

- Utilize the DQN component to explore the state-action space and learn optimal task scheduling policies.
- Train the DQN model using historical data to estimate the Q-values, which represent the expected future rewards for taking specific actions in given states.
- Employ techniques such as experience replay and target network updates to stabilize training and improve convergence.

Chronological Geese Migration Optimization (CGMO)

- Implement the CGMO component inspired by the behavior of geese migrating in chronological order.

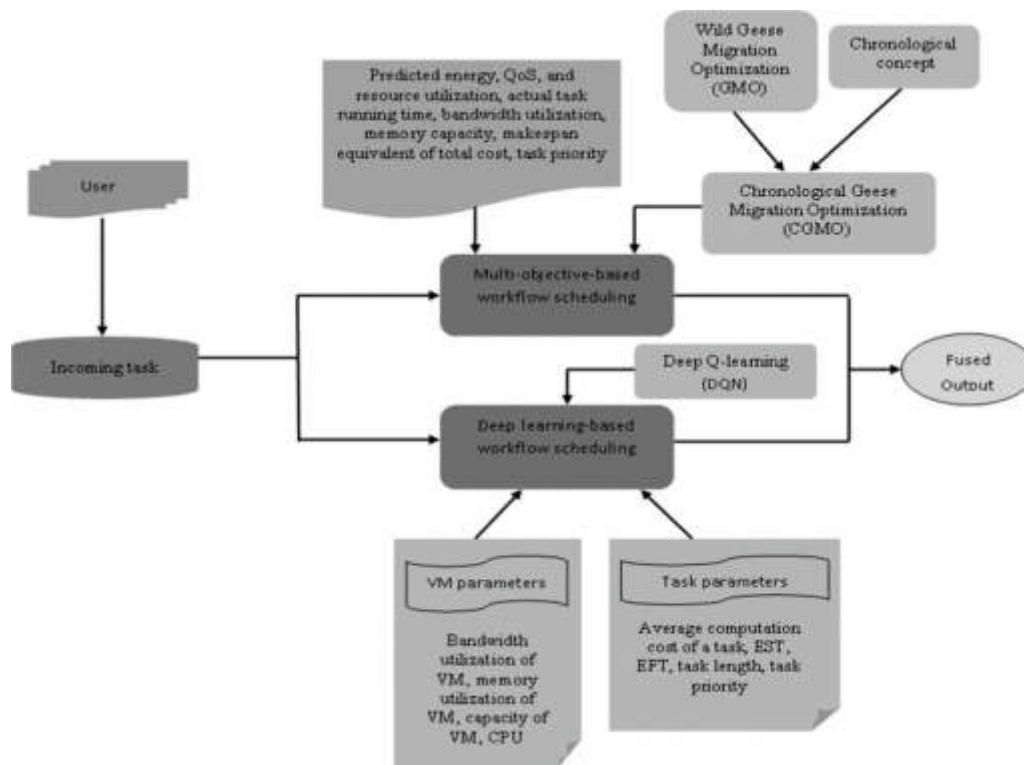


Fig. 4.1: Modeled view of CGMO for multi-objective and DQN based workflow scheduling

- Define migration paths as potential solutions to the task scheduling problem, where each path represents a sequence of task assignments over time.
- Use CGMO to iteratively update and optimize the migration paths based on historical patterns and environmental conditions.
- Leverage the collective intelligence of the population of "geese" (i.e., candidate solutions) to guide the search towards promising regions of the solution space.

Hybridization and Policy Integration

- Integrate the learned policies from the DQN with the optimized migration paths generated by CGMO.
- Combine the strengths of both algorithms to adaptively adjust task scheduling decisions in response to changing conditions and uncertainties.

Evaluation and Fine-Tuning

- evaluate the hybrid algorithm's performance using simulation or real-world experiments.
- Fine-tune the algorithm parameters and hyperparameters based on performance metrics such as resource utilization, energy efficiency, and task completion time.

Deployment and Continuous Learning

- Deploy the optimized scheduling algorithm in the dew-powered cloud environment.
- Monitor system performance and collect feedback data to facilitate continuous learning and adaptation.
- Update the algorithm periodically to accommodate evolving workload patterns and environmental dynamics.

4.1. Multi-objective based task scheduling. The system of task scheduling is significant for the enhancement of resources and the effectualness of server and also by improving the assessment of the analyzed nodes. In the multi-objective task set, the numerous tasks of multi-objective are defined time during the time processing of CC. In accordance with the execution time, the entire tasks of multi-objective on VM evaluate

SL. NO.	Pseudo Code of CGMO
1.	Input: Population size M , maximum iteration Max_{iter}
2.	Output: A_j^{r+1}
3.	Begin
4.	Initialize the population
5.	for $r = 1: Max_{iter}$
6.	for $n = 1: K$
7.	for $j = 1: M$
8.	Generate the migration groups using Eq. [17]
9.	end for
10.	end for
11.	if $rand > 0.5$
12.	for $n = 1: K$
13.	for ($j = a * (n - 1) + 1: (a * n)$)
14.	Upgrade the new solution employing Eq. [24] and Eq. [25]
15.	end for
16.	end for
17.	else
18.	for $n = 1: K$
19.	for ($j = a * (n - 1) + 1: (a * n)$)
20.	Evaluate free foraging utilizing Eq. [26]
21.	end for
22.	end for
23.	else
24.	Compute the radius of migration group by Eq. [27]
25.	end for
25.	Return
25.	Terminate

the load balance difference of task scheduling of multi-objective and create the intent function of CC scheduling of multi-objective task.

4.2. Deep Learning based task scheduling. Recurrent Neural Networks (RNNs) and long short-term memories (LSTMs) (RNN-LSTMs) are used to process incoming tasks for DL-based task scheduling. VM and task settings will determine which job is chosen. Let's assume for the time being that the task's computing cost, EST, EFT, duration, and priority are all fixed. Memory, CPU, and ram are the virtual machine settings.

5. Comparative Discussion. In this section, the effectiveness of the provided technique is shown in relation to earlier models. Maximum QoS (0.785), minimum energy usage (0.022), and maximum resource utilization (1.000) were all reached using the assessment methods used for DQN+CGMO. Table 5.1 presents the results of a comparative analysis of DQN and CGMO.

5.0.1. Deep Learning based task scheduling. Recurrent Neural Networks (RNNs) and long short-term memories (LSTMs) (RNN-LSTMs) are used to process incoming tasks for DL-based task scheduling. VM and task settings will determine which job is chosen. Let's assume for the time being that the task's computing cost, EST, EFT, duration, and priority are all fixed. Memory, CPU, and ram are the virtual machine settings.

In DQN [23] the basic procedure is to allow the initial state to the Neural network (NN). Moreover, NN precedes the entire tasks at the output stage. When the current value and deep-TD target values are similar, the upgrade regulations of

$J(Z, \psi)$ is not able to upgrade the values. Therefore, $J(Z, \psi)$ congregates to the original values and observes the preferred objective.

6. Results and Discussion. This section discusses not just the end result of DQN+CGMO, but also the experimental setup, simulation settings, metrics used, and comparative evaluation. The quality of service (QoS), energy consumption (EC), and resource utilization (RU) are the metrics used by the proposed DCGM

Table 5.1: Comparative Discussion (*DQN+CGMO)

Dataset	Metrics	RWWO_D MN+DRL	RWO +DRL	WOA +DRL	PSO+ DRL	ACO +DRL	RWWTD O+DRL	Proposed*
PM=5, VM=10	QoS	0.174	0.273	0.501	0.529	0.574	0.673	0.750
PM=5, VM=10	Energy Consumption	0.485	0.399	0.043	0.040	0.039	0.038	0.035
PM=5, VM=10	Resource utilization	0.947	0.960	0.961	0.964	0.974	0.975	0.979
PM=10, VM=15	QoS	0.179	0.279	0.526	0.538	0.579	0.679	0.761
PM=10, VM=15	Energy Consumption	0.489	0.405	0.045	0.040	0.037	0.0297	0.022
PM=10, VM=15	Resource utilization	0.953	0.961	0.963	0.975	0.977	0.980	0.989
PM=15, VM=20	QoS	0.194	0.294	0.525	0.550	0.589	0.693	0.771
PM=15, VM=20	Energy Consumption	0.507	0.408	0.047	0.041	0.038	0.030	0.021
PM=15, VM=20	Resource utilization	0.963	0.968	0.970	0.971	0.980	0.983	0.987
PM=20, VM=25	QoS	0.204	0.305	0.529	0.565	0.599	0.704	0.785
PM=20, VM=25	Energy Consumption	0.508	0.420	0.048	0.042	0.040	0.031	0.22
PM=20, VM=25	Resource utilization	0.968	0.997	0.997	0.998	1.000	1.000	1.000

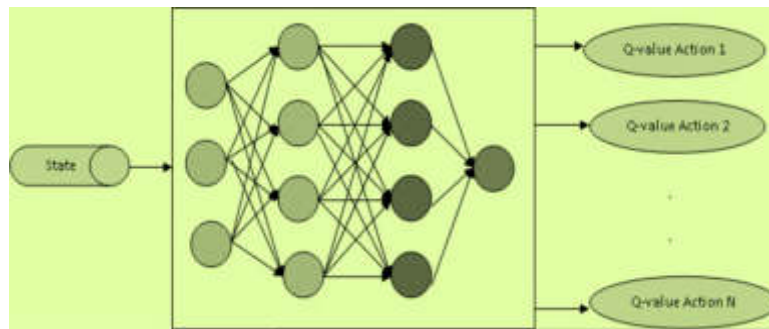


Fig. 5.1: Structure of DQN

method. The RWWO_ DMN+DRL [27] [28], RWO+DRL[5], WOA+DRL[18], PSO+DRL[19], ACO+DRL[9], and RWWTDO+DRL [3] comparison methods are used for DQN+CGMO.

The assessment of DCGM for varied job sizes with PM=5, VM=10, is shown in Figure 6.1,6.2,6.3. The DCGM based on QoS are discussed in Figure 6.1. While the previous approaches, namely RWWO_ DMN+DRL, RWO+DRL, WOA+DRL, PSO+DRL, ACO+DRL, and RWWTDO+DRL, obtained 0.174, 0.273, 0.501, 0.529, 0.574, and 0.673, the DCGM achieved the QoS of 0.750 when the task size=400. Figure 6.2 shows how much energy DCGM uses. The classic techniques such as RWWO_ DMN+DRL had 0.485, RWO+DRL had 0.399, WOA+DRL had 0.043, PSO+DRL had 0.040, AC+DRL had 0.039, and RWWTDO+DRL had 0.038 with 400 as the energy consumption if the DCGM obtained 0.035. Figure 6.3 explains the DCGM in terms of resource consumption. The traditional techniques achieved 0.947, 0.960, 0.961, 0.964, 0.974, 0.975 for RWWO_ DMN+DRL, RWO+DRL, WOA+DRL, PSO+DRL, ACO + DRL, and RWWTDO +DRL, while the DCGM earned 0.979 based on resource utilization with a work size of 400.

6.1. Assessment of DCGM based on task size with PM=20, VM=25. Figure 6.4,6.5,6.6 shows the worth of DCGM for different work sizes (PM=20, VM=25). Figure 6.4 depicts the assembly of a DCGM supporting QoS. When the task size was 400, the gains for the conventional methods were 0.204, 0.305, 0.529, 0.565, 0.599, and 0.704, respectively. The DCGM reached an all-time high of 0.785. The use of energy by DCGM is shown in Figure 6.5. For comparison, the energy consumption results for the RWWO_ DMN+DRL, RWO+DRL, WOA+DRL, PSO+DRL, ACO+DRL, and RWWTDO+DRL were 0.508, 0.420, 0.048, 0.042, 0.040, and 0.031, respectively, for a task size of 400. The DCGM achieved 0.022. Figure 6.6 depicts the typical approaches next to the DCGM with resource utilization at 1.000, including RWWO_ DMN+DRL = 0.968, RWO+DRL = 0.997, WOA+DRL = 0.997, PSO+DRL = 0.998, ACO+DRL = 1.000, and RWWTDO+DRL

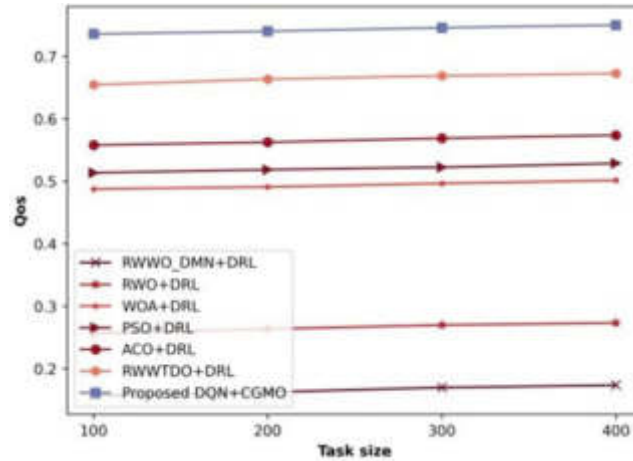


Fig. 6.1: Valuation of DCGM altering task size with PM=5, VM=10 QoS

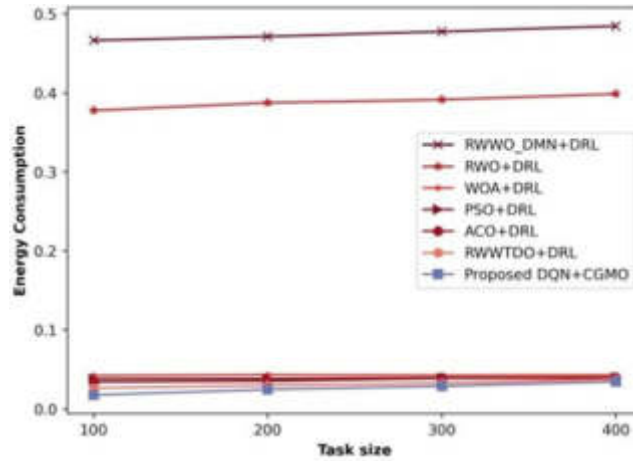


Fig. 6.2: Valuation of DCGM altering task size with PM=5, VM=10 Energy Consumption

= 1.000.

7. Comparative Discussion. In this section, the effectiveness of the provided technique is shown in relation to earlier models. Maximum QoS (0.785), minimum energy usage (0.022), and maximum resource utilization (1.000) were all reached using the assessment methods used for DQN + CGMO. Table 1 presents the results of a comparative analysis of DQN and CGMO.

7.1. Applications of Proposed algorithm in Real-world environment: An illustrative Scenario.

Envision a sophisticated city infrastructure with numerous IoT sensors used to monitor several elements of urban life, including traffic flow, air quality, and energy use. Various technologies, including as sensors on lampposts and smart meters in residences, produce large volumes of data that must be processed and evaluated immediately to allow city officials to make well-informed decisions. Dew computing is a promising approach that utilizes the computational capabilities of IoT devices to analyze data closer to the source, hence decreasing latency and bandwidth usage. Scheduling workflows effectively in a changing context is a major difficulty. Introduce the suggested method, DQN + CGMO, which aims to optimize workflow scheduling in Dew-enabled cloud computing settings. Let’s examine a real-life situation: To enhance traffic flow by evaluating live data

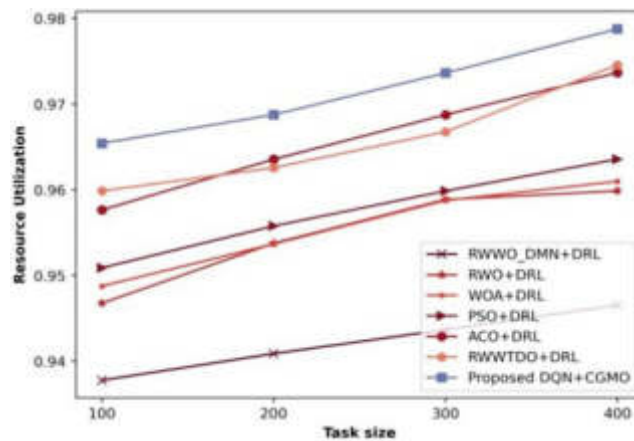


Fig. 6.3: Valuation of DCGM altering task size with PM=5, VM=10 Resource utilization

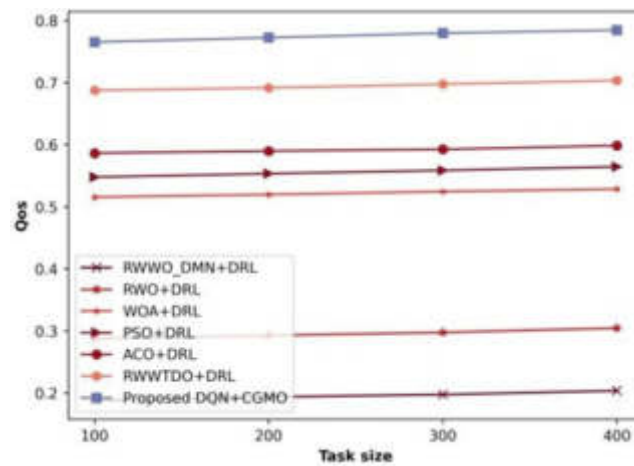


Fig. 6.4: Valuation of DCGM altering task size with PM=20, VM=25 QoS,

from traffic cameras, vehicle sensors, and GPS systems in automobiles.

- Data Collection: Traffic data is gathered from a variety of IoT devices located across the city, such as traffic cameras, car sensors, and GPS devices.
- Data Processing: The gathered data must undergo processing to identify traffic congestion, recognize traffic trends, and forecast traffic flow in various metropolitan locations.
- Workflow scheduling utilizes the DQN + CGMO algorithm. It efficiently organizes processes by taking into account aspects like anticipated energy use, Quality of Service (QoS), and resource usage.
- Implementation: Workflows are carried out using a distributed network consisting of IoT devices, cloud servers, and edge computing nodes. Tasks are assigned in real-time according on the devices' present status and the workload.
- Feedback and Optimization: The system continuously adjusts to evolving traffic circumstances and device capabilities. It modifies the schedule of workflow to provide the best performance and efficiency.

The proposed hybrid algorithm combining Deep Q-learning (DQN) with Chronological Geese Migration Optimization (CGMO) holds significant potential for addressing various challenges in real-world industry settings. Let's explore some concrete examples of its applications across different domains:

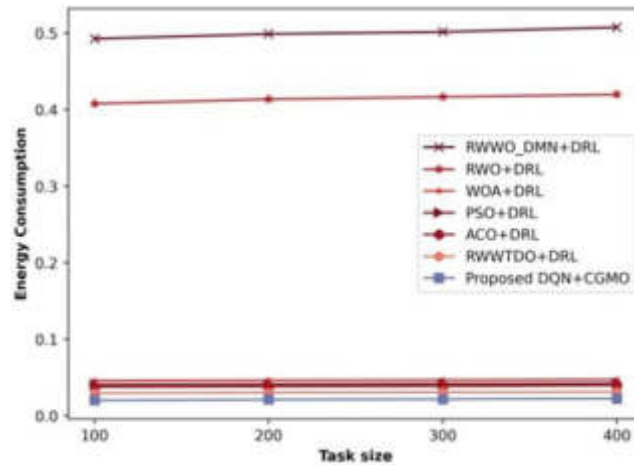


Fig. 6.5: Valuation of DCGM altering task size with PM=20, VM=25 b) Energy Consumption

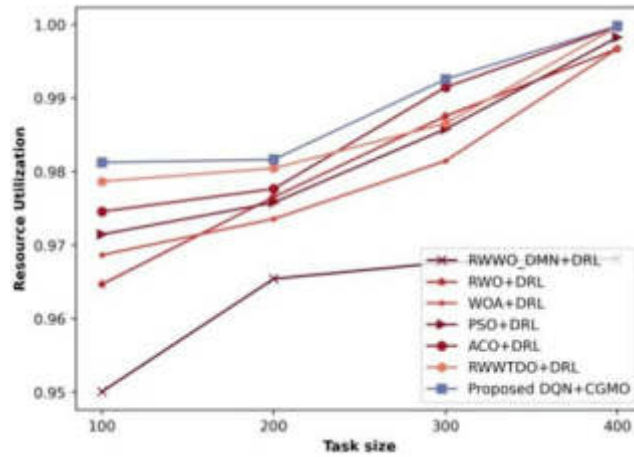


Fig. 6.6: Valuation of DCGM altering task size with PM=20, VM=25 Resource utilization

7.1.1. Cloud Computing and Data Centers.

- **Dynamic Resource Allocation:** In cloud computing environments, where resource demands fluctuate frequently, the hybrid algorithm can dynamically allocate resources based on workload patterns and environmental conditions. For example, in a data center serving multiple clients with varying processing requirements, the algorithm can optimize resource utilization by intelligently scheduling tasks on available servers, ensuring efficient use of computing resources while minimizing energy consumption.
- **Fault Tolerance and Load Balancing:** The algorithm can also enhance fault tolerance and load balancing in distributed systems. By continuously monitoring the health and performance of servers, it can redistribute tasks in response to failures or overloads, ensuring high availability and reliability of services. For instance, in a web hosting environment experiencing sudden spikes in traffic, the algorithm can automatically scale resources and distribute incoming requests across servers to maintain optimal performance.

7.1.2. Manufacturing and Supply Chain Management.

- **Production Scheduling:** In manufacturing facilities, the algorithm can optimize production schedules by

intelligently allocating resources (e.g., machines, workers) to different tasks based on demand forecasts and production constraints. For example, in a manufacturing plant producing automotive parts, the algorithm can dynamically adjust production schedules in response to changes in demand, inventory levels, and machine availability, optimizing throughput and minimizing production costs.

- **Inventory Management:** The algorithm can also improve inventory management in supply chain networks by optimizing order fulfillment and logistics operations. For instance, in a retail distribution center managing inventory across multiple warehouses, the algorithm can optimize the routing and scheduling of delivery trucks to minimize transportation costs and ensure timely delivery of goods to customers.

7.2. Implications and Potential Limitations of Proposed Work.

7.2.1. Implications.

- **Improved Efficiency and Resource Utilization:** The hybrid algorithm has the potential to significantly enhance efficiency and resource utilization in cloud environments by dynamically optimizing task scheduling decisions based on real-time data and environmental conditions. This can lead to cost savings, energy efficiency improvements, and better overall performance.
- **Enhanced Scalability and Flexibility:** By leveraging the capabilities of both DQN and CGMO, the algorithm can adapt to a wide range of workload patterns and system dynamics, making it well-suited for highly scalable and flexible cloud environments. It can handle diverse workloads and dynamically adjust scheduling policies to accommodate changing demands and resource availability.
- **Better Resilience and Fault Tolerance:** The algorithm's ability to dynamically adapt to changing conditions and redistribute tasks in response to failures or overloads can enhance system resilience and fault tolerance. It can help mitigate the impact of hardware failures, network outages, or sudden spikes in demand by quickly reallocating resources and rerouting tasks to unaffected components.

7.2.2. Potential Limitations.

- **Complexity and Computational Overhead:** Implementing and fine-tuning the hybrid algorithm may require significant computational resources and expertise, particularly in training the DQN model and optimizing CGMO parameters. The algorithm's complexity could pose challenges in real-time decision-making and scalability, especially in large-scale cloud environments with high workload variability.
- **Sensitivity to Environmental Factors:** The algorithm's performance may be influenced by the accuracy and reliability of environmental data, such as temperature, humidity, and renewable energy availability. Inaccurate or noisy data could lead to suboptimal scheduling decisions, highlighting the importance of robust data preprocessing and quality assurance mechanisms.

7.2.3. Evolution and Adaptability.

- **Continuous Learning and Optimization:** The hybrid algorithm can evolve over time through continuous learning and optimization, leveraging feedback data to refine its decision-making policies and adapt to evolving workload patterns and system dynamics. By incorporating reinforcement learning techniques, the algorithm can autonomously improve its performance and adaptability over time without manual intervention.
- **Customization for Different Cloud Environments:** The algorithm can be customized and fine-tuned to meet the specific requirements and characteristics of different cloud environments, such as public clouds, private clouds, and hybrid clouds. By adjusting model parameters, optimization objectives, and decision-making criteria, the algorithm can adapt to diverse deployment scenarios and optimize performance in various operational contexts.

8. Conclusion. DCGM is an innovative approach for scheduling workflow in CC that uses multi-objective and DL. This study incorporates two processes multi-objective workflow scheduling using DL, and vice versa. The fitness of workflow scheduling is measured using nine criteria in multi-objective scheduling. CGMO, which is derived from the combination of the GMO algorithm with the chronological idea, is taken into account throughout the workflow scheduling process. At the same time, DQN is fed the incoming task in the DL-based workflow scheduling process. Workflow scheduling also takes into account VM and task settings in real time.

Here, both the task parameters (such as the average computing cost of a task, the EST of a task, the EFT of a job, the duration of a task, and its priority) and the VM characteristics (such as the VM's bandwidth usage, memory utilization, capacity, and central processing unit) are taken into account. At the completion of this operation, the two outputs are combined. Maximum Qos (0.785), minimum energy usage (0.022), and maximum resource utilization (1.000) were all reached using the assessment methods used for DQN+CGMO. The study will be expanded to include additional hybrid networks, and the suggested method will eventually be included with cloud toolkits.

REFERENCES

- [1] MELL, P., AND GRANCE, T., *The NIST definition of cloud computing*, 2011.
- [2] ZHU, Z., ZHANG, G., LI, M. AND LIU, X., *Evolutionary multi-objective workflow scheduling in cloud*, IEEE Transactions on parallel and distributed Systems, vol. 27, no. 5, pp.1344-1357, 2015.
- [3] ABAZARI, F., ANALOUI, M., TAKABI, H. AND FU, S., *MOWS: multi-objective workflow scheduling in cloud computing based on heuristic algorithm*, Simulation Modelling Practice and Theory, vol. 93, pp.119-132, 2019.
- [4] YASSA, S., CHELOUAH, R., KADIMA, H. AND GRANADO, B., *Multi-objective approach for energy-aware workflow scheduling in cloud computing environments*, The Scientific World Journal, 2013.
- [5] NARENDRABABU REDDY, G. AND PHANI KUMAR, S., *Regressive whale optimization for workflow scheduling in cloud computing*, International Journal of Computational Intelligence and Applications, vol. 18, no. 04, pp.1950024, 2019.
- [6] MASDARI, M., VALIKARDAN, S., SHAHI, Z. AND AZAR, S.I., *Towards workflow scheduling in cloud computing: a comprehensive analysis*, Journal of Network and Computer Applications, vol. 66, pp.64-82, 2016.
- [7] CHEN, W.N. AND ZHANG, J., *An ant colony optimization approach to a grid workflow scheduling problem with various QoS requirements*, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 39, no. 1, pp.29-43, 2008.
- [8] ABRISHAMI, S., NAGHIBZADEH, M. AND EPEMA, D.H., *Deadline-constrained workflow scheduling algorithms for infrastructure as a service clouds* Future generation computer systems, vol. 29, no. 1, pp.158-169, 2013.
- [9] JIA, Y.H., CHEN, W.N., YUAN, H., GU, T., ZHANG, H., GAO, Y. AND ZHANG, J., *An intelligent cloud workflow scheduling system with time estimation and adaptive ant colony optimization*, IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 51, no. 1, pp.634-649, 2018.
- [10] ARABNEJAD, H. AND BARBOSA, J.G., *A budget constrained scheduling algorithm for workflow applications*, Journal of grid computing, vol. 12, pp.665-679, 2014.
- [11] LUDÄSCHER, B., ALTINTAS, I., BERKLEY, C., HIGGINS, D., JAEGER, E., JONES, M., LEE, E.A., TAO, J. AND ZHAO, Y., *Scientific workflow management and the Kepler system*, Concurrency and computation: Practice and experience, vol. 18, no. 10, pp.1039- 1065, 2006.
- [12] KAKKOTTAKATH VALAPPIL THEKKEPURYIL, J., SUSEELAN, D.P. AND KEERIKKATIL, P.M., *An effective meta-heuristic based multi-objective hybrid optimization method for workflow scheduling in cloud computing environment*, Cluster Computing, vol. 24, pp.2367-2384, 2021.
- [13] KAUR, P. AND MEHTA, S., *Resource provisioning and work flow scheduling in clouds using augmented Shuffled Frog Leaping Algorithm*, Journal of Parallel and Distributed Computing, vol. 101, pp.41-50, 2017.
- [14] THAIN, D., TANNENBAUM, T. AND LIVNY, M., *Condor and the Grid. Grid computing: Making the global infrastructure a reality*, pp.299-335, 2003.
- [15] BUYYA, R. AND VENUGOPAL, S., *The gridbus toolkit for service oriented grid and utility computing: An overview and status report*, In 1st IEEE International Workshop on Grid Economics and Business Models, GECON, pp. 19-66, IEEE, April 2004.
- [16] MCGOUGH, S., YOUNG, L., AFZAL, A., NEWHOUSE, S. AND DARLINGTON, J., *Workflow enactment in ICENI*, In UK e-Science All Hands Meeting, vol. 9, pp. 894-900, September 2004.
- [17] DEELMAN, E., BLYTHE, J., GIL, Y., KESSELMAN, C., MEHTA, G., PATIL, S., SU, M.H., VAHI, K. AND LIVNY, M., *Pegasus: Mapping scientific workflows onto the grid* In Grid Computing: Second European AcrossGrids Conference, AxGrids 2004, Nicosia, Cyprus, January 28-30, 2004. Revised Papers, pp. 11-20, Springer Berlin Heidelberg, 2004.
- [18] STRUMBERGER, I., BACANIN, N., TUBA, M. AND TUBA, E., *Resource scheduling in cloud computing based on a hybridized whale optimization algorithm*, Applied Sciences, vol. 9, no. 22, pp.4893, 2019.
- [19] STEPHANAKIS, I.M., CHOCHLIOUROS, I.P., CARIDAKIS, G. AND KOLLIAS, S., *A particle swarm optimization (PSO) model for scheduling nonlinear multimedia services in multicommodity fat-tree cloud networks*, In Engineering Applications of Neural Networks: 14th International Conference, EANN 2013, Halkidiki, Greece, September 13-16, 2013 Proceedings, Part II 14, pp. 257-268, Springer Berlin Heidelberg, 2013.
- [20] KAKKOTTAKATH VALAPPIL THEKKEPURYIL, J., SUSEELAN, D.P. AND KEERIKKATIL, P.M., *An effective meta-heuristic based multi-objective hybrid optimization method for workflow scheduling in cloud computing environment*, Cluster Computing, vol. 24, pp.2367-2384, 2021.
- [21] CHOUDHARY, A., GUPTA, I., SINGH, V. AND JANA, P.K., *A GSA based hybrid algorithm for bi-objective workflow scheduling in cloud computing*, Future Generation Computer Systems, vol.83, pp.14-26, 2018.
- [22] WU, H., ZHANG, X., SONG, L., ZHANG, Y., GU, L. AND ZHAO, X., *Wild Geese Migration Optimization Algorithm: A New Meta-Heuristic Algorithm for Solving Inverse Kinematics of Robot*, Computational Intelligence and Neuroscience, 2022.

- [23] KAUR, A., SINGH, P., SINGH BATTH, R. AND PENG LIM, C., *Deep Q learning based heterogeneous earliest finish time scheduling algorithm for scientific workflows in cloud*, Software: Practice and Experience, vol. 52, no. 3, pp.689-709, 2022.
- [24] MAZREKAJ, A., SHEHOLLI, A., MINAROLLI, D. AND FREISLEBEN, B., *The Experiential Heterogeneous Earliest Finish Time Algorithm for Task Scheduling in Clouds*, In CLOSER, pp. 371-379, May 2019.
- [25] GE, J., HE, Q. AND FANG, Y., *Cloud computing task scheduling strategy based on improved differential evolution algorithm*, In AIP Conference Proceedings, vol. 1834, no. 1, pp. 040038, AIP Publishing LLC, April 2017.
- [26] KUANG, L. AND ZHANG, L., *A new task scheduling algorithm based on value and time for cloud platform*, In AIP Conference Proceedings, vol. 1864, no. 1, pp. 020017, AIP Publishing LLC, August 2017.
- [27] SUN, W., SU, F. AND WANG, L., *Improving deep neural networks with multi-layer maxout networks and a novel initialization method*, Neurocomputing, vol.278, pp.34-40, 2018.
- [28] DONG, T., XUE, F., XIAO, C. AND ZHANG, J., *Workflow scheduling based on deep reinforcement learning in the cloud environment*, Journal of Ambient Intelligence and Humanized Computing, pp.1-13, 2021.
- [29] A V L N SUJITH, DR. A RAMA MOHAN REDDY, AND DR. K MADHAVI. , *Evaluating the QoS Cognizance in Composition of Cloud Services: A Systematic Literature Review* presented in International Conference on Advanced Machine Learning and Soft Computing-2018 and published in International Journal of Engineering and Technology, 7 (4.6) (2018) 141-149
- [30] A V L N SUJITH, DR. A RAMA MOHAN REDDY, AND DR. K MADHAVI., *EGCOPRAS: QoS-Aware Hybrid MCDM Model for Cloud Service Selection in Multi-Cloud Environment* Journal of Adv. Research in Dynamical & Control Systems, Vol. 11, 06, 2019
- [31] A V L N SUJITH, DR. A RAMA MOHAN REDDY, AND DR. K MADHAVI. , *QoS-Driven Optimal Multi- Cloud Service Composition Using Discrete and Fuzzy Integrated Cuckoo Search Algorithm* International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249- 8958, Volume-8 Issue- 5, June 2019.
- [32] RAMESH, G., GORANTLA, VENKATA ASHOK K AND GUDE, VENKATARAMAIAH., *A hybrid methodology with learning based approach for protecting systems from DDoS attacks*, Journal of Discrete Mathematical Sciences and Cryptography, 26:5, 13171325.

Edited by: Anil Kumar Budati

Special issue on: Soft Computing and Artificial Intelligence for wire/wireless Human-Machine Interface

Received: Jan 1, 2024

Accepted: Apr 17, 2024



A SECURE DATA STORAGE APPROACH FOR ONLINE EXAMINATION PLATFORM USING CLOUD DBAAS SERVICE

SRINU BANOTHU*, G. JANARDHAN†, G. SIRISHA‡, SRINIVASULU SHEPURI§, MADHAVI KARNAM¶, AND ALLAM BALARAM||

Abstract. For the time being, many government or private organizations for recruitment of staff or educational institutions moving towards online based tests. The online examination system is a software application used for conducting examination using computer systems. It helps to the recruitment agency or any govt. or private organizations for conducting any job recruitment examinations transparently. Due to this system results are processed without delay and efficiently evaluated to assess the candidates abilities. But the biggest challenge for online examination system is data integrity, security and privacy. The current system is resolving the privacy issue by providing authentication credentials such as user name, password to the candidates. So that only authorized users with proper credentials can login to the system and attempt the exam. But the data confidentiality and integrity are biggest challenges for the system. As the data stored in system database is in plain text format, hence it may be modified or misused by the internal staff of the organization. This paper presents the frame work for secure storage and management of candidates data using encryption scheme, distributed databases in cloud database system. The proposed framework enhances the data confidentiality, integrity and avoids any cheating by internal staff or third party institutions. This paper conducts experimental work on proposed framework and analyses the results of the system.

Key words: online examination system, data security, cloud database, encryption, distributed database

1. Introduction. Now a days most of the recruitment agencies such as government or semi government are conducting online examination as part of recruitment process for faster evaluation and recruitment process [1]. Many recruitment agencies are gradually replacing paper examinations by online examination systems due to various information technologies and rapid evolution of network technology. Online examination systems improve the efficiency and quality of the examination and make the examinations not limited to places and regions [2]. Existing online examination system modes consist mainly of Client/Server (C /S) and Browser/Server (B /S) structure [3,4]. For the C /S examination system, examination center is autonomous. During the examination, examination questions and examination information are pushed down to examination center by the administrator. The examinee is able to take the examination at examination center. This type of structure is mostly distributed in local area network. Candidates can only take the exams within the prescribed environment that is limited to some extent in terms of time and space. In addition, this type of system has a low carrying capacity, it is not easy to scale up, it is easy to lose candidates answers in emergency situations and there are some issues like disconnected examination systems and problems with data synchronization which make it difficult for B/S based examination systems to be widely used, which is why a new technology is needed to solve this dilemma [5].

The emergence of cloud computing is the result of the rapid evolution of the next generation Internet technology. It is a new form of neural computing mode [68] that allows computing to be distributed on many distributed computers instead of on local computers or on remote servers. It is a result of the integration of distributed computing and utility computing technologies, virtualization technologies, web services technologies, grid computing technologies, and others. The purpose of CC is to enable users to utilize virtual resource pools

*Dept. of CSE, Vignan Institute of Technology and Science, Deshmukhi(v), Yadadri Bhuvanagiri Dist, Telangana, India, 508284

†Dept. of CSE, Vignan Institute of Technology and Science, Deshmukhi(v), Yadadri Bhuvanagiri Dist, Telangana, India, 508284

‡Dept. of CSE, CVR College of Engineering, Mangalpally(V), Ranga Reddy Dist, Telangana, India, 501510

§Dept. of CSE (AIML), AVN Institute of Engineering and Technology, Hyderabad, Telangana state, India, 501505

¶Dept. of CSE, Gokaraju Rangaraju Institute of Engineering and Technology, Bachupally, Hyderabad, Telangana, India, 500090

||Department of CSE, MLR Institute of Technology, Hyderabad, Telangana - 500043. (bmadhaviranjjan@yahoo.com)

as much as possible at any time, anywhere on the network to solve large scale computing problems. One of the services offered by CC is Software as a Service (SaaS). SaaS is a type of software application mode which provides software services on the Internet and is the latest trend in software technology development [9, 10].

In addition to SaaS, double cache and adapter technologies are used to solve the problem of examinees answer loss and the problem of data synchronization between the examination systems in unexpected circumstances. As a result, CC offers a technical solution to the current design of the online examination system.

Conducting examination through online platform is having many advantages: it avoids the major issues of recruitment process such as paper leakage, recently it became biggest issue in telangana state(India), the internal staff of the Telangana State Public Service Commission soled the Group-I, Assistant Executive Engineering examination papers and earned the lot of black money. With this many peoples life whoever were sincerely prepared for getting job got spoiled. And also completes the recruitment process smoothly without any delays. Hence, many recruitment agencies are moving towards online examination platform. Besides advantages of this system, it also facing some issues related to user authentication and secure storage of candidate marks data. The marks obtained by the candidate plays a vital role in recruitment process, so biggest challenge is security and integrity of the marks data stored in database system. It became so significant to solve security issues; otherwise many qualified candidates lose opportunities. With these things in mind, a frame work for secure storage of marks data in cloud database system is proposed. [11] Cloud computing is a technology that provides data processing and storage services through internet on rental basis. One of cloud computing services is Database as a Service (DBaaS). The cloud Database as a Service (DBaaS) enables users to outsource data in cloud database system and access whenever required through any devices connected to internet. DBaaS provides organizations with unlimited data storage services cost-effectively with higher availability and easy deployment. Now a days most of the organizations or individuals are outsourcing their databases to the cloud environment.

The objective of this research is to develop a secure data storage and management system in cloud for online examination system using distributed databases and data encryption algorithms to meet the challenges facing by recruitment agencies.

2. Literature Survey. Any recruitment organization can plan, administer, and oversee exams in an online setting with the help of an online examination system. It helps the inspector by lessening the workload associated with administering tests, examining answer sheets, and generating results [12]. In light of this, online tests have become increasingly common in recent years. Although many young students disclose personal information on social media, Okada et al. [13] noted that their attitudes change when it comes to e-assessment since they are more worried about data privacy, security, and safety.

Cluskey et al. [14] studied all the feasible approaches to conduct online examinations without supervision. This paper presented the detail discussion about cheating scenarios used by students and also measures to avoid cheating by students. Authors have guided few methods for building an online testing plan and few online examination control techniques such as taking tests at one set time using Respondus Lockdown Browser (RDL), checking student ID, and so on.

Authors [15] proposed an approach to identify students movement in online examination using Convolution Neural Networks (CNN). The problem with this approach is that the position and orientation of students cannot be analyzed and requires much data to preview the data. And this approach is not focused on the security of data stored in database.

Mukta et. al. [16] worked on Adaptive Test Sheet Generation in E-Learning using Fuzzy Logic Approach. Authors guided the employing of an ambiguous technique of assessing students favorite tests in the e-learning.

Jung and Yeom [17] discussed about how to secure an online examination system using cryptography group. However, it needs the use of higher quality webcams and microphones. This is a disadvantage of the system. They have not concentrated about the protection of data from insider attackers.

Paul et al [18]. proposed a system where in the production manager of the questionnaire selects a percentage of complexity for questions that must be met. The program can generate papers in accordance with the format indicated by the administrator and subsequently save it in PDF format, enabling colleges to receive it upon clicking send.

Zhen and Su [19] suggested a methodology for designing a question paper template based on the input

requirements.

The paper [20] proposed a few strategies for the face identification system on this respect. The authors have defined how neural networks (NN), Support Vector Machines (SVM), and Algebraic characterization may be utilized in face reorganization systems. The colleagues and Jain presented the significance of a blockchain technology network in the online examination system. Authors introduced blockchain based online examination system. They employed the Smart contracts, it is one of the better applications and a Ethereum public blockchain platform [21]. They have additionally compared the general effectiveness of the blockchain-based method with the cloud-based scheme.

Lee and et al. developed in [22] a system that classifies student's VFOA information by capturing their head pose estimates and eye movement estimates using advanced technologies artificial intelligence approaches.

The papers [23, 24] proposed the approaches for user authentication and prevention of malpractices by users. Authors also proposed approaches for identification of misconduct of users during examination.

The authors of [24, 25, and 26] proposed the approaches for secure storage of data in cloud environments and performances of various database encryption algorithms for ensuring the confidentiality of data stored in cloud database systems. An unauthorized user cannot access data in an e-learning system. Only students who have been authenticated and granted permission can view exam data uploaded by teachers. One common technique for access control is encryption.

A session key establishment protocol was presented by Kausar et al. [27] for a predetermined duration, such as a class, seminar, or exam. The session key, which encrypts messages using symmetric cryptography, is distributed using a public key infrastructure, and message integrity is ensured by a hash-based message authentication code.

To properly finish the login procedure in Al-Hawari and Alshawabkeh's study [28], students must enter the exam instance session password correctly; the exam instance session password was produced automatically by the examination management system, and it wasn't made public until the instructor revealed it at the start of the relevant class. Few prior investigations [27, 28] have reported the possibility of a single point failure, the inability to prevent communication parties' repudiation actions, and the inability to resolve internal conflicts. In order to efficiently upload and download data in cloud-based systems, Sahaya et al. [29] presented a strategy that first encrypts the message using DES and then encodes it using Reed Solomon code in the data centers. However, it doesn't have precise access control.

The common problem identified from the above literature is that there are no approaches focused on security to the marks data stored database system. The marks obtained by the candidate may be modified by the internal staff of the organization by miss using authentication credentials. Instead our proposed model focuses on security and integrity of the data stored in database systems

3. Proposed System. To address security issues of marks data in online examinations systems and for better availability and scalability of the system. This work proposes a new framework using cloud services, cryptographic algorithms and data distribution. It is hopeful that proposed work overcomes the issues of data confidentiality, integrity and availability. These are three key measures of data security. Here a brief overview of proposed approach is presented with figure 1 for secure data storage and management of online examination systems using one of cloud service such as Cloud Database as a Service, data distribution and cryptographic algorithms. The objectives of proposed system are also discussed.

3.1. System Model. This system is having three major modules such as 1) users 2) administrator and 2) Cloud databases.

Users: The users are candidates who write the examination, every user is having authentication credentials such as username and password for login into the system. Once users are signed in they will write and submit examination. After submission users marks data and personal information will be stored in local database servers of the system.

Administrator: An administrator is a person who is responsible for whole data management such as data encryption, data outsourcing to cloud database servers, key generation, ensuring data privacy and security. An administrator plays a vital role and should be a main responsible person for ensuring privacy and security to online examination data.

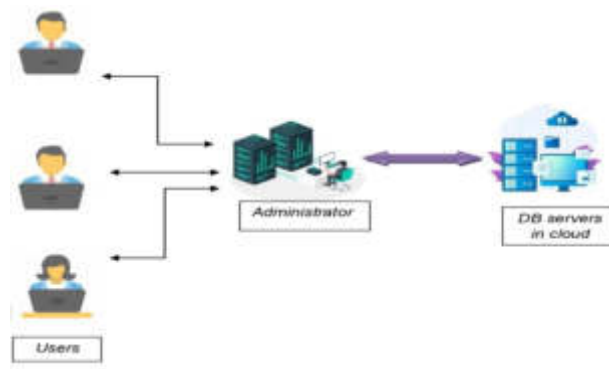


Fig. 3.1: Cloud based secure data storage system for storing online examinations marks data.

Cloud Database Servers: The proposed system uses the cloud DBaaS service for secure storage of candidate's data. Data storage in cloud provides many benefits such higher data availability, security and scalability etc.

3.2. Design Objectives. *Access Control:* The question paper submitted by any recruitment agency can be accessed by authenticated users only during given time period. Answer sheet and score obtained by candidates are accessible to administrator only. The administrator can provide access policies when cipher text of data is generated and stored in distributed data storage servers. Only the users who satisfy the access policy can view the data.

- 1) *Data Confidentiality:* System keeps data highly confidential from unauthorized people access.
- 2) *Data Tampering resistance or Ensuring Data Integrity:* Data integrity is a key issue for online examination system. Any unauthorized entity in the cloud environment or local servers should not modify the data. Otherwise, it violates the transparency of the online examination system.
- 3) *Collision Resistance:* when one key is not enough for data decryption, 2 or more unauthorized entities may try to combine their keys for decryption of data. System ensures that combine keys cannot decrypt correct plain text.
- 4) *Data availability:* As data is outsourced to cloud environment, cloud service providers always keeps data available to be accessed by authorized persons whenever needed. It provides higher scalability.

4. Methodology. This section presents the methodology of proposed system. Proposed approach achieves key elements of data security such as data confidentiality, data integrity and data availability. The aim of this approach is to securely outsource and manage the data of online examination using cryptography algorithms, data distribution and cloud services. The proposed approach uses cryptographic algorithms and vertical fragmentation feature of distributed database to achieve data confidentiality and data integrity. In this approach vertical fragmentation plays a vital role for achieving data confidentiality. Vertical fragmentation is a technique to split the database table vertically into two or more sub table fragments with chosen columns. As database tables are partitioned vertically with selected columns into two or more sub tables and distributed data into multiple database servers in cloud, the internal staff of the cloud service provider cannot get complete information about a record. So it keeps the data secure from insider attack in cloud environment. The cloud services are used for proving better data availability and scalability. Proposed system uses the Cloud Database as a Service (DBaaS) for storing data in cloud environment, Advanced Encryption Standards (AES) algorithm for cryptographic operations (i.e encryption and decryption of data) on data. As AES is more secure and robust algorithm, it is chosen for cryptography operations. AES uses keys of variable length such as 128,192 and 256 bits length keys, for 128 bit key, about 2128 attempts are required to break the cipher. This is very difficult task for the hacker to hack the data. The framework consists of two major modules: users and Administrators.

Users: users login to the online examination system using authorization credentials and attempt the test, after test is over submit the test. This test data will be stored in systems local database servers.

Administrator: Administrator is an owner of the data stored in systems local database servers, performs tasks such as data encryption and outsourcing to cloud environment for ensuring data privacy and security. For achieving data security issues, the proposed framework consists of two phases

Setup Phase. In this phase, administrator does the data pre-processing, outsourcing and user authorization

1. Administrator encrypts the database table attribute values using advanced encryption standard algorithm (AES) with a secret key (this key only knows to the administrator). Administrator uses SHA 256 authentication algorithm for generating 256 bit secret key.
2. Splits the table vertically into 2 or more sub tables with selected columns (i.e. vertical fragmentation). While splitting tables, in each sub table columns are included using a factor called key and data sensitivity. The maximum number of possible sub tables depends on number of columns of a table. Each sub table must includes at least two columns
3. Add row index column in each sub table and insert row index value in index column, it should be the same value for a row in every sub table. This helps to identify each sub records of a row of actual table (i.e. before splitting of table) and merge the records of fragmented tables into a record of original table (i.e. before partition). Row index column values are unique and not in encrypted form.
4. Finally, upload the vertically partitioned table data into multiple database server platforms of the same cloud environment. Choose different locations of data centres to store the data of partitioned table fragments. This makes very difficult to an attacker or internal staff to get complete record information of a table as records are partitioned and stored in different data centres.

Retrieval Phase. In this phase, the administrator sends the data retrieve request to all database servers where fragmented data is stored in cloud. The key attribute to select the records from fragmented tables is row index value. The cloud database server returns requested records data in encrypted format from multiple sub table fragments. An application merges the records data into a record of original table (before fragmentation). Then administrator gives input as secret key and then decrypted results are shown to the user. Only one key is used for data decryption The algorithms for secure data uploading to the cloud environment and retrieval are shown in Algorithm 1 and 2.

5. Implementations and Results.

Experimental Setup. For experimental results we have used the cloud service from cloudcluster.in and software technology PHP (Hypertext Preprocessor) for application development and MySQL database server for backend data store. For testing the results of our proposed model, initially we have created a account in cloudcluster.io, cloud cluster provides a complete managed open source application cloud service on kubernetes cloud for cloud DBaaS service. In which we have deployed two MySQL database servers with configurations of servers on cloud platform are:3(core) Processors, 4GB RAM, 100GB SSD and chosen data centers at two different locations for storing the fragmented table data. Then developed a small application in PHP and installed MySQL database server on local system for storing dataset. The data set is created with random values of marks in the range from 0 to 50. Data set includes fields such as candidate id, test id, marks scored. XAMPP server is installed on our local system to run an application. Our system is configured with Intel core i5 processor, 10GB RAM and 360GB hard disk space. System is connected to the internet of 150mbps Network speed. The results are shown in below figures.

For experimental purpose, we have created a database table named as testscore with fields such as id, candidateid, testid and testscore, in our local machine.

Figure 5.1 shows the records inserted into the testscore table in local system. The records in the database table are stored in plain text format or readable format. Then created two table fragments, one with fields id, candidateid, testid and other with id, testscore in two different cloud database servers. Here attribute id is a row index added in both table fragments. Later, run our application on our machine to encrypt, split and insert encrypted records into cloud database servers as shown in figure 5.2, in which records with selected columns are stored in encrypted format. Finally, figure 5.3 shows the encrypted data stored in fragmented table. Both table fragments are having same values for every row for row index attribute id.

5.1. Performance Analysis.

Performance Evaluation. The performance of proposed scheme is evaluated by considering the parameters such as time taken to encrypt, split and upload data into cloud environment and time taken to retrieve the

Algorithm 1 Database upload to cloud

```

Procedure databaseupload(dbname, tablename, secretkey )
{

Inputs: local database name, table name, secret key as inputs
Output: encrypted and vertically partitioned table
segments uploaded to database servers in cloud      ▷ generate the 256 bit hash code of secret key using SHA256
algorithm and save key in local database system

Hashed_key=sha256(secret_key);
                                                    ▷ Select the table data to be outsourced from local database server

if(num_of_rows ≥ 0)
{

while (all rows are processed)
{
    ▷ Select one row and encrypt all column values of selected row using AES256 algorithm with hashed secret key.

C1=encrypt (column-1)
C2=encrypt (column-2)
C3=encrypt (column-2)
C4=encrypt (column-2)
.
.
Cn= encrypt (column-n)
Insert (row-index, C1) into tablefragment1 in cloud
Insert (row-index, C2) into table fragment 2 in cloud
Insert (row-index, C3) into table fragment 3 in cloud
Insert (row-index, C4) into table fragment 4 in cloud
Insert (row-index, C5) into table fragment 5 in cloud
.
.
Insert (row-index, Cn) into table fragment
N in cloud
} } }

```

data from cloud database and view results to the users. The time to upload data (UT) considered the time of AES algorithm for data encryption (ET), communication cost (CC) and query execution time (QET).

$$UT = ET + CC + QET \quad (5.1)$$

The time to retrieve the data (RT) from cloud database servers, considered the time of AES algorithm for data decryption (DT), communication cost (CC) and SELECT query execution time (SQET). Here SELECT query without predicate is used for retrieving all the records from database.

$$RT = DT + CC + SQET \quad (5.2)$$

The performance of data upload time and retrieval time are tested by considering the data sets with variable number of records. The performance of the system in terms of time in seconds to upload data and retrieval data are shown in figure 5.4.

Security Alalysis. The security of our proposed approach is analyzed against to insider attacks and outsider attacks. The data is strongly protected against the insider and outsider attackers in such a way that as data is encrypted and distributed into multiple database systems if attacker compromises the data in one fragment,

Algorithm 2 For Secure retrieving of data from the cloud DB

```

Procedure database__retrieve (dbname, tablename, secretkey )
{
    ▷ Cipher text can be from your MySQL data or from a user input via a web form.
    ▷ This example will use user input cipher text to decrypt.
    ▷ generate the 256 bit hash code of secret key using SHA256 algorithm and compare with saved key in database

    Hashed_key=sha256(secret_key);
    ▷ compares decryption key

    if (decryption key matches with encryption key)
    {
        ▷ retrieve the records from table fragment1 in cipher text format

        Sql1=SELECT * from fagment1;

        if (num_rows_of_fragment1 ≥ 0)
        {
            ▷ select one record from fragment1

            While (all rows of fragment1 are processed)
            {
                ▷ select records from table fragment2 matching id with id of record selected from table fragment 1
                ▷ merge records retrieved from fragment 1 and 2

                $ row= $ row1+$ row2;
                ▷ decrypt all column values using AES decryption algorithm and secret key

                P1= decrypt (column-1)
                P2= decrypt (column-2)
                P3= decrypt (column-2)
                P4= decrypt (column-2)
                .
                .
                .
                Pn= decrypt (column-n)
                } } } }

```

cannot get the complete information and inside attackers also unaware about data format stored in database. The proposed scheme achieves data integrity such a way that as data is encrypted using symmetric encryption scheme using a secret key if any modification is done on cipher text, it cannot be decrypted properly.

As data is stored in one cloud vendor, the approach requires less cost and less communication delay for data upload and retrieval operations. So our model is more efficient and secure with less cost than available state of art models in literature. From literature, authors have focused on user authentication of online examination system to prevent from masquerade attacks, in this study we focused on confidentiality and integrity of marks data of the users after storing in database servers. This study ensure to protect data from internal staff of the organization, because internal staff may misuse their credentials and do update the marks. This is huge loss for the candidates those who prepare sincerely and write the examination.

Our proposed scheme, in addition to offering a higher security level, our scheme preserves better communication and computation performance while uploading data and retrieval phase. This is particularly important when it comes to privacy-preserving and avoiding single-point failure.

6. Conclusion and Future scope. In this paper, we propose a scheme for secure data storage and management of online examination systems using symmetric key cryptography algorithms, data distribution and cloud database as a service (DBaaS). In proposed scheme for preserving data confidentiality and integrity, database table records are encrypted using cryptographic AES-256 algorithm with 256 bit secret key generated

SELECT query execution time on UCA: 311.95521354675 Milliseconds

id	Candidate_id	Test_id	Score
48193	cand_00001	Test_01	34
48194	cand_00002	Test_01	28
48195	cand_00003	Test_01	38
48196	cand_00004	Test_01	27
48197	cand_00005	Test_01	24
48198	cand_00006	Test_01	22
48199	cand_00007	Test_01	38
48200	cand_00008	Test_01	34
48201	cand_00009	Test_01	23
48202	cand_00010	Test_01	37
48203	cand_00011	Test_01	27
48204	cand_00012	Test_01	34
48205	cand_00013	Test_01	32
48206	cand_00014	Test_01	39
48207	cand_00015	Test_01	37
48208	cand_00016	Test_01	29
48209	cand_00017	Test_01	34
48210	cand_00018	Test_01	34
48211	cand_00019	Test_01	30
48212	cand_00020	Test_01	22

Fig. 5.1: Data stored in local database system before encryption

SELECT * FROM "Test_Score"

id	Candidate_id	Test_id
48193	L3JP2GFh8FYTWUjZwD3H0pPMv06QTO90jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48194	NDhc0RIT0EjQ2owH02BEY1andkZz090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48195	NDMvUXZyY9IGT09UbuNwXZmfLUEQT090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48196	Vmh5a0hzSTdEd20MfdR0hdne91qz2090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48197	NuH0VDRQzVRRGJd1FNUDY2UEJwZz090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48198	Tmp0WWhIb0ndWKNWk4ZndbHkuU7090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48199	atH6S25xRABmp0R3c3NuQhZEs0dx090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48200	Rw6zyUp0ccGgUjXRxkU09kInYmRtUT090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48201	R2JL5tE1ZDFvVNFVNLDRSsmhZGT090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48202	Z1VWZmZW2hpKdFpaTXRnmE09W1TGT090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48203	CHM5JmQVONFE1Z2pWVlak1uZ2090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48204	ME1TY5Wf0Z0YUz0RTOJCeM0Zz090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48205	dRUR09T3lyU1FmWf958mdU228ZUT090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48206	b2F2bnY0NjF2cVZWzzyYbmc1L1R1td090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48207	NGSPM0xH1BmnpTvdPRFVMEHxZ090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48208	OUHJa1FjdW3bljWEIBRXZueXq0T090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48209	Uz06FzYyTn1c2hQUkSt05MvgZz090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48210	d05Wm9aUkQW4bFEZY2Z0gF0dc090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48211	YX14c2dZG4NtTevYvCRS9razAxQ1090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48212	MTJ6OUe0Y2VwS455VhVYUzka2g0U7090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48213	emh0FRQ1H4RUw5L2RBTmZJSXp5GT090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48214	hHU2VU1GW05abXE56ZyBtHRWkGT090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48215	UC9adm45UHW00WVhR203WmFNz090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48216	cm9kZU1STZTU1B3v0hWfZb5vYU090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD
48217	V0wOUZDUURybdMxvWnH3B609Q7090jowZDA4MDY3YTM4MD	OvisMFFlev13L1beV.J3dFyM0Ez2090jowZDA4MDY3YTM4MD

Fig. 5.2: Data stored in database after encryption and fragmentation (Table fragment1)

using SHA256 algorithm, and then table is vertically partitioned with selected columns into multiple small table fragments. These small table fragments are inserted and stored in different database servers in cloud environment. As data is stored in cipher text format and only part of the entire record is stored in each database servers, the internal staff of the organization also cannot know what data is. So the proposed scheme

```
SELECT * FROM 'Test_Score2'
```

id	Score
48193	cFISmM0Z2piakNUdV.JjRH3T1lzZz09OjowZDA4MDY3YTM4MD...
48194	am1MTG1uZDgwOT.Jvd1p4RVdyNn2Zz09OjowZDA4MDY3YTM4MD...
48195	N3pl.eHRJSFgvM05oM2g2NIVFWm1qZz09OjowZDA4MDY3YTM4MD...
48196	K2kydmpH TWZZU1hVfHJa3lybIRyZz09OjowZDA4MDY3YTM4MD...
48197	b3zT3NDNlreVFCWF.dHWF.ZmMHY4Zz09OjowZDA4MDY3YTM4MD...
48198	Y2pXVmiHOFNZZ2NpcE5iNVhTTFZQUt09OjowZDA4MDY3YTM4MD...
48199	N3pl.eHRJSFgvM05oM. No row selected.
48200	cFISmM0Z2piakNUdV.JjRH3T1lzZz09OjowZDA4MDY3YTM4MD...
48201	MW5WK2d5K3VvVHdyYk.J5eXhnNktYzZz09OjowZDA4MDY3YTM4MD...
48202	QnhZYzNVK3dYY1AwMFkrNEhMXprZz09OjowZDA4MDY3YTM4MD...
48203	K2kydmpH TWZZU1hVfHJa3lybIRyZz09OjowZDA4MDY3YTM4MD...
48204	cFISmM0Z2piakNUdV.JjRH3T1lzZz09OjowZDA4MDY3YTM4MD...
48205	Z3YwN0MzOHRkcvRMTFIMMEVUImJwQT09OjowZDA4MDY3YTM4MD...
48206	LDF5Zk01bWEyOC85THJ5WWRXdlQT09OjowZDA4MDY3YTM4MD...
48207	QnhZYzNVK3dYY1AwMFkrNEhMXprZz09OjowZDA4MDY3YTM4MD...
48208	aW0xYTBehkvUjLWfVrUnk4Vjh4QT09OjowZDA4MDY3YTM4MD...
48209	cFISmM0Z2piakNUdV.JjRH3T1lzZz09OjowZDA4MDY3YTM4MD...
48210	cFISmM0Z2piakNUdV.JjRH3T1lzZz09OjowZDA4MDY3YTM4MD...
48211	eI3R21XZV4NVN1Z2uz5VJFT1BmZz09OjowZDA4MDY3YTM4MD...
48212	Y2pXVmiHOFNZZ2NpcE5iNVhTTFZQUt09OjowZDA4MDY3YTM4MD...
48213	QnhZYzNVK3dYY1AwMFkrNEhMXprZz09OjowZDA4MDY3YTM4MD...
48214	am1MTG1uZDgwOT.Jvd1p4RVdyNn2Zz09OjowZDA4MDY3YTM4MD...
48215	dDdIRdk58m5DeWp2NKNWOSfPbE12UT09OjowZDA4MDY3YTM4MD...
48216	K2kydmpH TWZZU1hVfHJa3lybIRyZz09OjowZDA4MDY3YTM4MD...
48217	aW0xYTBehkvUjLWfVrUnk4Vjh4QT09OjowZDA4MDY3YTM4MD...

Fig. 5.3: Data stored in database after encryption and fragmentation (Table fragment2)

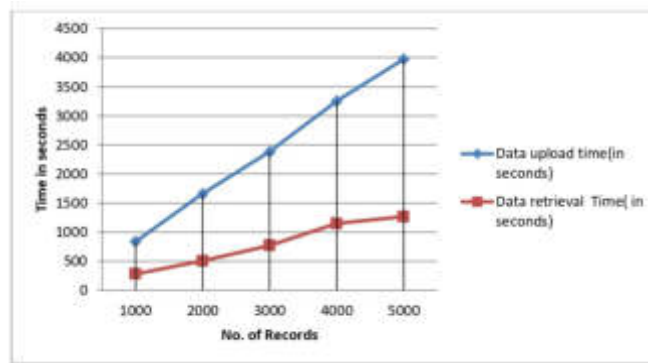


Fig. 5.4: Data upload and retrieval time of proposed scheme

preserves the data confidentially and integrity. As cloud is rental basis service, the cloud services providers always takes care about the security of physical infrastructure and keeps it always available. As cloud vendors maintain the infrastructure suitable for variable sized data, good scalability can be achieved with this approach. In our future research, scheme will incorporate other encryption algorithms and evaluate the performance.

REFERENCES

- [1] SATTAR, M. R. I., EFTY, M. T. B. H., RAFA, T. S., DAS, T., SAMAD, M. S., PATHAK, A., AND ULLAH, M. H. *An advanced and secure framework for conducting online examination using blockchain method*, Cyber Security and Applications, 2023.
- [2] AL-AQBI, A. T. Q., AL-TAIE, R. R. K., & IBRAHIM, S. K. *Design and Implementation of Online Examination System based on MSVS and SQL for University Students in Iraq*, Webology, 18(1). 2021.
- [3] MISTRY, B., PAREKH, H., DESAI, K., & SHAH, N. *Online Examination System with Measures for Prevention of Cheating along with Rapid Assessment and Automatic Grading*. In 2022 5th International Conference on Advances in Science and Technology (ICAST) (pp. 28-34). IEEE. 2022.
- [4] SEMLAMBO, A., ALMASI, K., & LIECHUKA, Y. *PERCEIVED Usefulness and ease of use of online examination system: A case of Institute of Accountancy Arusha*. International Journal of Scientific Research and Management (IJSRM), 10(04),

- 851-861. 2022.
- [5] QIANHUAZHU, *Design and testing of online examination system based on MyEclipse*, Software Engineering and Applications, vol. 08, no. 3, pp. 99103, 2019.
 - [6] GARIMA VERMA *Blockchain-based privacy preservation framework for healthcare data in cloud environment*, Journal of Experimental & Theoretical Artificial Intelligence, 36:1, 147-160, DOI: 10.1080/0952813X.2022.2135611, 2024.
 - [7] Z. HUANG, Y. ZHANG, Q. LI ET AL., *Unidirectional variation and deep CNN denoiser priors for simultaneously destriping and denoising optical remote sensing images*, International Journal of Remote Sensing, vol. 40, no. 15, pp. 57375748, 2019.
 - [8] X.-B. JIN, W.-Z. ZHENG, J.-L. KONG ET AL., *Deep-learning temporal predictor via bidirectional self-attentive encoderdecoder framework for IOT-based environmental sensing in intelligent greenhouse*, Agriculture, vol. 11, no. 8, p. 802, 2021.
 - [9] H. SHI, H. ZHANG, J. HUANG, AND Z. XU, *Design of examination system based on LabVIEW for pesticide detection staff*, Modern electronic technology, vol. 042, no. 2, pp. 4953, 2019.
 - [10] H. RU ZHANG *Application of cloud computing technology in the universitys information construction and development*, Software Engineering and Applications, vol. 8, no. 2, pp. 3237, 2019.
 - [11] BANOTHU, S., GOVARDHAN, A., & MADHAVI, K. *A Fully Distributed Secure Approach for Database Security in Cloud Computing*, In Computational Intelligence and Data Analytics: Proceedings of ICCIDA 2022 (pp. 523-531). Singapore: Springer Nature Singapore, 2022.
 - [12] D. V. KOTWAL, S. R. BHADKE, A. S. GUNJAL ET AL., *Online examination system*, International Research Journal of Engineering and Technology (IRJET), vol. 3, no. 1, pp. 115117, 2016.
 - [13] A. OKADA, D. WHITELOCK, W. HOLMES ET AL., *E-authentication for online assessment: a mixed-method study*, British Journal of Educational Technology, vol. 50, no. 2, pp. 861875, 2019.
 - [14] CLUSKEY JR, G. R., EHLEN, C. R., & RAIBORN, M. H. *Thwarting online exam cheating without proctor supervision.*, Journal of Academic and Business Ethics, 4, 1.,2011.
 - [15] KAVISH GARG, ET AL., *Convolutional neural network-based virtual exam controller*, 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), IEEE, 2020.
 - [16] MUKTA GOYAL, DIVAKAR YADAV, ALKA CHOUBEY, *Fuzzy logic approach for adaptive test sheet generation in e-learning*, 2012 IEEE International Conference on Technology Enhanced Education (ICTEE), IEEE, 2012.
 - [17] IM Y. JUNG, HEON Y. YEOM, *Enhanced security for online exams using group cryptography*, IEEE Trans. Educ. 52 (3) 340349.,2009.
 - [18] MOJITHA MOHANDAS, ET AL., *Automated question paper generator system*, Int. J. Adv. Res. Comput. Commun. Eng. 4 (12) 676678.,2015.
 - [19] DIMPLE V. PAUL, ET AL., *Use of an evolutionary approach for question paper template generation*, 2012 IEEE Fourth International Conference on Technology for Education, IEEE, 2012.
 - [20] CHENGGANG ZHEN, YINGMEI SU, *Research about human face recognition technology*, 2009 International Conference on Test and Measurement, IEEE, Vol. 1,2009.
 - [21] APOORV JAIN, ET AL., *Smart contract enabled online examination system based in blockchain network*, 2021 International Conference on Computer Communication and Informatics (ICCCI), IEEE, 2021.
 - [22] KYUNG MEE LEE, MIK FANGUY, *Online exam proctoring technologies: educational innovation or deterioration?*, Br. J. Educ. Technol.,2022.
 - [23] QUROTUL AINI, ET AL., *Digitalization online exam cards in the era of disruption 5.0 using the DevOps Method*, J. Educ. Sci. Technol. (EST) 7 (1) 6775.,2021.
 - [24] BANOTHU, SRINU, A. GOVARDHAN, AND KARNAM MADHAVI. *A Fully Distributed Secure Approach Using Nondeterministic Encryption For Database Security in Cloud.*, Journal of Theoretical and Applied Information Technology 100.7, 2022.
 - [25] KARANAM, MADHAVI, ET AL. *Performance Evaluation of Cryptographic Security Algorithms on Cloud.*, E3S Web of Conferences. Vol. 391. EDP Sciences, 2023.
 - [26] BANOTHU, SRINU, A. GOVARDHAN, AND KARNAM MADHAVI. *Performance evaluation of cloud database security algorithms.*, E3S Web of Conferences. Vol. 309. EDP Sciences, 2021
 - [27] S. KAUSAR, X. HUAHU, A. ULLAH ET AL., *Fog-assisted secure data exchange for examination and testing in E-learning System*, Mobile Networks and Applications, pp. 117, 2020.
 - [28] F. AL-HAWARI M. ALSHAWABKEH ET AL., *Integrated and secure web-based examination management system*, Computer Applications in Engineering Education, pp. 9941014, 2019.
 - [29] G. SAHAYA STALIN JOSE AND C. SELDEV CHRISTOPHE, *Secure cloud data storage approach in e-learning systems*, Cluster Computing, vol. 22, pp. S12857S12862, 2019.

Edited by: Anil Kumar Budati

Special issue on: Soft Computing and Artificial Intelligence for wire/wireless Human-Machine Interface

Received: Jan 1, 2024

Accepted: Mar 29, 2024



MACHINE LEARNING BASED TOOL WEAR PREDICTION FROM VARIABILITY OF ACOUSTIC SOUND EMISSION SIGNALS

N V. KRISHNAMOORTHY *AND JOSEPH VIJAY †

Abstract. A novel machine learning-based model is introduced in this research paper to forecast tool wear using acoustic emission (AE) signals. Adaptive boosting (AdaBoost) and a sophisticated feature engineering strategy are employed by the model to enhance the precision of its predictions. The proposed model, Machine Learning Tool Wear Prediction (MLTWP), analyzes AE signals generated during machining operations to distinguish between healthy and worn-out tool conditions with remarkable accuracy. The crux of our approach consists of meticulously eliminating and enhancing the temporal and spectral characteristics of the AE signals. We employ the Kolmogorov-Smirnov test to identify the most valuable classification features. We implemented AdaBoost with the objective of progressively enhancing a set of weak classifiers' ability to identify instances that were incorrectly classified in previous iterations. Utilizing this method increases the model's sensitivity to minute variations in tool wear conditions and its overall classification precision. The MLTWP model underwent extensive testing on a benchmark data set comprising 25,304 AE signal records from cutting mill tools, using a training tool split of 9,989 worn (positive) and 8,990 benign (negative) instances. The results of our experiments, validated through four-fold cross-validation, indicate that the MLTWP model exhibits superior performance compared to the existing Tool Wear Prediction using Acoustic Emission Signals (TWPAE) model. To provide greater specificity, the MLTWP exhibited the following metrics on average: precision (92.2%), specificity (91.38%), sensitivity (90.42%), accuracy (90.9%), and MCC (81.72%). The fact that these metrics exhibit significant improvement over those of the TWPAE model demonstrates that our method of feature engineering and adaptive boosting is effective at precisely predicting tool wear. This research not only advances the existing understanding of tool wear prediction but also establishes a robust framework for the implementation of machine learning in manufacturing predictive maintenance.

Key words: Machine Learning; Tool wear Prediction; Acoustic Emission Signal; and Artificial Neural Network; Kolmogorov-Smirnov (KS) Test

1. Introduction. Tool wear has a direct impact on the energy required to remove metal that in turn has an impact on dimensional accuracy, surface roughness, as well as the cutting operations [1]. Tool wear results from chemical, thermal, and mechanical interactions between the materials of the tool and the workpiece. The two primary forms of tool wear, which might signal the end of a tool's life are crater wear as well as flank wear (Vb), which are both caused by these interactions [2]. The efficiency of the procedure relies on flank wear. This variable shows wear from the industry. It has been extensively researched that flank wear raises cutting forces [3], in turn which raises the power tool of machine tools. Forces, power use, and the dynamics of tool-material interaction are all impacted by flank wear. Acoustic emissions (AE) are used to measure flank wear during turning movements, as discussed in [4].

Because of the tool's variable working environment, complex operating conditions, and various cutting settings, tool wear prediction is challenging. Tool wear but also noise signals are only two examples of the information-rich sensor signals with in mechanical tool process. To determine the level of tool wear, signals should be normalised and features extracted.

Noise Monitoring the wear of a tool requires signal processing. The processing of data via the time-frequency analysis is common [5]. The preprocessing of the signal and manual extraction of tool wear features from the signal are common signals for traditional wear detection systems. Data on several cutting parameters under various operating situations must be obtained through numerous tests and manual analysis. These techniques won't work in all situations and have limitations in actual application.

Plastic deformation, corrosion, fracture propagation, impact, erosion, or leakage are the main causes of AE

* Department of Mechanical Engineering, Sri Krishna College of Engineering and Technology, Coimbatore (Corresponding author, nvkrishnamoorthy1968@gmail.com)

†Karunya Institute of Technology and Sciences Coimbatore (vijayjoseph@karunya.edu).

[6]. For non-destructive testing, AE is used in refineries, pipelines, nuclear and conventional power production, aero planes, offshore oil platforms, and paper mills, including structures such as bridges and cranes [7]. For the last 30 years, tribologists, who study the interaction between moving surfaces (such as during machining operations), have concentrated on AE due to the precision with which it can detect surface wear.

AE signals are produced by friction processes range from 50 kHz to 1 MHz [8]. The AE sensor, to avoid signal attenuation, must be attached to the cutting tool or the workpiece. Since both the cutting tool as well as the workpiece are transient, the sensor in practical applications must be positioned in the tool holder.

A simple machine learning model is provided to forecast tool wear. It was inspired by recent promising work in tool wear monitoring using machine learning [9], [10].

2. Related Research. An electronic-mechanical system was created by Weller et al., [11] that uses sonic signals to assess the cutting edge wear of turning tools. In tests, the system is able to identify used cemented carbide tools that can cut AISI 1045 steel. A cutting tool quality check based on sound was developed by Mannan et al., [12]. In order to identify tool wear related sound patterns, the suggested approach enables process of machining sound signals. This technique has been proven to locate cutting instruments that are sharp, semi-sharp, and dull. Chatter was discovered by Delio et al., [13] utilising sound signals. Experiments demonstrate that the chatter detection technique based on audio signals performs as well to dynamometers, accelerometers, as well as displacement probes. A method for legitimate tool status monitoring specifically for turning was created by Salgado and Alonso [14]. One-dimensional spectral analysis of feeding motor current but also audio signals showed direction-changing data. By analysing retrieved characteristics and applying a SVM (support vector machine) technique, tool wear was calculated. To measure tool wear, Aliustaoglu et al., [15] employed microphone-captured audio signals and two-stage fuzzy logic. Testing for drilling was done on a four-axis Cnc turning centre. Sound signals were recorded with the microphone. In studies, two-stage fuzzy logic may identify tool wear. The audio signal processing was employed by Ubhayaratne et al., [16] to track the tool wear in sheet metal stamping. Stamping audio signals were cleaned and preprocessed using semi-blind signal extraction. A technique for determining the degree of tool wear based on vibrations from the machine's spindle was developed by Seemuang et al., [17]. The drilling sound signals were recorded using a low-cost microphone. Support vector machines were utilised by Kothuru et al., [18] to categorise tool wear problems (SVM). Features in the frequency domain were sanded out of audio signals. Over 90% prediction accuracy.

ANN (Artificial Neural Network) should be used to assess feeder machine current, cutting, as well as source of feed, according to Alonso et al (ANN). Maximum vibration length is inversely related to unidirectional tool wear, as shown by Sadettin et al., [19]. Ming-Chyuan et al., [20] as well as Alonso F.J. et al., [21] discovered that the sole audible sign of tool flank wear is when a machine produces a noise when cutting. In order to determine tool flank wear rapidly, According to Peng et al., [22], if the signal is treated linearly and travels, obtaining the Fourier spectrum might produce a considerably lower physical sensation. According to Huang et al., [23], Fourier processing produces global signal qualities as opposed to local ones. Another technique for estimating outcomes is the Hilbert-Huang transformation [24]. Prakash et al., [25], [26] has shown how streamlining a procedure produces good outcomes and a simpler way.

A contemporary model that relates to the objective of this article is projected by Ferrando Chacon et al., [27], which is referred as Tool Wear Prediction using Acoustic Emission Signals (TWP AE) in further discussions. The contemporary model TWP AE has used acoustic emission signal frequencies as features and focused on the process of eliminating redundant features as well as predicting worn state of the tools using machine learning. However, the feature engineering of this model is suboptimal as it is limited to eliminating redundant features and restricted only to the signal frequencies.

Yang, Cheng et al. [28] proposed a novel approach to understanding how milling tools wear down over time. This method combines an analysis of wear-related factors with a wear-prediction model, with a focus on how tool wear works and changes over time. The method improves the accuracy of tool wear predictions by combining experimental data and modeling. The goal is to improve the efficiency of machining processes and reduce downtime.

Perumal et al. [29] presented a simple machine learning method for monitoring tool wear in real time using sound signals. Their findings revealed that basic acoustic emission (AE) signals could be used to predict tool wear, making this a simple and effective method for tool condition monitoring (TCM) in manufacturing settings.

The study also demonstrated that linear regression algorithms can be used to calculate tool wear based on AE data. This could be a low-cost way to increase machining efficiency and tool life.

Qian et al. [30], investigated how BiLSTMA networks can be used to predict tool wear and how difficult it is to demonstrate how wear properties change over time. Their model examines sensor data and makes more accurate predictions than older methods by utilizing bidirectional long short-term memory (BiLSTM) architectures and attention mechanisms. This method demonstrates the importance of advanced neural network models in understanding and predicting how tool wear changes over time. He et al. [31] developed a deep multi-task network that uses sparse feature learning to monitor tool wear and machining quality simultaneously. Their framework combines deep learning methods to extract and connect features from various types of data. This gives them a comprehensive tool for determining how tools are performing and ensuring the quality of the products they produce. The findings show that multi-task learning and sparse representation work well together to improve prediction accuracy and operational efficiency in manufacturing systems.

Twardowski et al. [32] used acoustic emission techniques and machine learning methods to detect tool wear. Their findings show that AE sensors are effective at detecting changes in the condition of tools, and that machine learning algorithms can be used to analyze complex sensor data to determine how much wear something is experiencing. This study adds to the growing body of evidence arguing that AE monitoring and artificial intelligence should be used in tandem for proactive cutting tool maintenance.

The predicting and monitoring tool wear advances the field by introducing new approaches, such as combining wear-related factors, employing basic to advanced machine learning methods, and using acoustic emission signals for real-time monitoring. Each study demonstrates the potential for improving tool condition assessment accuracy and operational efficiency. However, they also highlight issues such as the need for a large amount of experimental data, the difficulty and computational requirements of advanced neural networks, and the feasibility of implementing these technologies in a variety of manufacturing environments. A thorough examination reveals that more research should be conducted on how to make these methods work in the real world with various types of machining operations, how to make them more scalable, and how to make complex models easier to understand. As this body of work demonstrates, tool wear monitoring research is constantly evolving. It predicts a future in which technology is increasingly used in manufacturing processes.

Due to the low specificity and sensitivity of existing tool wear prediction methods, it was determined from this analysis that using acoustic emission signals as input to machine learning based tool wear prediction is an additional viable research aim. Here, feature engineering for educating machine learning models is the main concern.

To better forecast tool wear from acoustic emission signals, this research contributes by concentrating on temporal and spectral aspects and an optimization approach based on a statistical diversity evaluation technique to overcome these limitations.

3. Methods and Materials. The processes and resources that went into developing the suggested model are outlined below. The features that will be used throughout the proposal's learning phase, as well as the appropriate features extraction from the provided labelled data. This section also discusses the strategy used to compare the projected values of an attribute over the provided records of the both positive and negative class labels. Moreover, this part investigates the appropriate classifier for detecting acoustic emission signals. We have also looked at the proposal's testing phase (identifying acoustic emission signals).

3.1. Acoustic Emission Signals. Transient elastic waves are produced when a material undergoes a quick redistribution of stress, a phenomenon known as acoustic emission (AE). When a structure is stressed (by load, pressure, as well as temperature change), energy is released from localized sources as stress waves, which travel to the surface and yet are measured by sensors. Picometer (10-12 m) movements can be detected with the correct equipment. AE can be triggered by a variety of events, including earthquakes, rock explosions, fissures, slip and dislocation motions, melting, twinning, and phase transitions in metals. In the case of composites, acoustic emissions result from matrix cracking, fiber breaking, and debonding. There are other AEs in polymers, wood, and concrete.

A material discontinuity's origin and importance can be revealed via its AE signal. Besides its extensive usage in academic studies, AET (acoustic emission testing) is also routinely used in a variety of industrial

contexts (e.g., evaluating structural integrity, locating defects, determining leak rates, and keeping tabs on weld quality).

In testing to conventional NDT techniques, Acoustic Emission stands apart in two key aspects. To begin, let's talk about the origin of the signal. Instead of supplying energy, AET just listens for energy emitted by the item being studied. Acoustic emissions can be triggered and propagated in operational structures by applying appropriate loading during AE testing. The discipline of AET is dedicated to the investigation of material dynamics. Not inactive characteristics (like a crack that isn't expanding) are shown. It is critical to detect both new and persistent flaws. Defects could be unnoticed when any loading becomes too low to trigger an acoustic event. AE testing provides instant feedback on the reliability of a part. With AET, there is no need to disassemble or clean the specimen, and yet a thorough volumetric assessment can be performed quickly and accurately thanks to the use of several sensors and a permanent mounting system that also allows for process control.

3.2. Preprocessing of the AE signals. In the process of digitizing the gathered sound signals, they frequently mix in with signals from other sources whose waveforms and arrival timings are not known. In order to extract the physical characteristics that best correspond with tool wear, signal processing is employed as part of the pre-processing stage of TCM systems for machining. Sound characteristics are extracted using a variety of signal processing techniques [33]. In this investigation, 1-second samples were taken from sound signals and analysed in the temporal domain. Frequency domain analysis samples were processed using FFT (Fast Fourier Transform)-based DTFT (Discrete-Time Fourier Transform). This research used a single absolute value from the Fourier transform to build a decision-making model. In order to resolve the DTFT-PR (preset resolution) issue in future research, the WT (Wavelet Transform) will be used.

Sound signals under the same cutting settings and at good and tool wear levels are displayed as a frequency spectrum in Figure 3.1. The spectrum was shown logarithmically to draw attention to the low-frequency band, where signals include more information about tool wear. Multiple frequency amplitudes are shown to correlate to tool wear in the images, highlighting the need for a sophisticated decision-making model to accurately depict these relationships.

3.3. Features. Temporal characteristics are those in the time domain, including signal energy, less energy, 0% of crossing, and many more, that are easy to extract and have physical comprehension.

- Since the stochastic process stated as a mathematical approach for acoustic emission only considers random variables, acoustic emission spectrum analysis is often more advanced than deterministic signals analysis, such as sinusoids.
- Static acoustic emission signals require a spectral phase that is completely random and, thus, devoid of information. The use of constant acoustic emission signals may be to blame for this phenomenon; by definition, it cannot occur randomly at times. Therefore, the phase idea is irrelevant for acoustic emission signals, which is a major difference between these signals as well as deterministic spectrum signals.
- So, spectral properties like frequency components, spectral density, fundamental frequency, and many more have been achieved by transforming a time-based signal into the frequency domain using Fourier-transform (FT). In this context, these characteristics might be put to use in identifying tones, rhythms, melodies, and pitches.
- Audio signals are often complex, made up of many individual shock wave of a fixed wavelength that propagate as a single perturbation in the medium. While the sound is being captured, we record the amplitudes of the ensuing waves. To break down a signal into its frequency components, mathematicians use a concept known as the Fourier Transform (FT). FT also provides the magnitude of each frequency in the signal, in addition to the frequency spectrum.

In the study of sound, novices often employ the broad spectral characteristic, which may take the form of a rising, decreasing, or varying frequency. The FT creates harmonic sound generators called as signal's spectral structure in acoustic emissions. As said before, there are a variety of ways this might be demonstrated. In other words, if the initial partial harmonic of a periodic signal, whose vibrations have repeated throughout time, is 100 Hz, the subsequent partial harmonics will be 200 Hz, 300 Hz, etc. This may, however, be an unusual occurrence in the realm of sounds created by nature. Make a note to the effect that no spectral characteristics

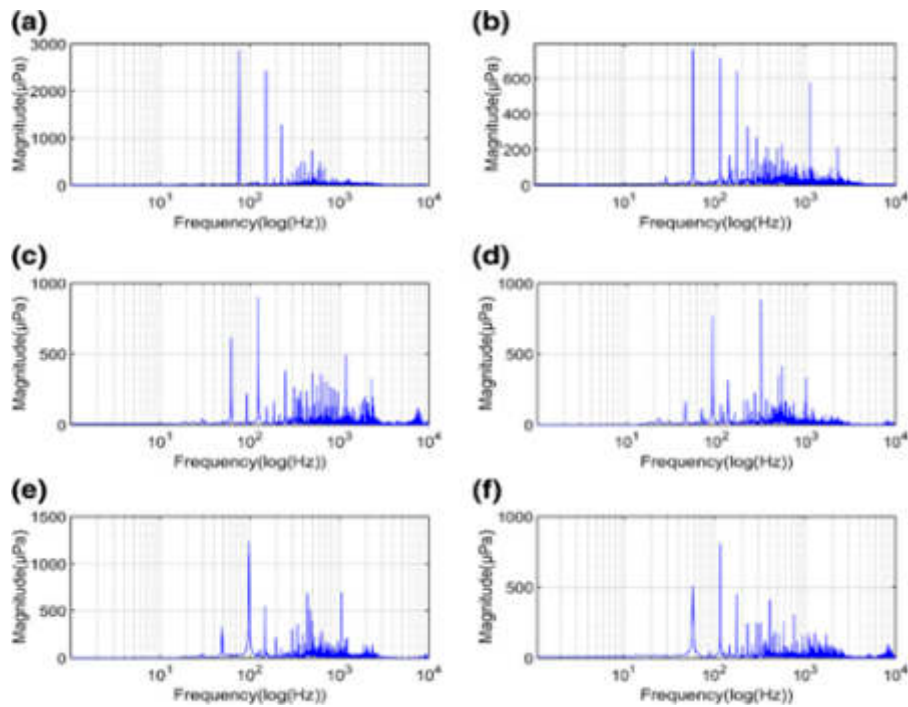


Fig. 3.1: Sound signals emitted from worn and benign tools at the same cutting settings

or devices are related to any suggested laws in the actual world where we could perceive sound. MFCC ("Mel cepstral frequency coefficients") as well as "predictive linear perceptive" (PLP) feature were the two best spectral features used in the acoustic emission. In this case, it may be understood that the frequency content of acoustic emissions is evaluated by the cochlea in the eardrum. Bandpass filters with a constant Q rank made it possible to modify the inspection of the basilar membrane. There is also the presence of essential bands, which rise to mask phenomena, where a burst or one strong acoustic emission may hide another lesser acoustic emission inside an essential band. The features of the auditory system that allow for the identification of acoustic emission are captured by both PLP and MFCC. Nonetheless, the strength of these characteristics was not very pronounced with respect to the acoustic emission.

The acoustic emission is also supported by a few time-domain characteristics, such as the plosion index and the highest coefficient of correlation.

In many complicated signals, such as acoustic emission, the signal's characteristics change over time. Small-time FT plays an important role in the interpretation of divergent acoustic emissions and may be more useful for referring to changes in frequency components over a finite time period. Integrating temporal envelopes over short time periods allows FT to derive spectral characteristics, while temporal-features are generated by turning envelopes into frequency modulation devices.

Temporal characteristics included things like ZCR, energy, etc., whereas spectral features included things like MFCC, GTCC, and several others in the frequency domain.

3.4. Kolmogorov-Smirnov test (KS Test). The significance of any differences between samples may be swiftly determined by using the Kolmogorov-Smirnov test [34]. As a uniformity, it ensures that all values in a distribution are consistent with one another. The Kolmogorov-Smirnov test allows one to examine the uniformity of a distribution, an important feature. In addition to comparing two multi-dimensional probability distributions, the Kolmogorov-Smirnov test may also evaluate the similarity of two one-dimensional distributions. It can immediately tell whether there is a difference between two samples. The Kolmogorov-Smirnov statistic evaluates the dissimilarity between the empirical distribution functions of two samples or between the

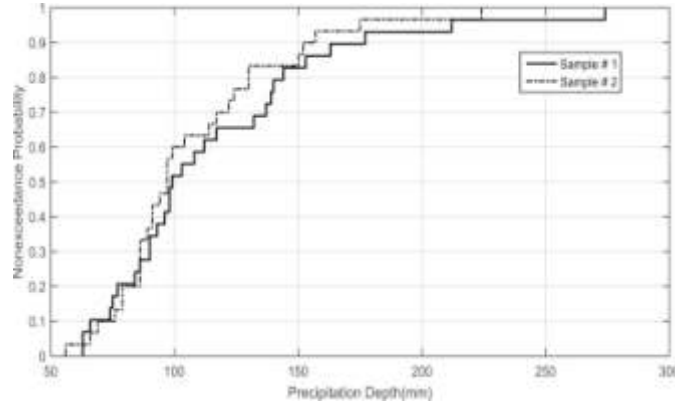


Fig. 3.2: Cumulative distribution function

sample and the reference distribution.

Distributions of observations in the two data sets may be compared using the Kolmogorov-Smirnov (KS) test [34]. Each dataset’s values are assumed to come from a given continuous distribution (the null hypothesis). As an alternate hypothesis, these data sets come from several continuous distributions. It is possible to do the hypothesis test with a 5% level of statistical significance. Cumulative distribution functions (CDF) for two different datasets are depicted in figure 3.2. The null hypothesis is supported by a two-sample KS test, which returns the D value of 0.1782 as well as a P-value of 0.694. At the 5% level of significance, the KS test on such two datasets failed to disprove the null hypothesis, showing that the given samples are drawn from the identical continuous distribution. This test can only conclude that, the distributions are distinct; it cannot reveal any changes to the mean, variance, or extremes.

The usage of KS-Test method explained in following description. In order to estimate the diversity between two vectors v_a, v_b representing the samples of two different distributions, $ks - test(v_a, v_b)$ Given two vectors v_a and v_b , the aggregate values of these vectors are represented as follows: Eq 3.1, Eq 3.2

$$\|v_a\| = \sum_{i=1}^{|v_a|} \{e_i \mid e_i \in v_a\} \tag{3.1}$$

$$\|v_b\| = \sum_{i=1}^{|v_b|} \{e_i \mid e_i \in v_b\} \tag{3.2}$$

The cumulative ratio for each sample in the vectors v_a and v_b is predicted as follows (Eq 3.3):

$$\forall_{(i=1)}^{|v_a|} \left(v_a^{cr} \leftarrow \sum_{j=1}^i \left(\frac{e_j}{\|v_a\|} \mid e_j \in v_a \right) \right) \tag{3.3}$$

Here, Eq 1 determines the cumulative ratio of each e_i of the elements listed in vector v_a . The $\|v_j\|$ indicates the aggregate value of the samples listed in the vector v_a , and v_a^{cr} is a set that comprises cumulative ratios of the samples listed in the vector v_a .

Further, it discovers cumulative ratios v_b^{cr} of the samples listed in vector v_b .

The absolute distance of "cumulative ratios" corresponding to values that are present at a similar index of both v_a and v_b vectors is found as: Eq 3.4

$$\forall_{(i=1)}^{(\max(\|v_a^{cr}\|, \|v_b^{cr}\|))} \{D[i] = \sqrt{(v_a^{cr}[i] - v_b^{cr}[i])^2}\} \tag{3.4}$$

Eq 2 for each index i , represents the "absolute distance $D[i]$ " of "cumulative ratios $v_a^{cr}[i]$, $v_b^{cr}[i]$ ".

The maximum value of the set D is found as d -stat. Then, find d -critic (from the KS-Table) at a given "degree of probability threshold" p_τ (0.01, 0.05, or 0.1) for vector aggregates $\|v_a\|$, $\|v_b\|$.

The return value is determined by the following condition:

$$\begin{cases} \text{false} & \text{if } d\text{-stat} > d\text{-critic} \\ \text{true} & \text{if } d\text{-stat} \leq d\text{-critic} \end{cases} \quad (3.5)$$

If d -stat is greater than d -critic, then there is no diversity between the given vectors' distribution. Hence, return false; else return true.

3.5. The classifier. Adaboost was designed to improve the performance of binary classifiers by using an ensemble learning approach. Adaboost uses iterative learning to learn the shortcomings of weak classifiers. The method of adaptive boosting [35] is utilised by the Adaboost classifier. Multiple, rather weak Boolean classifiers are combined to form this weak classifier. Every weak classifier has a single true/false criterion to divide the data. False-positives as well as false-negatives might be separated by other Weak Classifier. This continues until the WC calls time on the competition. The output of each WC will be combined to form the final output of the classification operation. In this section, the WC represents the proposed model's binary-classification optimal features, which are characteristics that are both consistent with the class and different from it.

Weak-Classifier iterations of the classification process allow the part of the corpus that didn't need to be categorized exactly to be boosted for the subsequent classifier iteration. Classification weight (WC) is utilised at each iteration to show the relative importance of different categories. Iterative calls to the WC will result in a standardization of classified records. Adaboost WC provide a concrete n-gram for precision classification and provide a justification for record polarity classification in the proposed approach. The suggested approach would use the WC to learn the suitability of each feature category for both positive and negative sentiment training corpora.

All data points are given equal consideration when Adaboost generates a model. Incorrectly categorized data are given more weight. In the following paradigm, more significant information is given more weight. Models will be trained until there is a noticeable improvement in accuracy. What follows is a description of the internal structure of this algorithm. Before doing binary classification, Adaboost assigns weights to data points. At the outset, irrespective of the overall data points, each data point has the same weight. In order to this, produce a decision stump from each of the features and then assess its tree accuracy using the Gini Index. In terms of the Gini Index, the first tree stump represents the worst possible result. Classified data points using this tree's alpha ("Importance" or "Influence") is based on the following Eq 3.6.

$$2^{-1} \cdot \log_e \left(\frac{1 - \|w\|}{\|w\|} \right) \quad (3.6)$$

Aggregating the all of the sample weights related to data points that are incorrectly labelled can define the total error $\|w\|$.

Consider a dataset having 5 data points as well as the assumption that there is 1 incorrect output; the total error then will be 1/5 (ranges between 0 and 1), then the stump's performance (alpha) will be: Eq 3.7, Eq 3.8, Eq 3.9, Eq 3.10

$$\alpha = 2^{-1} \cdot \log_e \left(\frac{1 - 5^{-1}}{5^{-1}} \right) \quad (3.7)$$

$$\alpha = 2^{-1} \cdot \log_e \left(\frac{0.8}{0.2} \right) \quad (3.8)$$

$$\alpha = 2^{-1} \cdot \log_e(4) = 2^{-1} \cdot (1.386) \quad (3.9)$$

$$\alpha = 0.693 \quad (3.10)$$

The error rate for alpha () ranges from 0 to 1, where 0 Indicates perfect stump and 1 indicates awful stump (see figure 3.3).

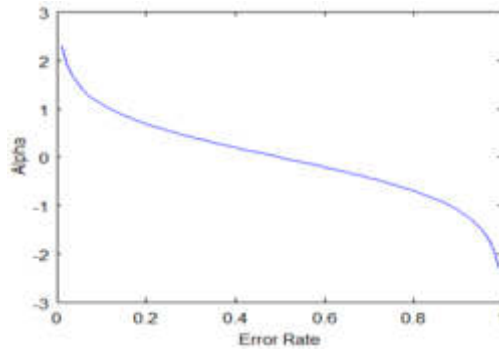


Fig. 3.3: Alpha versus error rate

As can be seen in the above graph (figure 3.3), whenever there is no misclassification, there is no error (Total Error = 0), and hence the alpha is quite high.

The classifier’s alpha will be zero if its predictions are half accurate and half wrong, with a total error of 0.5. A negative alpha value indicates that the classification was erroneous for all samples. If this is the case, the error is extremely large (approximate to 1), and the significance level will be low.

When the same weights being used to the next model, therefore the output obtained is identical to what’s been obtained in the first model, hence updating the weights is required in order to determine the total error $\|w\|$ as well as performance of a stump.

Correct forecasts will have their weights lowered while incorrect guesses will be given greater weight. After adjusting the weights, we will now give the data points with maximal weights more preference when building our next model.

It is important to update the weights that use the given formula after the classifier’s significance and the total error have been determined (Eq 3.11):

$$cw = w \cdot e^{\pm\alpha} \tag{3.11}$$

New weight cw , the alphas being negative whenever the sample is correctly classified. The alpha is being positive whenever the sample is miss-classified.

3.6. Tool Wear Prediction Model. The set SC of acoustic emission signals that are taken as input for training phase of the classification process shall be processed to digital format and then bi-part in to sets P and N , where set P represents the acoustic emission signals of the wearied tools (labeled as positive). In contrast, set N represents the acoustic emission signals of the benign tools (tools that are not wearied), which are labeled as negative.

Further, derives all the temporal and spectral features of acoustic emission signals listed in the set P as two different matrices named as tP, sP , which are representing temporal and spectral features in respective order. Here, each row of the both matrices representing the respective features of the corresponding acoustic emission signal listed in the set P . Similarly, derives all the temporal and spectral features of acoustic emission signals listed in the set N as two different matrices named as tN, sN , which are representing temporal and spectral features in respective order.

Further step of the learning phase discovers optimal temporal and spectral features of the both positive and negative labels as described below (Eq 3.12, Eq 3.13):

$$\bigvee_{(i=1)}^{\max(|tP|, |tN|)} \{tP[i], tN[i] \mid |tP[i]| > 0 \vee |tN[i]| > 0\} \tag{3.12}$$

// Eq 3.12 For each iteration, considers the vectors $tP[i], tN[i]$ representing the values listed in $i^{(th)}$ column of the matrices tP and tN in respective order.

$$\text{if}(\text{ks-test}(tP[i], tN[i]) \equiv \text{false}) \quad tP \setminus tP[i] \wedge tN \setminus tN[i] \quad (3.13)$$

This Eq 3.13 discovers the diversity state between two vectors $tP[i]$ and $tN[i]$ through KS-Test, $\text{ks-test}(tP[i], tN[i])$. If the diversity state is false, then the columns of the matrices tP , and tN represented by the vectors $tP[i]$, and $tN[i]$ in respective order are discarded. Conversely, if the diversity state is true, then the feature represented by the i th column of the matrices tP , and tN will be considered as an optimal feature towards the positive and negative labels.

The resultant matrices tP and tN retains the columns representing optimal features toward both positive as well as negative classes in respective order. Similarly, the KS-test shall be applied on each pair of the columns listed at the same index of the matrices sP and sN that are representing spectral features of the positive as well as negative classes in respective order, which results matrices sP and sN with optimal spectral features of the both lasbels. Further phase of the learning process builds ensemble of weak classifiers to perform adaptive boosting-based classification of the given acoustic emission signals as the signal emitted from the wearied tools and signals emitted from the benign tools.

The following diagram presented as figure 3.4 outlines the intricate process of the Tool Wear Prediction Model, illustrating the transformation of acoustic emission signals into a digital format and their subsequent classification. Through various stages, including feature extraction, optimization, and iterative adaptive boosting, the model aims to precisely predict the wear of tools. This visualization offers a comprehensive understanding of the complex mechanics underlying the process.

Acoustic Emission Signals Phase

- Acoustic emission signals are collected in set SC .
- These signals are then processed into a digital format.
- The digital signals are bifurcated into two sets, P and N , representing wearied (positive) and benign (negative) tools, respectively.

Temporal and Spectral Features Phase

- Temporal and spectral features are extracted from the signals in sets P and N .
- These features are organized into four different matrices: tP , sP for positive and tN , sN for negative classes, representing temporal and spectral features, respectively.

KS-Test & Optimal Features Phase

- The KS-test is applied to discover optimal temporal and spectral features for both positive and negative labels.
- Resulting optimal features are represented by matrices $tP\&tN$ and $sP\&sN$.

Weak Classifier Iterations Phase (Adaptive Boosting Process)

- AdaBoost WC (Weighted Classifier) is used to initiate the weak classifier iterations.
- A decision stump is created from each feature, and its tree accuracy is evaluated using the Gini Index.
- The alpha (importance or influence) and weights are updated, and weights are given to the incorrect guesses while reducing the correct ones.
- The iterative process continues, building upon the previous model and adjusting weights to give more preference to data points with maximal weights.
- The final model is created through iterative calls to the weak classifier, progressively emphasizing the importance of certain features.

Additional Notes

- There is an iterative updating of weights using a specific formula: $cw = w \cdot e^{(\pm\alpha)}$. This allows for the progressive adjustment of the model based on correctly and incorrectly classified samples.

3.7. MLTWP Algorithm Flow. Given a set of acoustic emission (AE) signals, the following algorithm outlines the steps to predict tool wear:

1. Collect AE signals into a set SC and process them into a digital format.
2. For positive set P and negative set N ,
 - Extract temporal and spectral features, and store in matrices tP , sP , tN , and sN respectively.
3. Apply the Kolmogorov-Smirnov (KS) test to identify optimal features for each i^{th} feature:

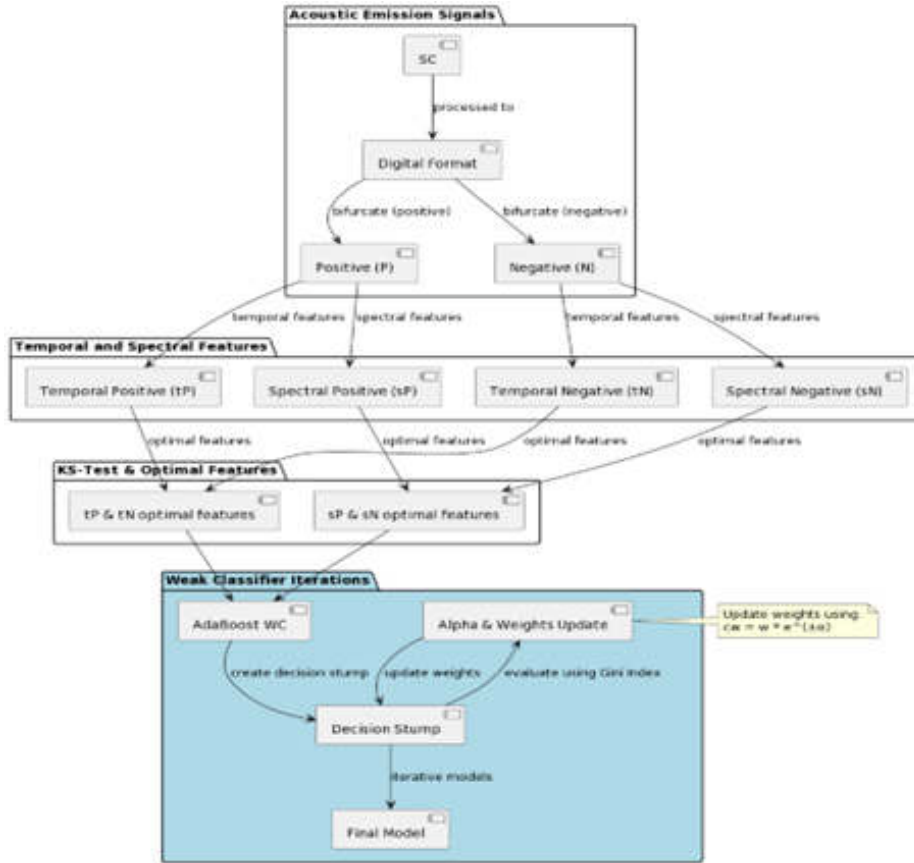


Fig. 3.4: The flow diagram of the MLTWP

- if $\text{KS_test}(tP[i], tN[i])$ is significant, retain $tP[i]$ and $tN[i]$.
 - if $\text{KS_test}(sP[i], sN[i])$ is significant, retain $sP[i]$ and $sN[i]$.
4. Initialize AdaBoost with weak classifiers derived from the optimal features.
 5. For each iteration k , perform:
 - (a) Create a decision stump h_k from the features.
 - (b) Calculate the error err_k of h_k .
 - (c) Compute the weight α_k of h_k :

$$\alpha_k = \frac{1}{2} \log \left(\frac{1 - err_k}{err_k} \right).$$

- (d) Update weights w for each data point:

$$w \leftarrow w \cdot e^{-\alpha_k y h_k(x)},$$

where y is the true label of x .

6. Combine the weighted decision stumps to form the final model M :

$$M(x) = \text{sign} \left(\sum_k \alpha_k h_k(x) \right).$$

7. Predict the wear condition of a new AE signal x using $M(x)$.

Table 4.1: Assumptions for Experimental Setup

	Description of Assumption
Assumption 1	All acoustic emission signals are accurately recorded without data loss.
Assumption 2	The dataset is fully representative of all possible states of tool wear.
Assumption 3	The labeling of the dataset into 'positive' and 'negative' is error-free.
Assumption 4	The temporal and spectral feature extraction processes are lossless.
Assumption 5	The Kolmogorov-Smirnov test reliably identifies the optimal features.
Assumption 6	The training and testing datasets are randomly partitioned to avoid bias.
Assumption 7	The four-fold cross-validation provides an unbiased estimate of performance.
Assumption 8	The AdaBoost algorithm converges to an optimal set of weights and stumps.
Assumption 9	The Gini Index is an adequate measure of decision stump performance.
Assumption 10	The MLTWP and TWPAE models are implemented under the same conditions.
Assumption 11	The performance metrics (precision, specificity, etc.) are computed without computational errors.
Assumption 12	The underlying distribution of the data remains stationary during the experiment.

The algorithm utilizes iterative boosting to refine the classification model, enhancing its predictive accuracy for tool wear based on AE signals.

4. Experimental Study. Experiments have been conducted on benchmark dataset to scale the performance of the proposed model MLTWP. The dataset adopted is a set of acoustic emission signals recorded from cutting mill tools both wearied and benign. Here, the acoustic emission signals of the wearied tools have been labelled as positive and the acoustic emission signals of the benign tools have been labeled as negative. The experimental data produced from the proposed model MLTWP and the state-of-the-art model Tool Wear Prediction using Acoustic Emission Signals (TWPAE) [27] have been compared in order to estimate the performance of the proposed model MLTWP. The dataset statistics have been explored in following section.

4.1. The dataset. The total records of the cutting mill dataset [36] are 25304, which possess the records of acoustic emission signals obtained from worn and benign tools. The acoustic emission signals of worn tools have been positive, which are of total 13318. The acoustic emission signals of unworn tools have been labeled as negative, which are of total 11986 records. 8990 negative records have been utilised for training, while there are 9989 positive records for training. Additionally, 2996 negative records and 3329 positive records have been tested. Further, each of the both positive and negative record sets will be partitioned in to four parts, such that 3 parts of the records from each label will be used to train the proposed model MLTWP and contemporary model TWPAE. The rest partition of the both positive and negative records will be used in testing phase. A table of assumptions for the experimental setup is as follows in Table 4.1

The table 4.1 summarizes the foundational assumptions made during the experimental phase of the study, ensuring clarity and transparency for replication and validation purposes.

4.2. The Performance Analysis. The performance of the proposed model MLTWP has been scaled through the results obtained for cross validation metrics from four-fold cross validation as shown in table 4.2, which have been compared with results of the fourfold cross validation metrics obtained from the experiments carried on the contemporary model TWPAE. The following discussion analyses the performance of the both proposed MLTWP and contemporary model TWPAE.

The obtained results for cross validation metrics demonstrate that the anticipated MLTWP model outperforms the existing TWPAE model. The precision statistic represents the percentage of correctly identified positive records relative to the total number of positive records. The accuracy of the predicted MLTWP model, as measured by four-fold cross validation, as 93%, 92%, 92, and 91%, in that sequence. On the other hand, the observed precision for the current technique TWPAE as 89.09%, 89%, 89%, and 88%. When comparing the suggested model MLTWP to the state-of-the-art TWPAE, the proposed model performs better on the cross validation metric called precision (see Figure 4.1).

Specificity might be defined as the fraction of records that were correctly categorised as negative relative to the total number of negative-records. Here, we see that the specificity for the MLTWP model were 92%,

Table 4.2: Statistics of cross validation metrics obtained from 4-fold classification scheme

PRECISION				
	FOLD#1	FOLD#2	FOLD#3	FOLD#4
MLTWP	0.9256	0.9201	0.9231	0.9191
TWPAE	0.8909	0.8851	0.8872	0.8839
SPECIFICITY				
MLTWP	0.9188	0.9132	0.9197	0.9118
TWPAE	0.8789	0.8714	0.8732	0.87
SENSITIVITY				
MLTWP	0.9085	0.8994	0.9022	0.9016
TWPAE	0.8908	0.8903	0.8975	0.8894
ACCURACY				
MLTWP	0.9134	0.9059	0.9089	0.9064
TWPAE	0.8851	0.8814	0.886	0.8803
F-MEASURE				
MLTWP	0.9222	0.9166	0.9199	0.9154
TWPAE	0.8849	0.8781	0.8801	0.8769
MCC				
MLTWP	0.8266	0.8118	0.8178	0.8127
TWPAE	0.7695	0.7621	0.7713	0.7599

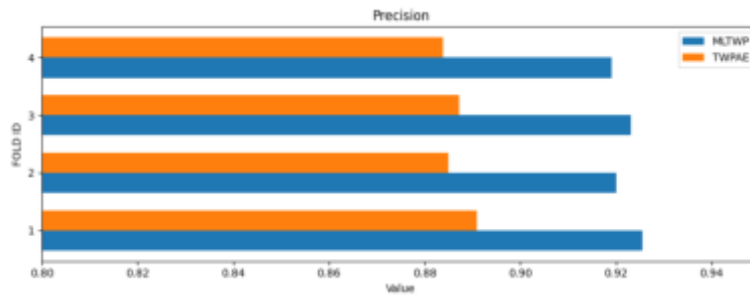


Fig. 4.1: Precision observed from four fold cross validation performed on MLTWP and TWPAE

91%, 92%, and 91% for four-fold cross validation; the corresponding figures for the TWPAE model were 88%, 87%, 87%, and 87%. In this case, it is assumed that the MLTWP approach outperforms the TWPAE model (see Figure 4.2).

In contrast to the actual number of delivered positive records for the testing, the test sensitivity measures the proportion of correctly identified positive records. Furthermore, sensitivity was observed from the MLTWP model over four-fold cross validation on different sets of records to be 91%, 90%, 90%, and 90% in that sequence, demonstrating optimal MLTWP performance when compared to the TWPAE technique (89%, 89%, 90%, and 89%) (see Figure 4.3).

The ratio of correctly identified positive and negative records relative to the total number of test records is an indicator of accuracy. The MLTWP method’s accuracy was shown to be optimum through four-fold cross validation (91%, 90%, 90%, and 90%), outperforming the accuracy of the other method TWPAE (89%, 89%, 89%, and 88%, respectively) (see Figure 4.4).

In Figure 4.5, we see a graph depicting a 4-fold increase in the F-measure metric between the anticipated MLTWP model and the existing TWPAE model. According to the results, the F-measure for the MLTWP model is 92%, 92%, 92%, as well as 92% after being subjected to four fold of cross-validation. In this case, the F-measures for the TWPAE model are as follows: 88%, 89%, 89%, and 88%. As a result, the F-measure for

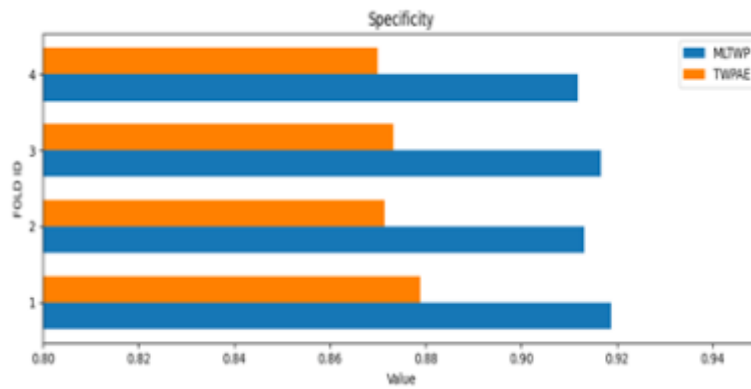


Fig. 4.2: Specificity observed from four fold cross validation performed on MLTWP and TWPAE

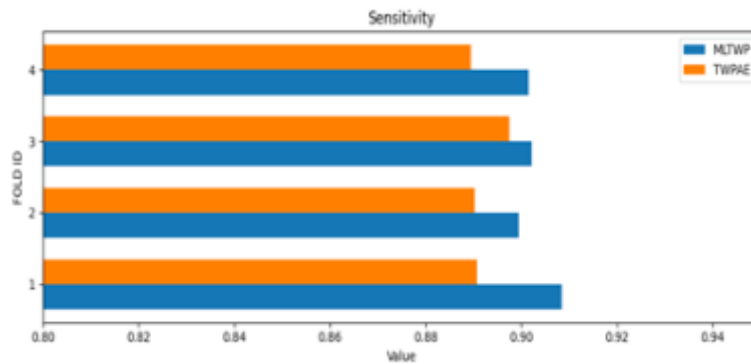


Fig. 4.3: Sensitivity observed from four fold cross validation performed on MLTWP and TWPAE

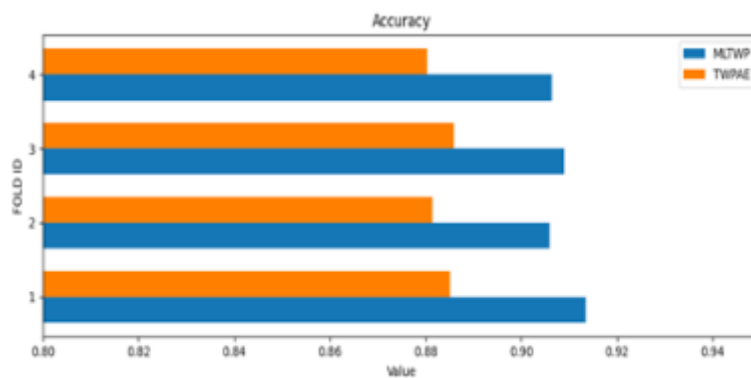


Fig. 4.4: Accuracy observed from four fold cross validation performed on MLTWP and TWPAE

the predicted MLTWP model is more significant than the conventional TWPAE technique.

Matthew's correlation coefficient (MCC) graph produced from four-fold cross validation of projected MLTWP model and current TWPAE model is shown in Figure 4.6. Using this data, we can determine that the MCC for

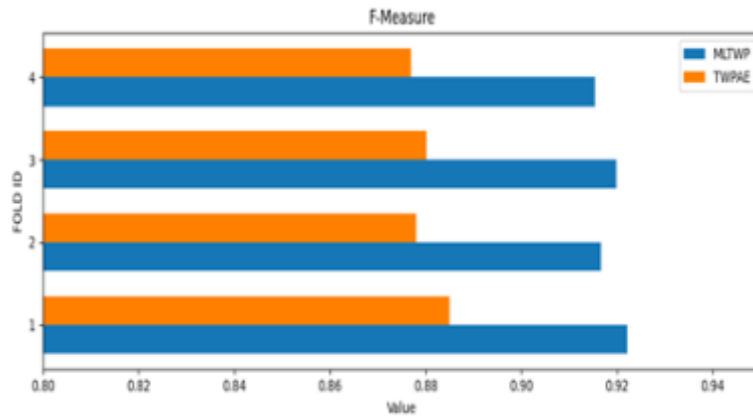


Fig. 4.5: F-measure observed from four fold cross validation performed on MLTWP and TWPAE methods

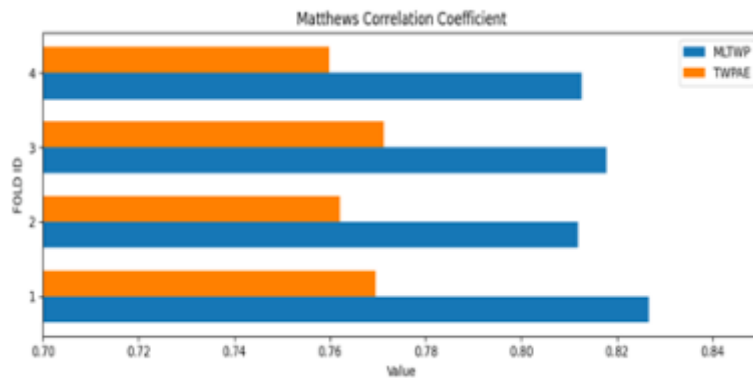


Fig. 4.6: Matthews correlation coefficient observed from four fold cross validation performed on MLTWP and TWPAE methods

the MLTWP model is, in sequence, 83%, 81%, 82%, and 81%. For the TWPAE model, the MCC is as follows: 77%, 76%, 77%, and 76%. It foresees unambiguously that MCC for the proposed MLTWP model is much more significant than the current TWPAE technique.

5. Conclusion. The Machine Learning Tool Wear Prediction (MLTWP) model, employing temporal and spectral features of acoustic emission signals, has demonstrated a marked improvement in predicting tool wear. Quantitatively, MLTWP has shown enhanced precision, specificity, sensitivity, accuracy, and Matthews correlation coefficient compared to the Tool Wear Prediction using Acoustic Emission Signals (TWPAE) model. In four-fold cross-validation, the MLTWP model has consistently outperformed TWPAE, affirming its robustness and efficacy. The significant gains in signal detection accuracy underscore the effectiveness of the feature selection and machine learning techniques implemented in the MLTWP model. Future work will delve into the analysis of quasi-identifiers from the optimal attributes identified by MLTWP to further refine tool wear prediction.

Data Availability Statement. The proposed Machine Learning based Tool Wear Prediction (MLTWP) from Variability of Acoustic Sound Emission Signals have been evaluated through cross validation using the benchmark data of cutting Mill dataset available at: [36]. <https://www.kaggle.com/datasets/shasun/tool-wear->

detection-in-cnc-mill

The results have been presented in the manuscript. The data used in this study is available publicly. However, the python code used to conduct the experiments is not publicly available. Nevertheless, the methodology and experimental setup used in this study have been described in detail, allowing for replication by other researchers using similar simulation tools or experimental setups.

Data and Materials Availability. The data and materials used in the proposed model, "Machine Learning based Tool Wear Prediction (MLTWP) from Variability of Acoustic Sound Emission Signals," have been described in detail in the manuscript. The study was conducted using the publicly available benchmark data, and the proposed model's performance was evaluated using 4-fold cross validation. The data and materials used in the proposed model are available in the manuscript.

REFERENCES

- [1] V. A. Pechenin, *et al.*, "Method of Controlling Cutting Tool Wear Based on Signal Analysis of Acoustic Emission for Milling," *Procedia Engineering*, vol. 176, pp. 246–252, 2017.
- [2] B. Li, "A Review of Tool Wear Estimation Using Theoretical Analysis and Numerical Simulation Technologies," *International Journal of Refractory Metals and Hard Materials*, vol. 35, pp. 143–151, Nov. 2012.
- [3] P. J. Arrazola, *et al.*, "Correlation between Tool Flank Wear, Force Signals and Surface Integrity When Turning Bars of Inconel 718 in Finishing Conditions," *International Journal of Machining and Machinability of Materials*, vol. 15, no. 1/2, p. 84, 2014.
- [4] A. Siddhpura and R. Paurobally, "A Review of Flank Wear Prediction Methods for Tool Condition Monitoring in a Turning Process," *The International Journal of Advanced Manufacturing Technology*, vol. 65, no. 1–4, pp. 371–393, Mar. 2013.
- [5] T. I. Liu, *et al.*, "On-Line Monitoring of Boring Tools for Control of Boring Operations," *Robotics and Computer-Integrated Manufacturing*, vol. 26, no. 3, pp. 230–239, June 2010.
- [6] R. Pullin, *et al.*, "Confidence of Detection of Fracture Signals Using Acoustic Emission," *Applied Mechanics and Materials*, vol. 7–8, pp. 147–152, Aug. 2007.
- [7] X. Li, "A Brief Review: Acoustic Emission Method for Tool Wear Monitoring during Turning," *International Journal of Machine Tools and Manufacture*, vol. 42, no. 2, pp. 157–165, Jan. 2002.
- [8] C. Zuluaga-Giraldo, *et al.*, "Acoustic Emission during Run-up and Run-down of a Power Generation Turbine," *Tribology International*, vol. 37, no. 5, pp. 415–422, May 2004.
- [9] Y. Zhang, *et al.*, "Tool Wear Condition Monitoring Method Based on Deep Learning with Force Signals," *Sensors*, vol. 23, no. 10, p. 4595, May 2023.
- [10] Y. Chen, *et al.*, "Predicting Tool Wear with Multi-Sensor Data Using Deep Belief Networks," *The International Journal of Advanced Manufacturing Technology*, vol. 99, no. 5–8, pp. 1917–1926, Nov. 2018.
- [11] E. J. Weller, *et al.*, "What Sound Can Be Expected From a Worn Tool?," *Journal of Engineering for Industry*, vol. 91, no. 3, pp. 525–534, Aug. 1969.
- [12] M. A. Mannan, *et al.*, "Application of Image and Sound Analysis Techniques to Monitor the Condition of Cutting Tools," *Pattern Recognition Letters*, vol. 21, no. 11, pp. 969–979, Oct. 2000.
- [13] T. Delio, *et al.*, "Use of Audio Signals for Chatter Detection and Control," *Journal of Engineering for Industry*, vol. 114, no. 2, pp. 146–157, May 1992.
- [14] D. R. Salgado and F. J. Alonso, "An Approach Based on Current and Sound Signals for In-Process Tool Wear Monitoring," *International Journal of Machine Tools and Manufacture*, vol. 47, no. 14, pp. 2140–2152, Nov. 2007.
- [15] C. Aliustaoglu, *et al.*, "Tool Wear Condition Monitoring Using a Sensor Fusion Model Based on Fuzzy Inference System," *Mechanical Systems and Signal Processing*, vol. 23, no. 2, pp. 539–546, Feb. 2009.
- [16] I. Ubhayaratne, *et al.*, "Audio Signal Analysis for Tool Wear Monitoring in Sheet Metal Stamping," *Mechanical Systems and Signal Processing*, vol. 85, pp. 809–826, Feb. 2017.
- [17] N. Seemuang, *et al.*, "Using Spindle Noise to Monitor Tool Wear in a Turning Process," *The International Journal of Advanced Manufacturing Technology*, vol. 86, no. 9–12, pp. 2781–2790, Oct. 2016.
- [18] A. Kothuru, *et al.*, "Application of Audible Sound Signals for Tool Wear Monitoring Using Machine Learning Techniques in End Milling," *The International Journal of Advanced Manufacturing Technology*, vol. 95, no. 9–12, pp. 3797–3808, Apr. 2018.
- [19] S. Orhan, *et al.*, "Tool Wear Evaluation by Vibration Analysis during End Milling of AISI D3 Cold Work Tool Steel with 35 HRC Hardness," *NDT & E International*, vol. 40, no. 2, pp. 121–126, Mar. 2007.
- [20] M.-C. Lu and E. Kannatey-Asibu, "Analysis of Sound Signal Generation Due to Flank Wear in Turning," *Manufacturing Engineering*, American Society of Mechanical Engineers, pp. 165–175, 2000.
- [21] F. J. Alonso and D. R. Salgado, "Application of Singular Spectrum Analysis to Tool Wear Detection Using Sound Signals," *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, vol. 219, no. 9, pp. 703–710, Sept. 2005.
- [22] Z. K. Peng, *et al.*, "An Improved HilbertHuang Transform and Its Application in Vibration Signal Analysis," *Journal of Sound and Vibration*, vol. 286, no. 1–2, pp. 187–205, Aug. 2005.
- [23] N. E. Huang, *et al.*, "The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time

- Series Analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, Mar. 1998.
- [24] A. G. Rehorn, J. Jiang, and P. E. Orban, "State-of-the-art methods and results in tool condition monitoring: a review," *International Journal of Advanced Manufacturing Technology*, vol. 26, pp. 693–710, 2005.
- [25] K. Prakash and A. Samraj, "Tool Flank Wears Estimation by Simplified SVD on Emitted Sound Signals," *2017 Conference on Emerging Devices and Smart Systems (ICEDSS)*, IEEE, 2017, pp. 1–5.
- [26] K. Prakash and A. Samraj, "Tool Wear Condition Monitoring using Acoustic Analysis of Emitted Sound Signals by Peak to Peak Analysis," *ISERD 93rd International Conference*, Hanoi, Vietnam, 2017, pp. 8–13.
- [27] J. L. Ferrando Chacón, *et al.*, "A Novel Machine Learning-Based Methodology for Tool Wear Prediction Using Acoustic Emission Signals," *Sensors*, vol. 21, no. 17, Sept. 2021, p. 5984.
- [28] C. Yang, *et al.*, "Tool wear prediction model based on wear influence factor," 2023.
- [29] C. L. Perumal, S. B. Bhadrinathan, and A. Samraj, "Tool Wear Condition Monitoring Using Emitted Sound Signals By Simple Machine Learning Technique," *DESIGN, CONSTRUCTION, MAINTENANCE*, vol. 2, 2022, pp. 168–172.
- [30] C. Qian, *et al.*, "Tool Wear Prediction Based on BiLSTMA Networks," *Proceedings of the 14th International Conference on Computer Modeling and Simulation*, 2022, pp. 103–108.
- [31] J. He, *et al.*, "Deep multi-task network based on sparse feature learning for tool wear prediction," *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, 2022.
- [32] P. Twardowski, *et al.*, "Identification of tool wear using acoustic emission signal and machine learning methods," *Precision Engineering*, vol. 72, 2021, pp. 738–744.
- [33] U. Zolzer, *Digital Audio Signal Processing*, 2nd ed, Wiley, 2008.
- [34] A. Ghasemi and S. Zahediasl, "Normality Tests for Statistical Analysis: A Guide for Non-Statisticians," *International Journal of Endocrinology and Metabolism*, vol. 10, no. 2, pp. 486–489, Dec. 2012.
- [35] T.-K. An and M.-H. Kim, "A new diverse AdaBoost classifier," *2010 International Conference on Artificial Intelligence and Computational Intelligence*, Vol. 1, IEEE, 2010, pp. 359–363.
- [36] "Tool Wear Detection in CNC Mill," available at Kaggle, 2023.

Edited by: Anil Kumar Budati

Special issue on: Soft Computing and Artificial Intelligence for wire/wireless Human-Machine Interface

Received: Jan 8, 2024

Accepted: Apr 8, 2024



RETRIEVAL OF TELUGU WORD FROM HAND WRITTEN TEXT USING DENSENET-CNN

RAJASEKHAR BODDU *AND EDARA SREENIVASA REDDY †

Abstract. The recognition of telugu hand written text is been one of the problems in many applications. To overcome the problem a deep learning technique is proposed in this work i.e. a Dense convolutional neural network (DCNN) model. A telugu dataset which is taken form IIIT-HW-Telugu is utilized to perform the proposed model. In this paper a four stage telugu word retrieval is performed, initially thinning of image is performed using morphological operation, secondly Densenet-CNN is applied for thinning image, thirdly perform OCR based image segmentation, finally two models like HARRIS and BRISK features to extract the features and evaluate information from the given input HWT images. The parameters evaluated are hamming distance, PSNR, MSE, Noise Sensitivity and rate of thinning. The proposed model outperformed well compared to other methods. The PSNR obtained using proposed model is 54.74, hamming distance is 1.2.

Key words: OCR, Morphological operations, Hilditch transform, CNN, DenseNet

1. Introduction. The never-ending desire to liberate information to flow in a digital format for easier access, dependence on archiving and preserving for longer term makes hand written text based word retrieval a highly intriguing subject of research. A large variety of digital libraries are emerging for the archiving of multimedia documents, including Universal Library (UL), Digital Library of India (DLI), and Google Books. These documents cannot always be saved as text. This increases the difficulty of finding pertinent papers. The cost of storage devices has decreased, and imaging devices are rising in popularity. This encourages scientists to work hard on creating effective methods for digitising and archiving massive amounts of multimedia material. Text, audio, picture, and video are all included in the multimedia data. Most of the items that have been archived so far are in books printed, while digital libraries are currently collections of document pictures. More specifically, digital material is saved as pictures that match to book pages.

Daily existence requires the use of images [1]. Every day, a large amount of data is created in the form of photographs by technological devices. The two main categories of image retrieval [2] are approaches based on text and second one is based in content. Each strategy has distinctive qualities [3] that may be applied to a variety of applications depending on the circumstance. Search by image content is discussed by author in [4] and is termed as content-based image retrieval (CBIR). With this method, the search examines the image's visual content rather than the keyword descriptions linked to it. Text-based image retrieval (TBIR) [5] relies only on keywords based on text and input to be descriptors, with text keywords also employed in the index pictures. The matched photos are fetched from the image repository by comparing the index images keywords with the supplied keywords. This strategy has the benefits of being simple to implement, quick retrieval, and web picture search. It is difficult to manually search through a vast collection of photographs for any image.

The author in [6] discussed about the study of computer vision and machine learning, where machine learning helps in developing the automated system of identification of scene from the text. This recognition of text in natural settings is a big challenge. English has been the primary language of research in the field of text categorization and domain identification. Regional languages, particularly Indian languages, have had far less of an impact. Telugu is a member of the Dravidian language family and is one of language which is older and traditional of Indians. Telugu is the sixteenth most spoken language in the world, with 93 million native

*Research Scholar, Department of Computer Science and Engineering, College of Engineering and Technology, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India. (rajsekhar.se15@gmail.com).

†Professor, Department of Computer Science and Engineering, College of Engineering and Technology, Acharya Nagarjuna University, Guntur, Andhra Pradesh, India.

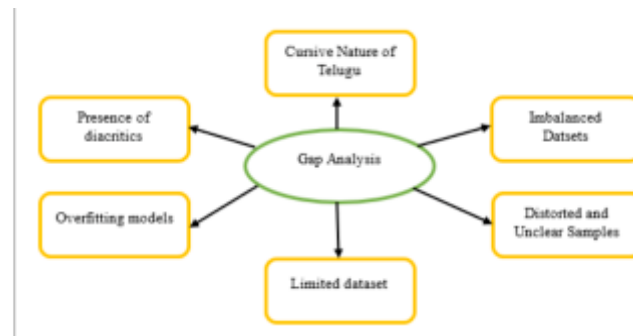


Fig. 1.1: Gap Analysis of Telugu Text Recognition [21]

speakers, according to the Ethnologic list. The gap analysis for any language is same as shown in figure 1.1.

It became a difficult process to retrieve telugu words from handwritten writing. Compared to other languages, telugu has a particular set of writing curves and strokes for its handwritten text. One of the cutting-edge methods for addressing issues in the sector of concern is deep learning. With the use of massive quantities of data and the Deep Learning technology, the nodes of one layer are connected to those of adjacent layers. The network is assumed to be more complex due to the number of tiers. Since deep learning systems handle massive volumes of data and perform challenging mathematical operations, they demand strong hardware. In this paper the utilization of DenseNet CNN is considered for effective retravel of telugu word from hand written text. Further in section 2 the discussion is about different techniques used previously. In section 3, the suggested framework is implemented and in section 4 the results obtained using proposed model is discussed.

2. Related Work. For the characterisation or acknowledgement of word pictures for different dialects, several writings are accounted for. When compared to the typical scanned texts in English writing, acknowledgment of terms from Telugu writings has not been studied as extensively. A part of the approaches mentioned are briefly discussed. The most popular highlights for creating bag-of-words (BoW) representation are those processed at interest focuses in scale-invariant feature transform (SIFT) [7] highlights. A histogram of the visual words serves as the word image's visual representation. "Data is lost when the highlights are quantized to a visual word. This is frequently believed to exhibit some measure of power (or invariance). In fact, there is still no consensus on how to select the vocabulary's range and retention techniques. Given a language, creating a BoW representation requires two key steps one is Coding and the other is pooling. With the use of the vocabulary words' histograms of repeated incidents, reports are properly expressed. The categorization and recovery of records are then carried out using these histograms.

Spatial pyramid matching (SPM) [8], which divides pictures into vertical and flat bearings, provides spatial request in common scene images. Word pictures are divided vertically into three portions and then recorded to provide order in representation. In [9], word pictures are shown as profile highlights, and Euclidean separation is unintentionally employed to imply comparison. Dynamic time wrapping (DTW) is used to coordinate word pictures in order to account for the variance in word image lengths. However, the focus of this study is to deduce an enforcement conspiracy using a back-end file structure. Due to this, DTW-based systems are not rational. Versatility in report recovery has also recently received significant attention. A list of 10 million pages using locally likely arrangement hashing (LLAH) was published in [10]. Recent efforts to recover strong records use the visual Bag of Words (BoVW) to represent and organise word pictures. One may quickly find significant papers from a million documents using BoVW representation and a reversed ordering scheme [11]. For the picture representation, feature points will be quantized, and a versatile representation is created by proposing a "vocabulary" over an element space. Ideally, code is generated from raw descriptors using vector quantization (VQ). The difficult issue of translating code words from the vocabulary to the feature vectors of an image is one limitation of the code-book technique. The challenging assignment presents the two problems of codeword credibility and vulnerability. The problem of selecting the proper codeword from at least two

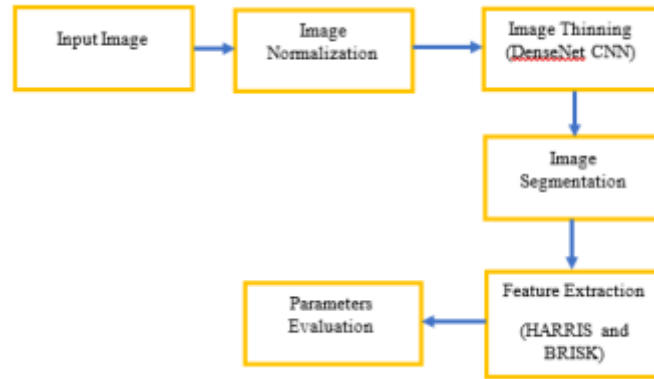


Fig. 3.1: Block Diagram of Proposed Work

significant candidates is known as codeword vulnerability. The VQ technique disregards the significance of several candidates and selects the most effective visual word. The problem of picking a codeword without a plausible hope in the lexicon is known as "codeword believability." The codebook method assigns the best-fitting codeword, although it is not a true representation. Authors in [12] proposed a soft assignment coding approach to overcome this restriction, in which each visual word is assigned a local characteristic based on its location.

In [13] author represented a spotting system of a Telugu word with improved performance over traditional system. It is based on correlation and hidden markov model (HMM) technology. In order to outperform BoVW and SIFT + BoVW, an algorithm is developed based on sped-up robust features (SURF) with the aid of BoVW. However, the word picture retrieval techniques described in the literature have some drawbacks, including inefficiency, increased complexity, and decreased precision with big data bases.

Contrarily, people in the south Telugu, an Indian alphabet, is composed of several elements, making the use of high-level feature extraction algorithms more difficult. On Indian languages, several techniques for domain identification and text classification have been developed, however only a small number of these studies have been reported on Telugu. This section provides an overview of a few strategies and approaches for text classification and domain identification. Automated text classification with a focus on Telugu is been developed in recent times. In his research, 800 Telugu news items were classified using supervised classification with the Naive Bayes classifier. KNN, Naive Bayes, and decision tree classifiers as text mining approaches to represent and categorise papers written in Indian. Telugu text documents can be categorised using the language-dependent and independent models suggested. Telugu texts were classified using a model for document organisation and categorization of texts using the word frequency ontology. The robustness of LSTM and CNN are combined in an attention-based multichannel CNN for text classification. In this network, CNN tracks word relationships while Bi-LSTM records word history and future information [14]. The author in [15] suggested a novel heuristic advanced neural network based telugu text categorization model (NHANNTCM) for extraction of telugu word and achieved an accuracy of 99%.

3. Methodology. In this paper thinning of the input image is the main concept. The thinning is performed using DenseNet CNN and followed by OCR based segmentation of letters and finally extraction of features using Harris and Brisk. The major goal of this study is to take use of the DL-CNN's robust performance in order to extract useful features with the least amount of time and effort possible. The convolutional phase known as CNN extracts features from pictures by acting as a visual descriptor. Each image is altered by the application of a series of filters, resulting in new image types known as convolution maps. The proposed model is shown in figure 3.1.

3.1. Input Dataset. Handwriting recognition (HWR) in Indic scripts is a challenging problem due to the inherent subtleties in the scripts, cursive nature of the handwriting and similar shape of the characters. Lack

విశ్వవిద్యాలయం అర్హత ఉత్తీర్ణులయిన

Fig. 3.2: Example of input words

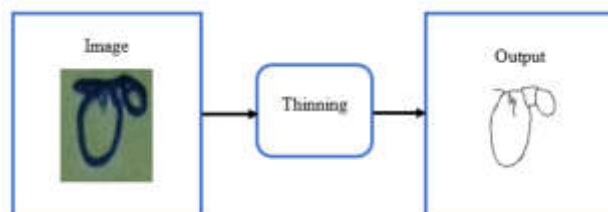


Fig. 3.3: Concept of Thinning

of publicly available handwriting datasets in Indic scripts has affected the development of handwritten word recognizers. In order to help resolve this problem, 2 handwritten word datasets: IIIT-HW-Dev, a Devanagari dataset and IIIT-HW-Telugu, a Telugu dataset [16]. In this paper, Telugu Dataset (IIIT-HW-Telugu) with a file size of 3.7 GB is considered to evaluate the proposed model. The samples of images considered as input is shown in figure 3.2.

3.2. Image Normalization. To extract various features from a picture on the same structure for image normalisation, all randomly sized images are downsized into the same size images. Here, using a bilinear standard transformation, photos of arbitrary sizes are normalised to be 100x200 size and the aspect ratio of image should not get disturbed. The numerous distortions, including missing segments with distortion, distortion caused randomly, effects of noise, and the segments which are missing in query terms, are also calculated, and analysed using this normalisation procedure.

3.3. Thinning of Image. Thinning is the technique of removing unused pixels from an image in order to recover the skeletons. It is also known as the Skeletonization process. Black foreground pixels are removed repeatedly, layer by layer, in this morphological procedure until a one-pixel-wide skeleton is reached. It entails reducing something to its tiniest size. Typically, binary pictures made up of black (foreground) and white (background) pixels are subjected to skeletonization. It accepts a binary picture as input and outputs another binary image, as seen in Figure 3.3.

The use of a deep convolutional neural network to thin an input picture is covered here. When compared to other algorithms, DCNN-based image thinning produces accurate Skelton estimates of the input picture. The deep network utilized in this work is densenet.

3.4. DenseNet CNN. Following is an explanation of how CNN operates: using the two-dimensional layer of convolution given input is paired up with sliding filter. The convolution of layers for the input is computed using the dot products of weights and input, in this process the set of filters are moved in vertical direction and horizontal direction towards the input. Threshold process is done in the ReLU layer by converting the values to zero which are lower than zero. Later down sampling is performed in the max pooling layer by identifying the maximum level of every zone by splitting the input into a rectangle pooling region. Bias vector is added to the fully connected layer, before performing this addition the input is multiplied by a weight matrix. Figure 3.4 depicts the overall architecture of deep CNN.

Here, we will provide a summary of the DenseNet design for convolutional networks, which stands for densely linked convolutional networks. The issue that the convolutional neural network is seeking to address with the density of the design is to deepen it. Dense nets are networks of convolution with plenty of connections.

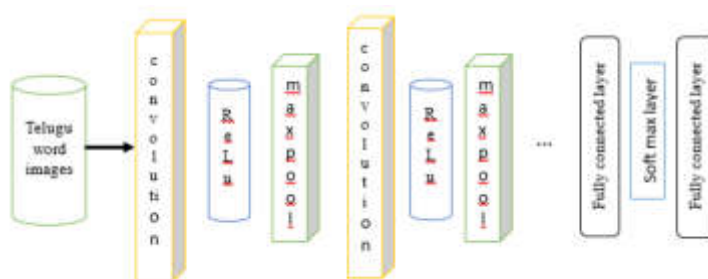


Fig. 3.4: Architecture of DCNN

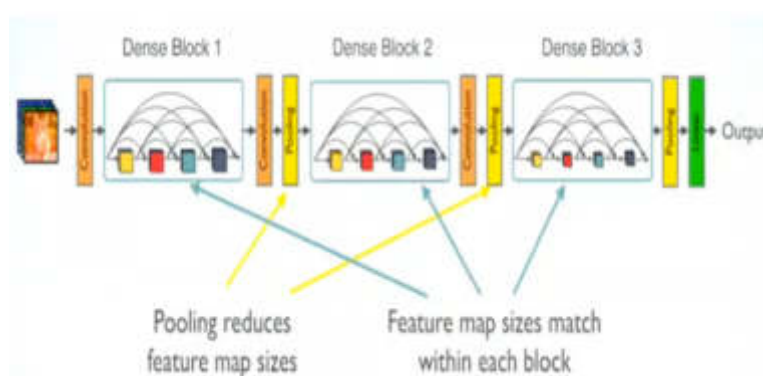


Fig. 3.5: DenseNet Architecture [16]

It is quite like a ResNet, with a few key differences. While ResNet employs an additive technique and accepts a previous output as an input for a subsequent layer, DenseNet utilizes every previous output as a source for a new layer. Figure 3.5 conveys the detailed structure of DCNN.

The drawback of this network is that it gets sort of unsustainable as we go further into it. For example, if the second layer is set for moving towards third layer, then the third layers need to be sourced with the second layer and the other layers from the initial state. In this network dense blocks are created for which different filters are created at every block but the size of feature map are well in constant for every block internally. The layer present in dense network is transition layer, these layers are handled by down sampling. This down sampling is performed by applying normalization of batch, one to one convolution and two to two layers of pooling.

The thinning of input image is performed deeply by carrying every layer of the telugu hand written word to the next layer until a better output is achieved. Figure 3.6 the densenet performance helps in thinning the image with more accurate output based on the working performance of densenet.

3.5. Image Segmentation. In this section, we will go over the segmentation techniques, which are yet another crucial stage of the OCR system. Simply divided into smaller segments for subsequent processing, segmentation is the act of taking a whole picture. Using word level segmentation, a picture is segmented. The input is divided into separate letters, making it easier to recognise the word.

We are given an image with a single line made up of a string of letters at this level of segmentation. As seen in Figure 3.7, the goal of word level segmentation (WLS) is to separate the picture into its component letters.

3.6. Feature Extraction. Feature extraction forms an important part of in retrieval of words from hand written text. In many applications, the speed of feature detection in a picture is critical. To compute the

Layers	Output Size	DenseNet-121	DenseNet-169	DenseNet-201	DenseNet-264
Convolution	112 × 112	7 × 7 conv, stride 2			
Pooling	56 × 56	3 × 3 max pool, stride 2			
Dense Block (1)	56 × 56	1 × 1 conv 3 × 3 conv × 6	1 × 1 conv 3 × 3 conv × 6	1 × 1 conv 3 × 3 conv × 6	1 × 1 conv 3 × 3 conv × 6
Transition Layer (1)	56 × 56 28 × 28	1 × 1 conv 2 × 2 average pool, stride 2			
Dense Block (2)	28 × 28	1 × 1 conv 3 × 3 conv × 12	1 × 1 conv 3 × 3 conv × 12	1 × 1 conv 3 × 3 conv × 12	1 × 1 conv 3 × 3 conv × 12
Transition Layer (2)	28 × 28 14 × 14	1 × 1 conv 2 × 2 average pool, stride 2			
Dense Block (3)	14 × 14	1 × 1 conv 3 × 3 conv × 24	1 × 1 conv 3 × 3 conv × 32	1 × 1 conv 3 × 3 conv × 48	1 × 1 conv 3 × 3 conv × 64
Transition Layer (3)	14 × 14 7 × 7	1 × 1 conv 2 × 2 average pool, stride 2			
Dense Block (4)	7 × 7	1 × 1 conv 3 × 3 conv × 16	1 × 1 conv 3 × 3 conv × 32	1 × 1 conv 3 × 3 conv × 32	1 × 1 conv 3 × 3 conv × 48
Classification Layer	1 × 1	7 × 7 global average pool 1000D fully-connected, softmax			

Fig. 3.6: DenseNet Description [20]



Fig. 3.7: OCR Image Segmentation

correspondence between numerous perspectives effectively and accurately, the detected feature points must be represented independently. Fast feature recognition, description, and matching are necessary for real-time processing of the pictures. The points used to characterise the images must meet two crucial requirements in order to achieve better feature matching of image pairs: first, the feature points of the same strokes in various perspectives, viewpoints, or lighting conditions must be the same; second, the points must have enough information to match with one another. The finest characteristics for matching are corners. The most crucial aspect of a corner is that if one exists in a picture, the surrounding area will abruptly alter in intensity.

The information of pixels which is present locally are explained using local feature descriptors. These local features which are need to be evaluated meet various criteria like blurriness, presence of noise, translation invariant, rotation, scale and transformation based on affine. An effective feature detection operator that has seen widespread application is the Harris corner detector and Brisk corner detection. The rotation invariant Harris corner detector has sufficient data for feature matching.

Due to the Harris corner detector’s great invariance to rotation, scale, illumination fluctuation, and noise in image it is a well-liked interest point detector. The local autocorrelation function of a signal serves as the foundation for the Harris corner detector, which monitors local variations in the signal with patches that have been slightly displaced in various directions. The Harris approach looks at the intensity which is average and to be directional, to locate the corners in the input picture. The approach of detecting corners mathematical formulation essentially determines the intensity difference in every direction using a displacement of (u, v).

For the pixel with displacement (u,v) the grey intensity is termed as $I(x,y)$. Here the variation of the pixel

Table 4.1: Requirement of Design Environment

Description	Requirement
RAM	8GB
Processor	Intel i7
Matlab version	2021a
Image format	JPEG

that is gray is (x,y) with a shift range of (u,v) is given by equation (3.1).

$$H(u, v) = \sum_{x,y} w_f(x, y) [I(x + u, y + v) - I(x, y)]^2 \quad (3.1)$$

The term $w_f(x, y)$ denotes the windowing function, the shifted intensity value which is termed as $I(x + u, y + v)$ and the intensity value is termed as $I(x, y)$.

Key point descriptor with scale Scale-spacing and binary description are both handled by the BRISK approach [17]. In the picture pyramid's octave layers, key points are found. Quadratic function fitting is used to translate the coordinates and scale of each key point into representation of a continuous domain. The BRISK descriptor is generated as a binary string in two steps after the BRISK characteristics have been detected. The first stage aids in the creation of a rotation-invariant description by estimating the key points' orientation. To effectively and quickly construct a description that captures regional attributes, the second stage utilises rigorous brightness comparisons. The BRISK descriptor uses a concentric circle sampling method to specify N locations.

The smoothing of intensity at every point s_{pi} is done performed using a gaussian function for preventing of effects like aliasing. The sample points with N number are paired into (s_{pi}, s_{pj}) and are bifurcated into two degrees of classes: one is short pair in which the distance condition should be $(s_{pi}, s_{pj}) < T_{max}$ and the other one is long pair with a distance condition of $(s_{pi}, s_{pj}) > T_{min}$. These two pairs perform individual action like estimation of rotation using the short pair and building of descriptor after correction of rotation is performed using the long pair. The computation of local gradients of BRISK descriptor is given by

$$\nabla(s_{pi}, s_{pj}) = (s_{pj} - s_{pi}) \frac{I(s_{pj} - \sigma_j) - I(s_{pi}, \sigma_i)}{\|s_{pj} - s_{pi}\|^2} \quad (3.2)$$

The local gradient is termed as $\nabla(s_{pi}, s_{pj})$ which is the sampled pair and the intensity that is smoothed at x at scaling factor σ . In the average gradients of x and y direction the rotation angle θ is calculated. To get the descriptor that is rotation-invariant, the short pairs are rotated by an angle of $-\theta$. The binary descriptor, which serves as a description for each keypoint, is an encoded binary string.

Algorithm

- Step1. Loading the image from the dataset
- Step2. Resize all the images and reduce noise
- Step3. Removing unused pixels from an image in order to recover the skeletons.
- Step4. Using DCNN for extraction of features
- Step5. Divide the letters for the given input word
- Step6. Evaluate the BRISK features and HARRIS corner points
- Step7. Calculate all the parameters like hamming distance between the pixels of the word, MSE, PSNR etc..

4. Experimental Results. The analysis of suggested model is detailed with the help of matlab simulation. The entire simulation is performed using the image of hand written telugu words. The results shown below gives the effectiveness of the proposed model. The consideration for performing simulation is shown in table 4.1 and some of the assumption in simulation environment is shown in table 4.2.

Table 4.2: Assumption in Simulation Environment

Assumption	Description
Stability in performance	The evaluation does not undergo any adverse changes that effect the experimental findings
Range of Consistency	The environment of simulation remains consistency and do the parameters
Constant parameters	The parameters used for evaluation are same for all set of input images
No Interference from external source	As there is not external interference the simulation results will be accurate



Fig. 4.1: Input Image



Fig. 4.2: Thinning Image

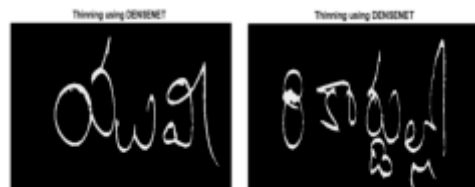


Fig. 4.3: DenseNet based thinning

The input is normalised and sent for next stage of processing that is thinning can be considered from Figure 4.1. The thinning image which is obtained using different techniques is shown in Fig. 4.2 and Fig. 4.3. In this process the front end of the image is highly viewed for achieving better set of features in next stage.

Every letter in the word is been segmented using OCR word segmentation model and the results achieved is shown in fig 4.4.

Harris features and Brisk features are been extracted and is shown in fig 4.5. The features are extracted for the image which is thinned using DenseNet CNN. The duration of telugu data retrieval is less when compared to other extraction of thinning image. The performance will be improved when the features are extracted for DCNN thinned image.

Certain parameters are considered for showing the performance of the suggested model with other existing techniques. The parameters are measure of connectivity, MSE, PSNR, Rate of thinning, RMSE, time of execution, Noise sensitivity and hamming distance. These parameter values are shown in table 4.3.

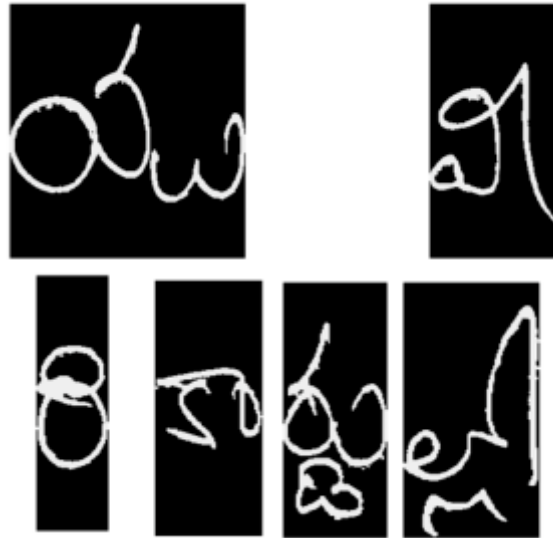


Fig. 4.4: Segmentation of DCNN thinned image

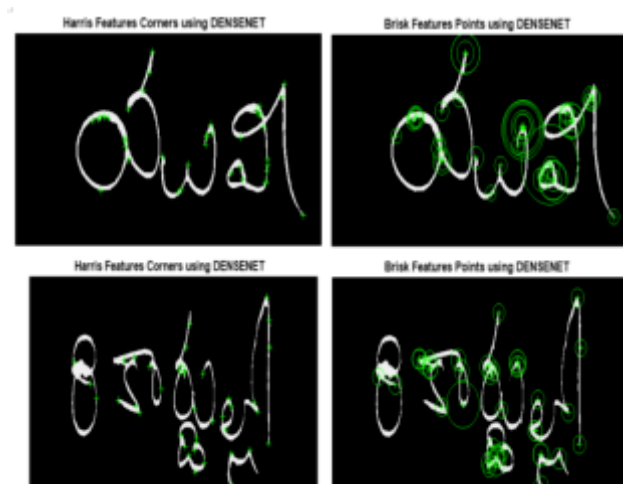


Fig. 4.5: Outputs of feature extraction

One of the crucial metrics for assessing the effectiveness of the employed strategies is PSNR. The PSNR, which affects pixel quality, is the ratio of the highest pixel value to the noise (MSE). The error's value, which is expressed on a logarithmic decibel scale, decreases with increasing PSNR. Figure 4.6 shows the contrast in PSNR.

5. Conclusion. In this study, an efficient model for Telugu word recognition is presented through the testing of several neural network designs. The provided input is first normalised and thinned before being submitted to the deep convolutional neural network model to extract the feature maps. Using DenseNet convolutional neural networks, a model for Telugu text extraction and identification is built in this article. Additional corner characteristics are retrieved using the Harris and Brisk techniques. The simulation research

Table 4.3: Performance measures using different techniques

Measure Evaluted	Hilditch Algorithm	Morphological operations	RNCNN-BRHA [18]	Proposed HWTR-DCNN
Measure of Connectivity	4.0	4.0	4.0	4.01
Rate of Thinning (pixels)	1.0	1.0	1.0	1.0
MSE	0.022	0.025	0.0007	0.00018
PSNR	16.49	15.95	51.84	54.74
RMSE	0.149	0.159	0.0025	0.00135
Time for executing (Sec)	2.19	0.608	0.102	0.094
Noise Sensitivity	0.63	0.388	0.50	1
Hamming Distance (pixels)	7.15	6.20	2.71	1.20

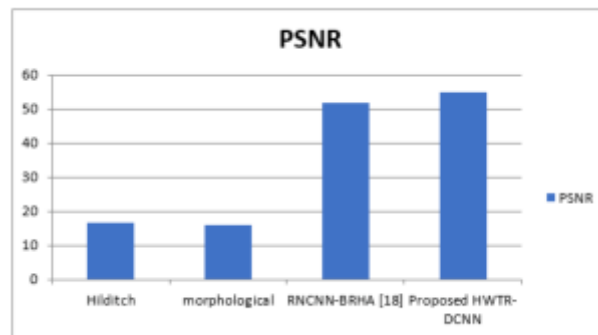


Fig. 4.6: PSNR Comparison

showed that in terms of PSNR, MSE, Noise sensitivity, and execution time, the new technique outperformed the traditional retrieval system. Additionally, the suggested HWTR-DCNN system's performance evaluation is shown using mAP and mAR and is contrasted with the current systems.

REFERENCES

- [1] Li, Ang, et al. "Generating holistic 3d scene abstractions for text-based image retrieval." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.
- [2] Unar, Salahuddin, et al. "Detected text-based image retrieval approach for textual images." IET Image Processing 13.3 (2019): 515-521.
- [3] MK, Yanti Idaya Aspura, and Shahrul Azman Mohd Noah. "Semantic text-based image retrieval with multi-modality ontology and DBpedia." The Electronic Library (2017).
- [4] Zeng, Mengqi, et al. "CATIRI: An efficient method for content-and-text based image retrieval." Journal of Computer Science and Technology 34.2 (2019): 287-304.
- [5] Estrela, Vania Vieira, and Albany E. Herrmann. "Content-based image retrieval (CBIR) in remote clinical diagnosis and healthcare." Encyclopedia of E-Health and Telemedicine. IGI Global, 2016. 495-520.
- [6] Harmandeep Kaur, and Munish Kumar, "A Comprehensive Survey on Word Recognition for Non-Indic And Indic Scripts," Pattern Anal Applic, vol. 21, pp. 897-929, 2018. Crossref, <https://doi.org/10.1007/s10044-018-0731-2>
- [7] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", International Journal of Computer Vision, Vol. 60, No. 2, pp. 91-110, 2004.
- [8] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories", In: Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, 2006.
- [9] T. M. Rath and R. Manmatha, "Word spotting for historical documents", International Journal of Document Analysis and Research, Vol. 9, No. 2-4, pp.139-152, 2007.
- [10] K. Takeda, K. Kise, and M. Iwamura, "Real-time document image retrieval for a 10 Million pages database with a memory efficient and stability improved LLAH", In: Proc. of the International Conf. on Document Analysis and Recognition, pp. 1054-1058, 2011.

- [11] R. Shekhar and C. V. Jawahar, "Word Image Retrieval Using Bag of Visual Words", In: Proc. of the International Workshop on Document Analysis Systems, pp. 297-301, 2012.
- [12] J. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders, "Kernel Codebooks for Scene Categorization", In: Proc. of European Conf. on Computer Vision, pp. 696-709, 2008.
- [13] D. Nagasudha and Y. M. Latha, "Keyword Spotting using HMM in Printed Telugu Documents", In: Proc. of International Conf. on Signal Processing, Communication, Power and Embedded Systems, pp. 1997-2000, 2016.
- [14] Zhenyu Liu, Haiwei Huang, Chaohong Lu, and Shengfei Lyu. 2020. Multichannel cnn with attention for text classification. ArXiv, abs/2006.16174.
- [15] Rajasekhar Boddu, Edara Sreenivasa Reddy (2023). Novel Heuristic Recurrent Neural Network Framework to Handle Automatic Telugu Text Categorization from Handwritten Text Image. International journal of recent and innovation trends in computing and communication, vol.11, No.4, pp.296-305.
- [16] <http://cvit.iiit.ac.in/research/projects/cvit-projects/indic-hw-data>
- [17] Leutenegger S, Chli. M, Siegwart RY (2011) Brisk: Binary robust invariant scalable keypoints. In: Proceedings of the 2011 International Conference on Computer Vision, IEEE Computer Society, Washington, DC, USA, ICCV '11, pp 2548–2555.
- [18] Boddu, R., Reddy, E.S. (2023). Fusion of RNCNN-BRHA for recognition of telugu word from handwritten text. Revue d'Intelligence Artificielle, Vol. 37, No. 1, pp. 215-221.

Edited by: Anil Kumar Budati

Special issue on: Soft Computing and Artificial Intelligence for wire/wireless Human-Machine Interface

Received: Jan 12, 2024

Accepted: Mar 20, 2024



UAV PATH PLANNING MODEL LEVERAGING MACHINE LEARNING AND SWARM INTELLIGENCE FOR SMART AGRICULTURE

ROBERTO E. ROQUE-CLAROS*, DEIVI P. FLORES-LLANOS†, ABEL R. MAQUERA-HUMPIRI‡, VIJAYA KRISHNA SONTHI§, SUDHAKAR SENGAN¶, AND RAJASEKAR RANGASAMY||

Abstract. Smart agriculture, through precision farming, is revolutionizing traditional farming methods by optimizing resource use and enhancing yields. With the integration of technology, especially the advent of Unmanned Aerial Vehicles (UAVs) or drones, modern agriculture has attained new heights in efficient crop management, real-time data collection, and sustainable practices. UAVs play a pivotal role, offering aerial insights into crop health, soil conditions, and targeted resource application, promoting sustainable farming. However, navigating UAVs efficiently across dynamic agricultural terrains presents challenges, particularly in path planning. While traditional grid-based models have their merits, the complexities of modern farms demand more adaptive models. This work introduces a hierarchical path planning framework for UAVs, combining the “Enhanced Genetic Algorithm using Fuzzy Logic” for global planning and the “Improved D* Algorithm” for real-time local adjustments. This dual-layered approach ensures efficient, safe, and energy-conserving UAV trajectories, marking a significant advancement in UAV-based smart agriculture.

Key words: Unmanned Aerial Vehicles, Precision Farming, Grid-Based Models, D* Algorithm, Smart Agriculture

1. Introduction. Smart agriculture has reshaped the way we perceive and practice traditional farming. Through the lens of precision farming, the nuances of crop management are being meticulously addressed, leveraging data-driven insights to optimize resource utilization and maximize yields. The influence of autonomous vehicles in this domain underscores the potential of technology to enhance and streamline agricultural operations, bringing about a seamless integration of mechanization and intelligence [11]. This synergy promises to lead the farm sector toward unprecedented efficiency and sustainability. Unmanned Aerial Vehicles (UAVs), commonly known as drones, have emerged as revolutionary tools in modern agriculture. Their ability to swiftly traverse vast expanses of land, capturing high-resolution imagery and providing real-time data, has redefined precision farming. UAVs can efficiently monitor crop health, assess soil moisture levels, and detect pest infestations, all from an aerial vantage point. This bird’s-eye view enables farmers to make informed decisions, leading to reduced input costs and increased crop yields. Furthermore, UAVs facilitate targeted applications of pesticides and fertilizers, ensuring that resources are used judiciously, minimizing environmental impact [12]. The integration of UAV technology in agriculture optimizes farm management practices and paves the way for sustainable and environmentally-conscious farming, marking a significant stride toward the future of agriculture.

In the growing field of smart agriculture, integrating UAVs offers immense potential but is not without challenges. One of the paramount issues is the intricacy of path planning for UAVs. Given agricultural landscapes’ diverse and dynamic nature, plotting an efficient and safe drone route necessitates advanced algorithms and real-time data processing [7]. Factors like varying crop heights, obstacles like trees or infrastructures, and changing weather conditions can significantly influence the UAV’s trajectory. While the objective is to cover

*Universidad Nacional del Altiplano de Puno, P.O. Box 291, Puno - Perú. (reroque@unap.edu.pe).

†Universidad Nacional del Altiplano de Puno, P.O. Box 291, Puno - Perú. (dflores@unap.edu.pe).

‡Universidad Nacional de Moquegua - Perú. (20230101120@unam.edu.pe).

§Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram 522502, Andhra Pradesh, India. (vijayakrishna1990@gmail.com).

¶Department of Computer Science and Engineering, PSN College of Engineering and Technology, Tirunelveli, Tamil Nadu, India. (Corresponding Author, sudhasengan@gmail.com).

||Department of Computer science and Engineering, GITAM School of Technology, GITAM University, Bengaluru Campus, India. (rrangasa@gitam.edu).

maximum ground efficiently to gather data, avoiding collisions and ensuring the UAV's energy conservation becomes equally vital. Therefore, a practical path planning mechanism is imperative not only for the operational success of UAVs in smart agriculture but also to ensure the safety and sustainability of their application in such a crucial sector.

In addressing the path planning challenge for UAVs in smart agriculture, several existing models have been proposed and explored. Traditional approaches have primarily revolved around grid-based methods, where the agricultural field is divided into uniform cells, and the UAV's path is determined by traversing these cells based on predefined algorithms. A* and Dijkstra are classic examples renowned for their efficiency in obstacle-free environments. However, the need for adaptive and dynamic models grew as agricultural terrains became more complex. Genetic Algorithms (GAs) and Particle Swarm Optimization (PSO) have been introduced as heuristic methods to navigate intricate landscapes [8]. These algorithms simulate natural processes and behaviors to find optimal or near-optimal paths, making them more resilient to dynamic environmental changes. More recently, machine learning techniques, particularly deep learning, have been incorporated to predict and adjust UAV paths in real-time, leveraging vast datasets from past flights. While these models have highlighted promising results, a comprehensive solution that seamlessly integrates responsiveness, accuracy, and efficiency remains a subject of ongoing research.

The challenge of UAV path planning in smart agriculture demands a model that is accurate and adaptive to the varying nuances of an agricultural landscape. To tackle this, this work introduces a novel, hierarchical framework. Firstly, the farm terrain is meticulously represented through a grid environment. This grid is formed by discretizing the field into a two-dimensional lattice, where distinct cells denote either navigable or obstructed zones. To enhance clarity and reduce computational overhead, morphological operations refine this grid, highlighting only essential path-planning elements. This dynamic grid adjusts to changing agricultural conditions, ensuring the UAV's viability throughout different agricultural phases. Following the grid formation, our model sequentially integrates two sophisticated algorithms to provide optimal UAV navigation.

The first phase involves "Global path planning", utilizing the "Enhanced Genetic Algorithm using Fuzzy Logic for Global Path Planning". This algorithm evaluates various pathways throughout the grid, identifying the most efficient route from the beginning point to the target by selecting the paths with the highest fitness values. This ensures that the UAV is provided with an efficient and energy-conserving trajectory. Following the global path planning, critical path nodes are extracted, marking significant waypoints or transitions in the path. Subsequently, the "Improved D* Algorithm for Local Path Planning" is applied in the second phase. This algorithm focuses on the UAV's immediate surroundings, adjusting its real-time trajectory based on detected obstacles or unforeseen environmental changes. By doing so, the UAV can adapt quickly to ensure safe and effective local maneuvering. The culmination of these two processes yields "The optimal trajectory", guiding the UAV seamlessly from its initial point to its destination. Once this trajectory is successfully followed, the model confirms Path finding success.

The article is framed as follows: Section 2 presents the literature review, Section 3 presents the proposed model, Section 5 presents the experimental analysis, and Section 6 presents the conclusion of the work.

2. Literature Review. The arena of UAV path planning has seen extensive research and development in recent years, focusing primarily on optimization, collision avoidance, and adaptability to the varying complexities of the UAVs' environment. Aggarwal *et al.*, comprehensively analyzed various path-planning techniques used for UAVs over the years [1]. They broadly classified these techniques into representative, cooperative, and non-cooperative, emphasizing the path's optimality, shortness, and collision-free nature. An essential contribution of their work is the exhaustive comparative tables and identification of open research problems in UAV path planning, emphasizing factors such as energy efficiency, time efficiency, and robustness.ents.

Bai *et al.*, proposed a path-planning algorithm harmoniously integrated with the A and DWA algorithms [2]. Their approach gave prominence to global path optimization while considering UAVs' security and speed requirements. By preprocessing the map for obstacles, they addressed inherent limitations of the standard algorithms, ensuring the UAV path is efficient and safe. Delving into the potential of reinforcement learning in UAV path planning, Tu *et al.*, highlighted its application in aquaculture cage detection [13]. They employed the Q-learning algorithm, comparing it with the SARSA algorithm. Their use case underscored the importance of energy conservation and efficiency, given the vast expanse of the sea and the scattered nature of net cages.

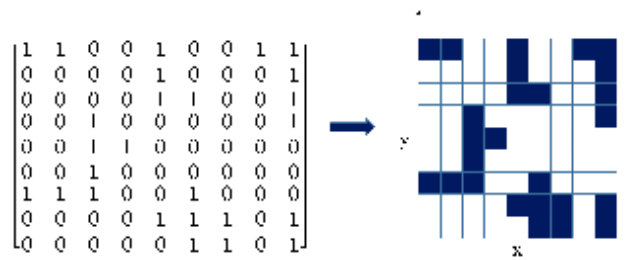


Fig. 3.1: Grid Representation

Chen *et al.*, addressed challenges in agricultural irrigation by introducing an intelligent irrigation robot [3]. Their work is pivotal for its emphasis on precision in irrigation using an improved path planning algorithm leveraging Bayesian theory. They aimed for full irrigation coverage in the complex agricultural environment, ensuring no area was inspected. Li *et al.*, brought forward an innovative integration of the improved artificial fish swarm algorithm with Bézier curves for mobile robot path planning and smoothing [6]. Their method promises enhanced planning accuracy and path continuity, meeting the kinematic demands of mobile robots.

Highlighting the potential of reinforcement learning in multi-layered path planning, Cui *et al.*, introduced a unique algorithm that assimilated local and global information for superior performance [4]. Their approach utilized B-spline curves for real-time path smoothing, proving its efficacy through various simulations. Qu *et al.*, proposed a hybrid algorithm, HSGWO-MSOS, by combining the strengths of the simplified grey wolf optimizer and the modified symbiotic organism's search [9]. Their algorithm emphasized efficiency in exploration and exploitation, offering an enhanced route for UAVs that is feasible and effective. Yan *et al.*, ventured into Deep Reinforcement Learning (DRL) for UAV path planning in dynamic and potentially threatening environments [14]. Their model simulated the UAV's survival probability against threats like missile attacks, using the D3QN algorithm for improved performance. Shao *et al.*, tackled the issue of autonomous UAV formation system path planning, proposing a comprehensively enhanced particle swarm optimization technique [10]. Their methodology emphasized rapidity and solution optimality, addressing terrain and threat constraints. Lastly, Han *et al.*, concentrated on UAV indoor path planning in complex environments [5]. They introduced a set of grid-optimized algorithms that considerably reduced computational complexity, efficiently tackled dead zone airspaces, and assured efficient and flyable path planning in intricate 3D indoor airspace.

3. Proposed Model.

3.1. Task Model. In configuring the computational representation of the agricultural terrain for UAV path planning, a meticulous grid structuring process is undertaken. This grid acts as a virtual model, aiding the UAV in discerning navigable paths from obstructed zones within the agricultural landscape. The field is discretized into a two-dimensional grid, G , where each cell, C_{ij} , corresponds to a specific area in the farmland. Cells that represent obstacles are assigned a value of 0 , indicating areas that are off-limits for the UAV. In contrast, cells expressing free space are assigned a value of 1 , delineating safe flight zones (Figure 3.1). The binary values create a stark contrast on the grid, forming a map of passable and impassable regions for the UAV.

To ensure the UAV path planning model does not become overburdened by environmental intricacies, a combination of morphological operations, specifically dilation and erosion, is applied. These operations facilitate feature extraction, resulting in a refined grid, G' , which highlights critical path planning information while negating redundant details. For the UAV to accurately locate itself within the grid, each cell is indexed with a unique coordinate pair, (x_i, y_j) . The relationship between a cell's linear index, k , and its two-dimensional coordinate pair is governed by EUQ (3.1) and EQU (3.2)

$$x_i = \left\lfloor \frac{k-1}{N_x} \right\rfloor + 1 \quad (3.1)$$

$$y_j = ((k - 1) \bmod N_y) + 1 \quad (3.2)$$

where N_x and N_y represent the total number of rows and columns in the grid, respectively. The functions $[\cdot]$ and \bmod denote the floor operation and modulus operation, instrumental in mapping the linear index to the grid coordinates. The grid environment, G' , is designed to be dynamic, accommodating changes in agricultural conditions, such as seasonal crop growth or temporary obstructions like farming equipment. This dynamic aspect ensures that the path planning model remains viable throughout varying stages of the agricultural lifecycle.

3.2. Objective Function Formulation. In smart agriculture, the path-planning model aims to blend global and local methodologies to derive an optimal trajectory for UAVs. The quantification of this trajectory's optimality is captured in the objective function, F_{obj} .

Formally, this function is given by EQU (3.3)

$$F_{obj}(P) = w_1 \cdot L_{\text{global}}(P) + w_2 \cdot E_{\text{global}}(P) + w_3 \cdot T_{\text{local}}(P) - w_4 \cdot C_{\text{local}}(P) \quad (3.3)$$

where:

- P represents the UAV's path.
- $L_{\text{global}}(P)$ signifies the total length from the global path planning perspective.
- $E_{\text{global}}(P)$ corresponds to the energy consumed during the global path traversal.
- $T_{\text{local}}(P)$ denotes the time taken to focus on finer, local path intricacies.
- $C_{\text{local}}(P)$ captures the coverage of specific areas of interest from a local planning standpoint.
- w_1, w_2, w_3 , and w_4 are weighting coefficients reflecting the relative significance of each component to the holistic path planning objectives.

The optimal path, while considering both global and local aspects, should satisfy the following constraints:

- *Safety Constraints:* The UAV's trajectory must bypass obstacles, notably the 0value cells in the grid environment.
- *Flight Dynamics Constraints:* Considering the UAV's inherent physical capabilities, the path is limited by its turning radius and maximum flight speed.
- *Coverage Constraints:* Ensuring complete and detailed coverage of agricultural areas is pivotal, especially zones marked for close monitoring. The route must minimize overlaps and redundancies.

3.3. Optimization Strategy. For the global path planning phase, a fuzzy-based Genetic Algorithm (GA) is employed to navigate the broader aspects of the agricultural terrain. The GA's intrinsic evolutionary process, when enhanced with fuzzy logic, provides a robust mechanism to discern optimal paths by considering the global dynamics of the agricultural landscape. In the local path planning phase, Swarm Intelligence is utilized. Swarm Intelligence, inspired by the collective behavior of decentralized systems, excels in refining trajectories. It accounts for intricate details and unexpected hindrances in the agricultural setting, ensuring the UAV can navigate tighter spaces and rapidly adjust its trajectory when faced with unforeseen challenges. Formally, the proposed model optimizes the objective function F_{obj} using these methods, targeting the following goal:

$$P^* = \arg \min_P F_{obj}(P) \quad (3.4)$$

With this optimization strategy, the hierarchical UAV path planning model seeks to determine the best possible trajectory. This trajectory balances the broad strokes of global path planning with the finer nuances of local planning, ensuring the UAV meets the demands of smart agriculture. The emphasis is firmly on precision, efficiency, and adaptability.

3.4. Proposed Enhanced Genetic Algorithm Using Fuzzy Logic for Global Path Planning. In UAV-based smart agriculture, achieving comprehensive field coverage while minimizing energy consumption and traversal time is paramount. The Genetic Algorithm (GA) has been a popular method for this task due to its inherent ability to search vast solution spaces efficiently. However, to better address an agricultural field's dynamic and often uncertain environment, integrating fuzzy logic into GA can offer significant advantages.

3.4.1. Representation and Initial Population. For the given agricultural grid, denoted as G , of dimensions $M \times N$, each chromosome in the GA represents a potential path P that the UAV can follow. This path starts at a designated point S and ends at a predetermined destination D . Utilizing the principle of random walks, we initialize our population of chromosomes. In this approach, each UAV path is generated by letting it ‘walk’ randomly across the grid from the start point S to the destination D , ensuring it stays within the boundaries and constraints of the grid. Through this stochastic method, many diverse pathways are conceived, providing a broad spectrum of starting solutions for the GA to refine and optimize. These random walks, while unguided, produce routes that capture the vast complexities of the agricultural landscape, laying a strong foundation for the subsequent genetic algorithm optimization.

3.4.2. Fuzzy-based Fitness Evaluation. The fitness of each path is determined by several factors: path length $L(P)$, energy consumed $E(P)$, and the agricultural area covered $A(P)$. Instead of rigid thresholds, fuzzy logic can handle the imprecision in measurements. This includes the unpredictability of dynamic obstacles that may suddenly block the path, uncertain wind conditions affecting energy consumption, and varying crop heights impacting the coverage assessment. The fuzzy-enhanced fitness function may be expressed as (3.5)

$$F_{\text{fitness}}(P) = w_1 \times \mu_{\text{length}}(L(P)) + w_2 \times \mu_{\text{energy}}(E(P)) + w_3 \times \mu_{\text{area}}(A(P)) \tag{3.5}$$

Here, μ denotes the membership function in fuzzy logic, defining the degree of truth of each component.

3.4.3. Fuzzy-enhanced Selection. Given a set of chromosomes $C = \{c_1, c_2, \dots, c_n\}$, each chromosome c_i has an associated fitness value $F_{\text{fitness}}(c_i)$. In the fuzzy-enhanced GA, a fuzzy membership function $\mu_{\text{suitability}}(c_i)$ evaluates the suitability of chromosome c_i for selection. The uncertain factor, such as obstacle, is represented as ω . If the challenges are uncertain, the adjusted fitness can be defined as EQU (3.6)

$$F'_{\text{fitness}}(c_i) = F_{\text{fitness}}(c_i) + \alpha \times \mu_{\text{wind}}(\omega) \tag{3.6}$$

where α is a factor determining the effect of obstacle uncertainty.

3.4.4. Fuzzy-enhanced Selection. For two parent chromosomes c_p and c_q , the compatibility for a crossover at a gene g_i is given by $\mu_{\text{compatibility}}(c_p[g_i], c_q[g_i])$.

The crossover point(s) X is determined as EQU (3.7)

$$X = \arg \max_i \mu_{\text{compatibility}}(c_p[g_i], c_q[g_i]) \tag{3.7}$$

This ensures that genes around point X have the highest compatibility between parents.

3.4.5. Fuzzy-enhanced Selection. For a chromosome c_i , the mutation likelihood for a gene g_j is determined by $\mu_{\text{effectiveness}}(c_i[g_j])$. If the mutation threshold is θ , a gene g_j is mutated if: $\mu_{\text{effectiveness}}(c_i[g_j]) > \theta$. This ensures targeted mutations based on the effectiveness of individual genes in the chromosome.

3.4.6. Optimization and Convergence. Optimization and Convergence: Let the optimal path after k generations be P^* and the fuzzy-adjusted path quality be $Q(P)$ considering the specific challenges and requirements of smart agriculture. As the generations progress, the optimal path is refined as EQU (3.8)

$$P^{k+1*} = \arg \max_{P \in C^{k+1}} Q(P) \tag{3.8}$$

The algorithm converges when the quality difference between consecutive generations is below a predefined threshold, EQU (3.9)

$$\left| Q(P^{k*}) - Q(P^{k+1*}) \right| < \epsilon \tag{3.9}$$

The following algorithm presents the steps in the proposed enhanced GA algorithm.

Algorithm 1: Enhanced Genetic Algorithm (EGA) using Fuzzy Logic**Inputs:**

- Agricultural grid G of dimensions $M \times N$.
- Start point S .
- Destination point D .
- Population size: PopSize.
- Maximum number of generations: MaxGen,
- Crossover probability: P_c .
- Mutation probability: P_m .
- Convergence threshold: ε .

Output: Optimal path P^* .**Algorithm:****1. Initialization**

- (a) Set generation to 0.
- (b) Initialize an empty set Population.
- (c) For i from 1 to PopSize:
 - i. Generate a path P_i using random walks from S to D within grid G .
 - ii. Add P_i to Population.

2. Evaluation

- (a) For each path P_i in Population:
 - i. Compute the fitness value:

$$F_{\text{fitness}}(P_i) = w_1 \times \mu_{\text{length}}(L(P_i)) + w_2 \times \mu_{\text{energy}}(E(P_i)) + w_3 \times \mu_{\text{area}}(A(P_i))$$

3. Selection

- (a) For each chromosome c_i in Population:
 - i. Evaluate the suitability using $\mu_{\text{suitability}}(c_i)$.
 - ii. Adjust fitness considering uncertain obstacles:

$$F'_{\text{fitness}}(c_i) = F_{\text{fitness}}(c_i) + \alpha \times \mu_{\text{obstacle}}(\omega)$$

4. Crossover

- (a) For two parent chromosomes c_p and c_q :
 - i. Determine compatibility for crossover at a gene g_i by $\mu_{\text{compatibility}}(c_p[g_i], c_q[g_i])$.
 - ii. Choose crossover points X maximizing compatibility.

5. Mutation

- (a) For Each chromosome c_i :
 - i. Determine mutation likelihood for a gene g_j by $\mu_{\text{effectiveness}}(c_i[g_j])$.
 - ii. If $\mu_{\text{effectiveness}}(c_i[g_j]) > \theta$, mutate gene g_j .

6. Optimization and Convergence

- (a) Determine the optimal path after k generations as P_k^* and the fuzzy-adjusted path quality as $Q(P)$.
- (b) Refine the optimal path:

$$P_{k+1}^* = \arg \max_{P \in \text{Population}_{k+1}} Q(P)$$

- (c) Check for convergence: If $|Q(P_k^*) - Q(P_{k+1}^*)| < \varepsilon$, end the algorithm.

4. Improved D* Algorithm for Local Path Planning. In the context of UAV-based smart agriculture, while global path planning designs an optimal route for comprehensive field coverage, local path planning adapts to dynamic obstacles and sudden environmental changes. The traditional D* algorithm has been effective for this purpose, but we propose an enhanced approach incorporating refined UAV safety distance determination.

- **Enhanced Cost Function:** While the basic D relies on a static cost between nodes, our improved model integrates dynamic costs influenced by numerous factors:
- **Energy Cost:** Borrowing from our global path planning model, the energy E required to traverse a path segment is included in the cost function.
- **Safety Cost:** Given that a UAV requires a safe distance from obstacles, especially in unpredictable agricultural terrains, a cost component S that penalizes paths too close to detected obstacles is introduced. This is determined as EQU (4.1)

$$S = k \times e^{-d} \quad (4.1)$$

where k is a constant, d is the distance from the obstacle, and e is the base of the natural logarithm. This ensures a higher cost for paths closer to blocks and vice versa. The final cost C for an edge can be represented as EQU (4.2)

$$C = w_1 \times L + w_2 \times E + w_3 \times S \quad (4.2)$$

where w_1 , w_2 , and w_3 are weighting factors, and L denotes path length.

- **UAV Safety Distance Determination:** The safety distance determination ensures that the UAV maintains a safe distance from obstacles (Figure 4.1). This involves geometric calculations to determine the distance between the UAV's path and nearby obstructions.

Given:

- Coordinates of a current node P are (x_p, y_p) .
- Coordinates of the destination node Q in its path are (x_q, y_q) .
- Coordinates of an obstacle node R are (x_r, y_r) .

The projection of obstacle R onto the path segment PQ is denoted by R' . The y-coordinate, $y_{r'}$ of this projection can be determined as EQU (4.3)

$$y_{r'} = \frac{(y_q - y_p)}{x_q - x_p} (x_r - x_p) + y_p \quad (4.3)$$

The angle θ between the path segment PQ and the x-axis can be found as EQU (4.4)

$$\theta = \arctan \left(\frac{y_p - y_q}{x_p - x_q} \right) \quad (4.4)$$

Now, the distance s from the obstacle R to its projection R' on the path, PQ is EQU (4.5)

$$s = |y_r - y_{r'}| \quad (4.5)$$

- **Safety Threshold and Path Consideration:** A predefined safety threshold, T , is essential. If s is less than T , the path might be too close to the obstacle and should be reconsidered. However, considering environmental factors and UAV specifications, a dynamic adjustment factor, Δ , is introduced. This adjustment factor modifies the safety threshold based on real-time data, EQU (4.6)

$$T' = T + \Delta \quad (4.6)$$

Now, if $s < T'$, the UAV should reconsider the path. Otherwise, it can continue on the path PQ . The above equation, Δ , represents the dynamic adjustment factor influenced by UAV speed, wind conditions, and obstacle mobility. For instance, the safety threshold should be increased if an obstacle moves.

- **Path Smoothing Optimization:** One of the challenges with any path planning algorithm is that the generated path might not always be smooth. Sharp turns or zigzag patterns may be introduced in the path, especially when navigating through dense obstacle environments. Such paths are inefficient

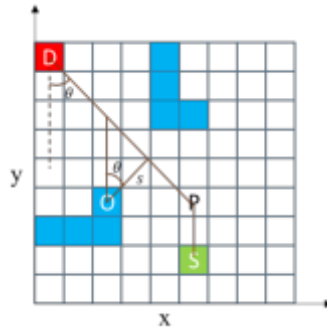


Fig. 4.1: Determining Safety Distance

for a UAV, leading to rapid battery drain, instability, and reduced safety. Thus, path smoothing optimization ensures that the UAV’s path is as streamlined as possible, reducing unnecessary movements and providing a more efficient trajectory.

Bezier curves are a mathematical tool in computer graphics designed to generate smooth curves. By employing Bezier curves, we can transform a series of linear segments into a smooth curve that preserves the original waypoints and reduces inflections. Given two control points A and B and two endpoints, P_0 and P_1 , the Bezier curve $B(t)$ is defined as EQU (4.7)

$$B(t) = (1 - t)^3 P_0 + 3(1 - t)^2 t A + 3(1 - t) t^2 B + t^3 P_1 \tag{4.7}$$

where $0 \leq t \leq 1$. A cost function can be defined to evaluate the smoothness of a path. The cost is higher for paths with sharp turns or abrupt changes in direction. Given a path segment s , the cost $C(s)$ could be related to the derivative of the path concerning distance, squared, EQU (4.8)

$$C(s) = \int_{\text{path}} \left(\frac{d^2 s}{dt^2} \right)^2 dt \tag{4.8}$$

The goal is to minimize $C(s)$ to achieve the smoothest path. Often, a single pass of optimization might not yield the best results. Iterative refinement involves running the smoothing algorithm multiple times, tweaking parameters, and adjusting waypoints as needed until a desired level of smoothness and efficiency is achieved.

- **Collision Check after Smoothing:** Path smoothing optimizes UAV trajectories, eliminating abrupt changes and ensuring energy-efficient movement. However, as trajectories are modified, the risk of infringing upon safety margins around obstacles may inadvertently increase. Thus, postsmoothing collision checks are indispensable. A UAV operating in a cluttered environment may encounter multiple obstacles, static (e.g., infrastructure) and dynamic (e.g., other UAVs). As the UAV’s path undergoes smoothing, ensuring that it remains collision-free at every point becomes paramount. To this end, we introduce a distance function, D . For any point p on the UAV’s trajectory, D calculates the shortest distance to the closest obstacle, facilitating instantaneous hazard proximity assessment. Let O denote the set of all obstacles in the environment, and $d(p, o)$ represent the Euclidean distance between point p and obstacle o . Then, the distance function $D(p)$ is articulated as, EQU (4.9)

$$D(p) = \text{Min}_{o \in O} d(p, o) \tag{4.9}$$

To ensure safety, for all points p on the smoothed path: $D(p) > R_{\text{safe}}$; where R_{safe} is not merely a predefined radius. Instead, it is dynamically computed based on the UAV’s relative positioning to nearby obstacles, considering both distance and angular considerations.

Given the UAV's position as p_{UAV} , and the angle $\alpha(p_{UAV}, o)$ between its heading direction and an obstacle o , we compute distances and angles to all obstacles, EQU (4.10) and EQU (4.11)

$$D(o) = d(p_{UAV}, o) \forall o \in O \quad (4.10)$$

$$A(o) = \alpha(p_{UAV}, o) \forall o \in O \quad (4.11)$$

This work introduces weighting factors, w_d and w_α , which dictate the significance of distance and angular considerations. R_{safe} is then defined as EQU (4.12)

$$R_{\text{safe}} = \frac{\sum_{o \in O} w_d \cdot D(o) + w_\alpha \cdot A(o)}{|O|} \quad (4.12)$$

This approach ensures that the UAV maintains an adaptive safety radius, considering its distance from and orientation to potential obstacles. Should any point on the smoothed path breach the condition mentioned above, the path is deemed unsafe and requires further refinement. This rigorous framework guarantees that the UAV's route is not only visually smooth but also technically secure for traversal. The following steps present the proposed local path algorithm.

Algorithm 2: Improved D* for Local Path Planning

Inputs:

- Start node S
- Goal node G
- Global Path Planning Model
- Weights w_1, w_2, w_3
- Safety factor k
- Safety threshold adjustment Δ

Output:

- The smoothed path from S to G or "UNSAFE" notification.

Steps:

1. **Initialization:**
 - (a) Current node $\leftarrow S$
 - (b) Initialize $\text{open_list} = \{\}$ and $\text{closed_list} = \{\}$
2. **Enhanced Cost Function:**
 - (a) Compute E , which represents the energy from the Global Path Planning Model.
 - (b) Calculate the safety cost as: $\text{Safety_Cost}(n) = k \times \exp(\text{distance to nearest obstacle}(n))$
 - (c) Determine the path length from node n to the goal node: $\text{Path_Length}(n) = \text{distance}(n, G)$
 - (d) Return $w_1 \times \text{Path_Length}(n) + w_2 \times EE + w_3 \times \text{Safety_Cost}(n)$
3. **UAV Safety Distance Determination**
 - (a) Calculate $\text{Safety_a_Distance}(P, Q, R)$:

$$R' = \text{projection of } R \text{ onto segment } PQ$$

$$s = \underline{\text{distance}(R \text{ to } R')}$$

- (b) Return s
4. **Safety Threshold and Path Consideration:**
 - (a) $\text{Check_Safety_Threshold}(s, T)$:

$$T' = T + \Delta$$

- (b) If $s < T'$ Then return "UNSAFE"
- (c) Else, return "SAFE".
5. **Path Smoothing Optimization:**

- (a) Bezier Smoothing(path):
 - i. For each segment s in the path: Apply Bezier curves using control points and endpoints If $C(s)$ (based on the derivative) is too high, refine segment s
 - ii. Return smoothed path
- 6. Collision Check after Smoothing:**
 - (a) Collision Check (path):
 - i. For each point p in the path: Calculate $D(p)$ as the shortest distance to the closest obstacle If $D(p) < R_{\text{safe}}$, return “UNSAFE”
 - ii. Return “SAFE”
- 7. Path Planning Procedure:**
 - (a) While current node is not goal node:
 - (b) Find neighbors of current node
 - (c) For each neighbor:
 - i. Calculate cost using Calculate Cost_Cost
 - ii. Calculate safety distance using Calculate Safettocos _Distance
 - iii. Check safety threshold using Check_Safety_Threshold
 - iv. If the node is safe and has a reasonable cost, add to opena list
 - (d) Move current node to closed list
 - (e) Set the node with the lowest cost in open list as current node
- 8. After Reaching Goal_Node:**
 - (a) Traceback path from goal node to start node
 - (b) Apply Bezier. Smoothing to the path
 - (c) Conduct Collision Check on the smoothed path
 - (d) If the path is “UNSAFE”, reiterate path planning or refine the smoothing
 - (e) Else, execute the path.

Having delved into the intricacies of both the Enhanced Genetic Algorithm using Fuzzy Logic for global path planning and the Improved D* Algorithm for local path planning, it is crucial to understand their harmonized implementation in UAV navigation. The Enhanced Genetic Algorithm using Fuzzy Logic, renowned for its adeptness in combining genetic algorithms with fuzzy logic principles, sketches an optimal path from the start point to the goal by evaluating multiple routes and selecting the best fitness value. This provides the UAV with a broad overview of its trajectory, ensuring efficient and energy-conservative navigation. However, the ever-changing nature of real-world scenarios requires a more adaptive approach to immediate obstacles and dynamic environments. This is where the Improved D* Algorithm comes into play. It operates in the UAV’s immediate surroundings, dynamically adjusting its real-time trajectory based on the sensed obstacles and environmental changes. By rapidly updating the robot’s path as the environment changes, the Improved D* Algorithm ensures safe and adaptive local maneuvering. When these two algorithms are sequentially integrated, the UAV benefits from the foresight of the Enhanced Genetic Algorithm using Fuzzy Logic for long-range planning while relying on the agility and responsiveness of the Improved D* Algorithm for short-range adjustments. The synergy of these algorithms equips the UAV with a comprehensive and adaptable navigation blueprint, ensuring safe and efficient flight.

5. Experimental Analysis. The principal simulation platform for our experimental analysis was configured around an Intel® Core™i7 – 9700 K CPU @ 3.60GHz with Windows 10 as the operating system. The algorithm’s performance was evaluated using MATLAB simulation tools. We designed a comprehensive path planning domain measuring 90 km by 90 km. The environment is partitioned into cells, each encompassing a 10 km by 10 km area. This results in a grid configuration of 9×9 cells.

Key simulation parameters were adjusted as follows:

- The UAV’s maximum permissible turning angle was set to $\pi/6$ to account for a more agile flight profile, enhancing the fidelity of the simulation in representing the dynamic maneuvering capabilities of contemporary UAVs.
- The grid granularity was established with a resolution of $N = 0.5$ km, improving the precision of our

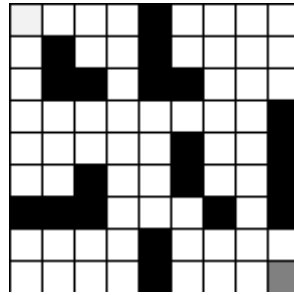


Fig. 5.1: Initial airspace setup

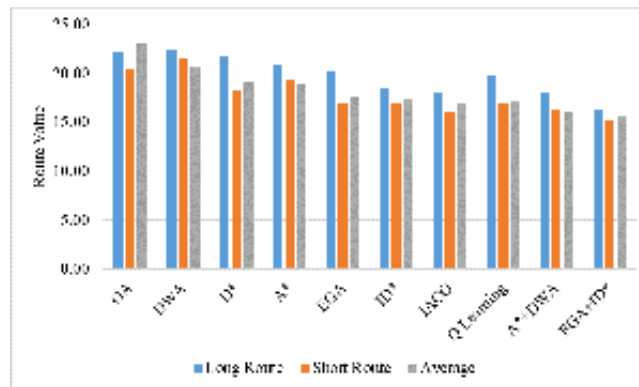


Fig. 5.2: Route length analysis

spatial analysis and the solution of the path-planning process.

- A safety margin was defined at 1 km, ensuring a conservative operational envelope for the UAV to prevent potential collisions with obstacles.
- The UAV's initial and target coordinates were plotted at (10, 10) and (90, 90), respectively, providing a diverse range of trajectory planning scenarios across the simulation space.

These simulation adjustments have been carefully chosen to test the boundaries of the proposed UAV path planning model, ensuring robustness, adaptability, and a high degree of environmental fidelity. Through this altered configuration, the model is subject to a wider range of test scenarios representative of complex agricultural terrains. Figure 5.1 shows the UAV's simulated airspace.

From Figures 5.2, 5.3, and Table 5.1, When the performance of the algorithms is compared using the average route values, the EGA+ID* approach emerges as the better standard, serving as our reference point. Relative to this, the Particle Swarm Optimization (PSO) method exhibited routes that were approximately 51.73% longer. Following closely, the Ant Colony Optimization (ACO) generated paths that were 47.48% longer than the EGA+ID*. The Genetic Algorithm (GA) trailed slightly behind, with 46.59% more extended paths than the proposed combined approach. The Dynamic Window Approach (DWA) displayed a marked improvement over the previous algorithms but still had routes approximately 31.57% longer than the EGA+ID*. When we consider other conventional algorithms, the disparity in performance becomes even more palpable: D* exhibited routes that were 22.34% longer, while A* demonstrated paths that were 20.39% more extended. When considering the state-of-the-art models, the Improved Ant Colony Optimization (IACO) and Q Learning had 7.80% and 9.20% longer routes, respectively. Impressively, the fusion of A* with the Dynamic Window Approach (A*+DWA) came remarkably close to the top-performing model, with only a 2.49% increase in path length over the EGA+ID*.

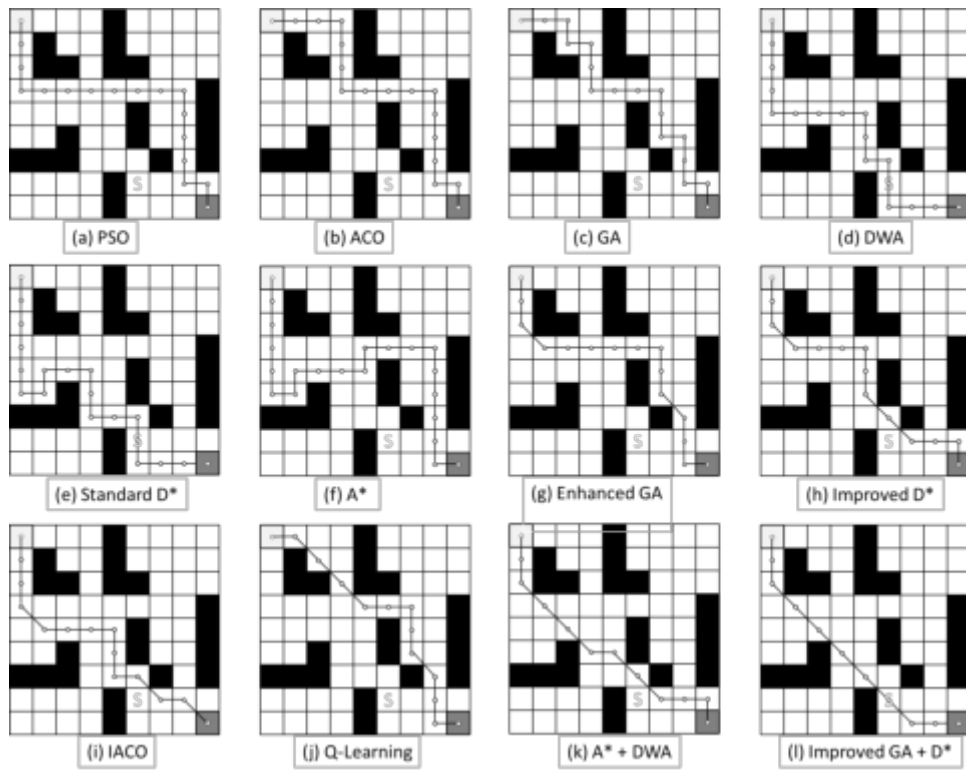


Fig. 5.3: Path planning results

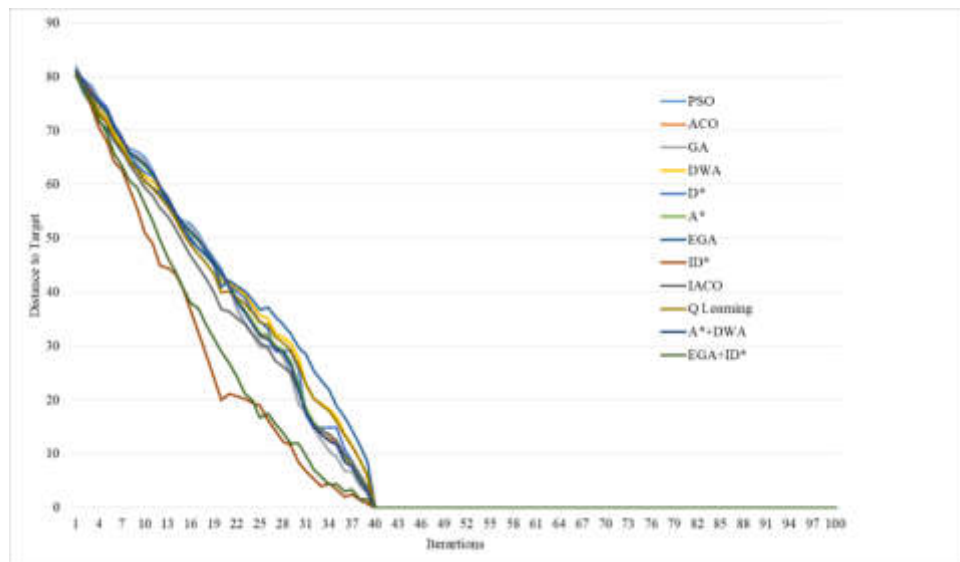


Fig. 5.4: Distance to target analysis

Table 5.1: Performance analysis of the compared models

Algorithms	Inflection Count	Time (s)	Accuracy
PSO	16	12.35	48.7
ACO	15	10.8	46.6
GA	15	11.63	48.3
DWA	15	10.46	45.5
D*	17	8.93	58.8
A*	17	9.02	56.4
EGA	12	8.83	74.1
ID*	11	9.57	74.83
IACO	13	7.96	73.36
Q Learning	12	7.07	81.23
A*+DWA	10	6.86	84.6
EGA+ID*	9	5.54	91.1

The data reveals a fascinating interplay between inflection count, time efficiency, and accuracy of various path-planning algorithms. The combined EGA+ID* model distinguishes itself as a frontrunner, completing its tasks in a mere 5.55 seconds and achieving an impressive accuracy of 91.1%. This superior performance is realized with a minimal inflection count of 9, implying a trajectory with fewer directional changes and a smoother path. In contrast, the D* and A* algorithms, despite having the highest inflection counts of 17, manage to hold their own. Specifically, D* highlights a respectable accuracy of 58.8%, slightly edging out A*. On the other end of the spectrum, algorithms like PSO and ACO, while being relatively faster than some counterparts, lag in accuracy, hovering around the mid-40s range.

Meanwhile, the A*+DWA fusion strikes a commendable balance, ensuring quick path planning in 6.863 seconds and delivering an accuracy of 84.6% with a mere 10 inflections. Overall, the results underscore the process of integrating global and local pathplanning models, as demonstrated by the unmatched efficiency and accuracy of EGA+ID*. The distance to the target for each iteration is displayed in Figure 6. Starting at 59.18 in the initial iteration, EGA+ID* demonstrates a steady decrease in distance values, reflecting its effective optimization capability. Compared to other algorithms, EGA+ID* maintains consistent performance, converging closer to the target as iterations progress. This pattern displays the efficiency and potential of the EGA+ID* model in optimization tasks, making it a promising choice for such applications.

6. Conclusion. The rapid evolution of smart agriculture, underpinned by the convergence of technology and traditional farming practices, underscores a transformative shift in the agricultural domain. Unmanned Aerial Vehicles (UAVs), central to this transformation, change how agricultural operations are visualized and executed. While their potential is undeniable, ensuring their efficient and safe operation in the dynamic environment of a farm presents considerable challenges. The key lies in the intricate process of path planning. Our proposed hierarchical framework, which amalgamates the strengths of the “Enhanced Genetic Algorithm using Fuzzy Logic” for broad trajectory planning and the “Improved D* Algorithm” for nuanced, real-time adjustments, is a step forward in addressing this challenge. This integrated approach not only guarantees efficient navigation but also bolsters the safety and energy conservation of UAVs in real-world agricultural settings. As we look to the future, the fusion of such sophisticated algorithms with UAV technology holds the promise of further elevating the standards of smart agriculture, driving the sector towards heightened sustainability and productivity.

REFERENCES

- [1] S. AGGARWAL AND N. KUMAR, *Path planning techniques for unmanned aerial vehicles: A review, solutions, and challenges*, Computer Communications, 149 (2020), pp. 270–299.
- [2] X. BAI, H. JIANG, J. CUI, K. LU, P. CHEN, AND M. ZHANG, *UAV Path Planning Based on Improved A* and DWA Algorithms*, International journal of aerospace engineering, 2021 (2021), pp. 1–12.

- [3] M. CHEN, Y. SUN, X. CAI, B. LIU, AND T. REN, *Design and implementation of a novel precision irrigation robot based on an intelligent path planning algorithm*, arXiv preprint arXiv:2003.00676, (2020).
- [4] Z. CUI AND Y. WANG, *UAV path planning based on multi-layer reinforcement learning technique*, Ieee Access, 9 (2021), pp. 59486–59497.
- [5] B. HAN, T. QU, X. TONG, J. JIANG, S. ZLATANOVA, H. WANG, AND C. CHENG, *Grid-optimized UAV indoor path planning algorithms in a complex environment*, International Journal of Applied Earth Observation and Geoinformation, 111 (2022), p. 102857.
- [6] F.-F. LI, Y. DU, AND K.-J. JIA, *Path planning and smoothing of mobile robot based on improved artificial fish swarm algorithm*, Scientific reports, 12 (2022), p. 659.
- [7] S. A. H. MOHSAN, N. Q. H. OTHMAN, Y. LI, M. H. ALSHARIF, AND M. A. KHAN, *Unmanned aerial vehicles (UAVs): Practical aspects, applications, open challenges, security issues, and future trends*, Intelligent Service Robotics, 16 (2023), pp. 109–137.
- [8] H. QIN, S. SHAO, T. WANG, X. YU, Y. JIANG, AND Z. CAO, *Review of autonomous path planning algorithms for mobile robots*, Drones, 7 (2023), p. 211.
- [9] C. QU, W. GAI, J. ZHANG, AND M. ZHONG, *A novel hybrid grey wolf optimizer algorithm for unmanned aerial vehicle (UAV) path planning*, Knowledge-Based Systems, 194 (2020), p. 105530.
- [10] S. SHAO, Y. PENG, C. HE, AND Y. DU, *Efficient path planning for UAV formation via comprehensively improved particle swarm optimization*, ISA transactions, 97 (2020), pp. 415–430.
- [11] B. B. SINHA AND R. DHANALAKSHMI, *Recent advancements and challenges of Internet of Things in smart agriculture: A survey*, Future Generation Computer Systems, 126 (2022), pp. 169–184.
- [12] D. C. TSOUROS, S. BIBI, AND P. G. SARIGIANNIDIS, *A review on UAV-based applications for precision agriculture*, Information, 10 (2019), p. 349.
- [13] G.-T. TU AND J.-G. JUANG, *UAV Path Planning and Obstacle Avoidance Based on Reinforcement Learning in 3D Environments*, in Actuators, vol. 12, MDPI, 2023, p. 57.
- [14] C. YAN, X. XIANG, AND C. WANG, *Towards real-time path planning through deep reinforcement learning for a UAV in dynamic environments*, Journal of Intelligent & Robotic Systems, 98 (2020), pp. 297–309.

Edited by: Vadivel Ayyasamy

Special issue on: Internet of Things and Autonomous Unmanned Aerial Vehicle Technologies for Smart Agriculture Research and Practice

Received: Jan 3, 2024

Accepted: Mar 11, 2024



SMART FERTILIZING USING IOT MULTI-SENSOR AND VARIABLE RATE SPRAYER INTEGRATED UAV

HAYDER M. A. GHANIMI*, R. SUGUNA†, JOSEPHINE PON GLORIA JEYARAJ‡, K SREEKANTH§, RAJASEKAR RANGASAMY¶, AND SUDHAKAR SENGAN||

Abstract. This paper introduces a “Smart Fertilizing Using Internet of Things (IoT) Multi-Sensors” system to enhance fertilizer management in agriculture. The system has four main parts: the Nutro Determining Unit (NDU), the Nutro Sensing Unit (NSU), the Nutro UAV Variable Fertiliser Spray System, and a Variable Rate Unmanned Aerial Vehicle (UAV) Sprayer model. The NDU collects vital data on Soil Moisture (SM) and Environmental Conditions (EnC) using advanced IoT cameras, while the NSU consolidates and normalises the data for advanced analysis using Heuristic Decision Trees (HDT) and Random Forest (RF) algorithms. In India, a data-driven UAV system uses IoT and UAV technologies to determine nutrient needs and create a prescription map for fertilizer application. The approach caused increases in the efficient utilisation of resources, Crop Yield (CY), and ecological footprint when it underwent evaluation in a crop maize field that was 14 hectares in size. A fresh benchmark for Smart Farming (SF) techniques has been set up by this method of operation, which is motivated by data and symbolises an important innovation in modern and ecologically conscious SF methods.

Key words: UAV, IoT, Sprayer Model, Smart Fertilizing, Crop Yield, Smart Farming

1. Introduction. The major developments that have currently taken advantage of the field of agriculture in the past few decades have caused the origin of revolutionary ideas like “Smart Farming” (SF) and “Precision Agriculture” (PA). New technologically focused SF approaches have succeeded conventional farming methods as an impact of these changes. The modern-era SF procedures leverage revolutionary technologies like autonomous devices, statistical analysis of data, and the Global Positioning System (GPS) in order to improve farmer productivity and effectiveness. High crop yields (CY), better consumption of resources, less ecological damage, and maximized use of resources are feasible because PA’s data is updated constantly. Precise SF tasks become possible with this method of treatment because it enables selective application on particular areas or crops [7].

The Internet of Things (IoT) generates an evolution in the agricultural sector by implementing an online network of connected devices, including cameras, sensor networks, and Unmanned Aerial Vehicles (UAV) [6]. The main objective of IoT devices is to contribute to making SF more productive by collecting and analyzing data about the crop and the Earth’s Soil Moisture (SM). Some examples of inventions that showcase this include autonomous irrigation and precise monitoring systems. By pairing to the World Wide Web in real-time, farm IoT devices provide agriculturalists access to additional valid data, which enables them to make more accurate selections [10].

Ethical practices in ecological awareness finance can now develop owing to this better connectivity, which improves profitability and helps in the effective control of resources. IoT technology plays an integral part in

*Information Technology Department, College of Science, University of Warith Al-Anbiyaa, Karbala, Iraq; Computer Science Department, College of Computer Science and Information Technology, University of Kerbala, Karbala, Iraq (hayder.alghanami@uowa.edu.iq).

†Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai 600062, Tamil Nadu, India (drsuguna@veltech.edu.in).

‡Department of ECE, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai 600062, Tamil Nadu, India. (josephineraj90@gmail.com).

§Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram 522502, Andhra Pradesh, India. (ksreekanth@kluniversity.in).

¶Department of Computer Science and Engineering, Alliance School of Advance Computing, Alliance University, Bengaluru 562106, India. (rajasekar.r@alliance.edu.in).

||Department of Computer Science and Engineering, PSN College of Engineering and Technology, Tirunelveli 627152, Tamil Nadu, India. (Corresponding Author, sudhasengan@gmail.com).

monitoring SM and environmental conditions (EnC), namely the levels of nutrients and SM, which is necessary for optimizing fertilizer consumption for successful economic crop development. Periodic data analysis makes it possible to identify the proper proportion of fertiliser to use, which, in consequence, serves to prevent spraying less or excess. Highly accurate and effective fertiliser treatments can be generated by using data pertaining to soil and crop requirements collected by IoT-enabled devices across a period of time frames [5].

A form of application that simplifies the use of pesticides is the Variable Rate Sprayer (VRS), which regulates the volume of fertiliser or pesticide sprayed based on particular regions and data in real-time. Better crop production, better conditions for crops, fewer waste products, and less negative environmental impact are the results of this approach's use.

More and more innovative farmers are employing VRS systems to integrate PA methods into their farming practices. This has the potential to minimize waste and the adverse impact of agricultural products on the natural world [3]. A multi-sensor IoT architecture and a prototype for SF fertilisation using a UAV-based VRS have been recommended in the present investigation on the "Smart Agriculture using IoT Multi-Sensors (SA-IoT-MS)" of the entire system.

The Nutro Sensing Unit (NSU), Nutro Gateway Unit (NGU), Nutro Decision Unit (NDU), and Nutro UAV (NUAV)-Variable Fertiliser Spray System (VFSS) are the 4 portions that make up the entirety of the system. The efficient functioning of the Sustainable Agriculture System (SAS) is enhanced by continually tracking both the soil quality of a crop and outside factors.

An array of parameters are recorded and digested by a unified system. These factors include soil water content, temperature, humidity, and NPK levels. The Decision-Making Process (DMP) revolves around the NDU, which generates a treatment map and forecasts nutrient requirements via a Heuristic Decision Tree (HDT) and Random Forest (RF) method. In order to achieve optimal use of resources while improving CY and good health, this map has been used to guide the accurate placement of fertiliser by the UAV system. By this creative approach, the developers have achieved significant progress towards attaining SAS standards.

This study article outlines the following: Chapter 2 starts with the existing literature analysis; Chapter 3 presents the framework hypothesis; Chapter 4 includes experimental research, and Chapter 5 concludes the research.

2. Literature Review. In order to effectively analyze soil nutrients, [1] developed an IoT-based system using a novel Nitrogen (N), phosphorus (P), and potassium (K) (NPK) sensor. Their use of fuzzy logic for data interpretation demonstrates the growing trend of incorporating complex data processing methods into SF technology.

[8] highlighted the importance of IoT in monitoring SM and EnC, specifically indoor plants, by measuring SM and NPK values and providing user feedback through an online data display, showcasing the integration of IoT with user-friendly interfaces in agriculture.

The authors emphasize the significance of precise data collection and analysis in agriculture, as highlighted by [2, 12]. They propose an SF system integrating Artificial Intelligence (AI) and sensor technology, focusing on energy-efficient deployment and sustainable SF. They also discuss the technical aspects of field data attainment systems in PA, emphasizing the need for accurate data in fertilizer and irrigation systems.

UAV technology has gained significant applications in SF [13, 4, 9], particularly precision farming. They have developed a pulse width modulation variable spray system, which an STM32 chip controls, showcasing the innovative integration of precise control mechanisms in UAV systems for spray farming.

3. System Model. The work puts forward the SA-IoT-MS system, an original concept developed with a clear goal to provide optimal fertiliser management in PA farms. The NUAU-Variable Fertiliser Spray System, the NSU, the NGU, and the NDU are the four interrelated elements that make up the basis of this proposed model. The NSU is not simply planned between the trees, but it has been connected with revolutionary IoT sensors [14, 15, 11, 16]. These sensors collect a wealth of data on the SM and EnC. The data obtained include temperature, humidity, pH, SM, and NPK. The primary objective of the distinct GSP-ID assigned to every NSU is to collect and monitor real-time data that is important in comprehending soil properties.

The NGU's function as a computing unit is to obtain the data collected from the NSUs and send it to them. The NGU may obtain data from numerous NSUs, do the initial analysis, and verify for reliability and consistency while sending it on for deeper analysis because of its architecture. It is the task of the NGU to

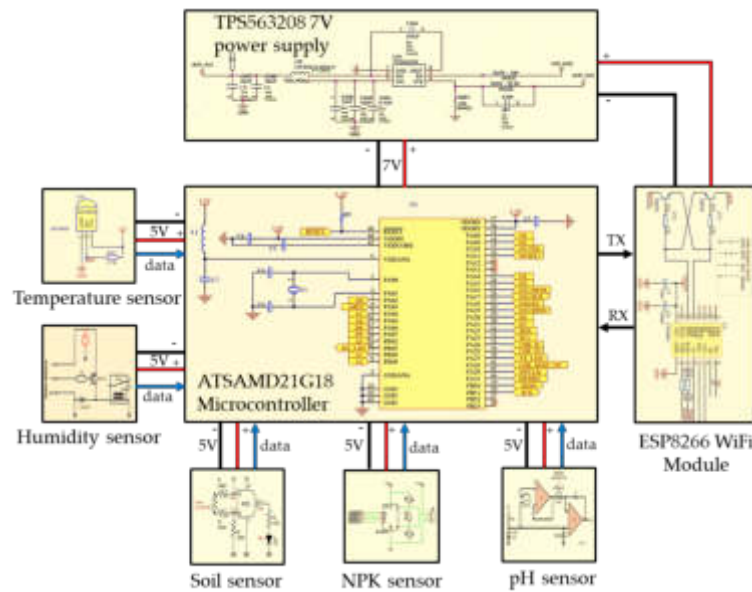


Fig. 3.1: NSU unit structure

ensure that the sensory data have access to the DMP entity uninterrupted any delays. The data is vulnerable to an HDT analysis by the NDU after the NGU stops receiving it. Using methods of empirical research, this elaborate study evaluates the temperature, SM, and NPK levels on an individual basis. In furtherance of analyzing the information that is presented, the NDU is assigned to detecting emerging patterns and trends. This attribute is vital in order to make intelligent DMP about fertiliser services, which will result in improved effectiveness and efficiency of deployments based on real-time and historical information.

In this framework, implementing the NUAV-VFSS is the final phase. This UAV-based system sprays fertilizer precisely where it is required, owing to the extensive instruction presented by the NDU. According to data analysis, the UAV employs VRS innovation to optimize the level of fertilizer sprayed in different regions of the crop. The productivity and efficacy of the fertilization method are significantly improved by this approach, which provides a personalized and efficient use of fertilizer.

3.1. NSU. Figure 3.1 illustrates the elements of the NSU that were developed for the purpose of this research to promote accurate data collection and effective communication in the proposed SF model. The NSU's primary operation is regulated by the ATSAM21G18 microcontroller, which effectively controls the unit's functions while using less power. The NSU collects an enormous amount of soil-based and environmental information using a number of unique sensors. The DSB18B20 temperature sensor is noted for its accuracy in measuring room temperature; the HR202 moisture sensor performs well at monitoring the level of SM in the atmosphere; and the Jxct soil sensor is great at analyzing the attributes of soil. Measuring SM and EnC in full is required for SF effectiveness. The I2C-SM sensor and the Sen0161 pH sensor have been used in the present study.

Reliable transfer of data for analysis is made possible by the ESP8266 Wi-Fi module, making it possible for simple internet access with the NGU. The NGU depends on the TPS563208 power system for reliable power control. This module provides all sensors and microcontrollers with stable and uninterrupted power, ensuring that they can function and collect data without delay in a number of crop settings.

1. **ATSAMD21G18 Microcontroller:** The NSU powers the ATSAM21G18 Microcontroller, a low-power, high-performance microchip on the ARM® Cortex®-M0+ platform that is ideal for home automation and industrial applications. Its 256 kb flash and 32KB SRAM provide ample memory for data processing. The ATSAM21G18 microcontroller is a versatile device with six configurable

Table 3.1: Microcontroller Configuration

Feature	Description
Microcontroller	ATSAMD21G18, ARM® Cortex®-M0+ based
Memory	256KB Flash, 32KB SRAM
Operating Frequency	Up to 48MHz
PWM Channels	20 channels
Power Supply Range	1.62V to 3.63V

Table 3.2: Wi-Fi Module Configuration

Feature	Specification
Processor	L106 32-bit RISC, 80MHz
Memory	32 KiB instruction RAM, 80 KiB user-data RAM, up to 16 MiB QSPI flash
WiFi	IEEE 802.11 b/g/n
GPIOs	16 pins
Interfaces	SPI, I ² C, I ² S, UART, ADC
Power Management	APSD for VoIP, Bluetooth co-existence
RF System	Self-calibrated

SERCOM modules, three 16-bit timers, a 32-bit real-time clock, and 20 PWM channels, enabling rapid data processing and accurate environmental monitoring. It also features a 14-channel 12-bit analog-to-digital converter and a 10-bit digital-to-analog. The ATSAMD21G18 microcontroller is a high-performance device that supports full-speed USB devices and embedded host functionality, can handle 120 touch channels, and operates within a power range of 1.62V to 3.63V, as detailed in Table 3.1.

2. **ESP8266 WiFi Module:** The ESP8266 Wi-Fi Module, a small and cost-effective System on Chip (SOC) with an integrated TCP/IP protocol stack, enables Arduino controllers to connect to Wi-Fi networks. This module additionally decreases the requirement for a CPU. The module's built-in processing and storage capabilities and high on-chip integration require less additional hardware, reducing the space required for the Printed Circuit Board (PCB). Because of its space-saving design, the ESP8266 is a good option for applications with limited free space for living. It is not essential to use external RF devices because the module's self-calibrated Radio Frequency (RF) system ensures reliable performance in a wide range of temperature and humidity conditions. Table 3.2 displays comprehensive setup details for the installed Wi-Fi module.
3. **Power Module:** The NSU devices are able to run properly because the TPS563208 input module provides them with power through its 3-A synchronous communication step-down inverters. The key objectives of this unit, enclosed in a SOT23 package and small in size, include easy operation, low idle current, and no reliance on external elements. It has a wide input voltage range (4.5V to 17V) and output voltage range (0.76V to 7V). It also has D-CAP2 mode control, which lets it respond quickly to transients, and continuous current mode, which works well with low loads. Additionally, the module ensures safety and dependability by including controls such as the current limit, UVP, and TSD. Furthermore, it performs well across a wide temperature range, ranging from -40° to 125°. Additionally, Table 3.3 contains the configuration information of the electric power module.

The following sensors are integrated into the NSU unit:

- *Temperature:* This investigation uses the DS18B20 sensor to measure the temperature in SF models. The precision, dependability, and ease of integration of this digital sensor led to its selection as the winner. It is a well-liked option in San Francisco because it accurately measures and controls the temperature in the agricultural environment.
- *Humidity:* SF applications, in which energy efficiency is a top priority, are a good fit for the HR202

Table 3.3: TPS563208 description

Feature	Specification
Output Current	3A
Control Mode	D-CAP2 Mode
Input Voltage Range	4.5V–17V
Output Voltage Range	0.76V–7V
Operating Mode	Continuous Current
Switching Frequency	580kHz
Shutdown Current	<10mA
Voltage Accuracy	2/cent at 25°
Soft Start	1.0mins
Package	6-pin SOT23 (1.6mm×2.9mm)
Temperature Range	-40° to 125°

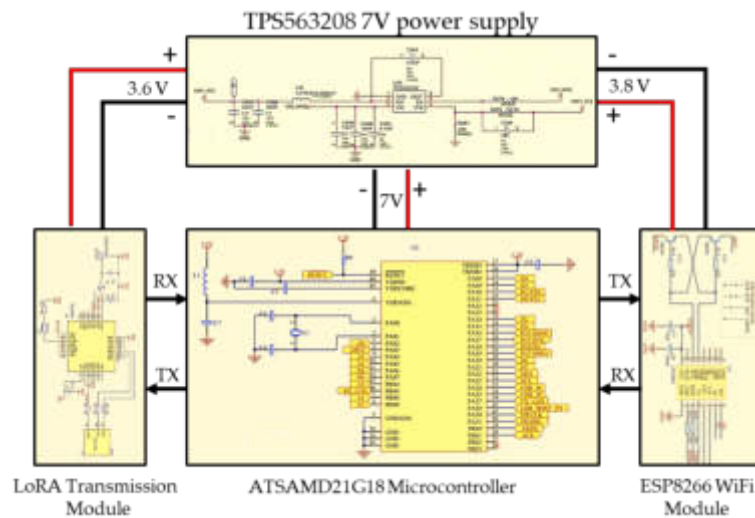


Fig. 3.2: NGU structure

sensor because of its improved moisture tracking capabilities, wide measurement range, and higher accuracy. As a result, it is an excellent option for SF applications.

- *SM*: The I2C-SM sensor employs the capacitive sensing method, resulting in an accurate and non-corrosive device with consistent long-term reliability, regardless of the predominant soil conditions. It is suitable for SF devices because of its low-voltage functioning, which promotes the use of energy.
- *pH*: Acidity in soils impacts crop growth, fertilizer absorption and soil health; the Sen0161 pH sensor is a significant device for monitoring this factor. Its investigations are vital for making smart choices about soil care along with growth because they are accurate and endure for an extended period of time.
- *NPK*: For precise soil mineral levels of difficulty, fertility tests, and SF use approaches, the LNPk-1 sensor is required. The system delivers accurate fertilizer data employing modern chemical tools for measurement.

3.2. NGU. Collecting and analysing data from NSU distributed around the trees is the task of the NGU, which is a vital part of the SF system. Its core is the ATSAM21G18 microcontroller, which has been designed for practical use. According to Figure 3.2, the NGU includes the ESP8266 Wi-Fi Module, which enables the creation of reliable wireless connections and the sending of data to a primary server or a cloud-based system.

Table 3.4: LoRA Module Description

Specification	Description
Frequency Range	433MHz
Modulation Techniques	FSK/GFSK/MSK/LoRa
Sensitivity	-136 dBm
Output Power	+20 dBm
Data Rate	<300 kbps
Operating Temperature	-40° to +80°
Standby Current	≤ 1mA
Supply Voltage	1.8V to 3.6V

The NGU has a LoRA transmission module that enables transmission over long distances. This technology is intended for use in low-power, wide-area network (LP-WAN) applications, making it especially useful for SF fields in which NSUs are spread across numerous land areas. While using minimal power, the module can send and receive data over enormous distances most efficiently. When it comes to ensuring that DMPs in remote or difficult locations have a link to an uninterrupted supply of data, the ability of this device to manage large distances without impacting battery efficiency or signal quality is paramount.

The capacity of the TPS563208 power module to provide support for the NGU, which comprises the LoRA module, implies that the operation will be stable and uninterrupted. The microcontroller and communication modules, such as Wi-Fi and LoRA, can function at their highest possible efficiency due to this power module, which provides an uninterrupted and reliable energy supply. This significantly enhances the overall reliability and performance of the NGU, as previously mentioned. As an outcome of the synergistic combination of cutting-edge microcontroller technology, numerous communication options, and a reliable electrical system, the NGU has been determined to be a vital and practical element within the overall structure of the SF.

1. **LoRa Transmission Module:** LoRa Transmission Modules, more specifically for use in SF, make it possible for IoT applications to participate in long-range wireless communication. These modules operate on the LoRaWAN network and use wavelengths that typically fall within the frequency spectrum of 864MHz to 915MHz, following the regulations of the different regions. LoRa modules can send data over ranges of up to 15 kilometres, even in areas that are susceptible to noise. LoRa modules are renowned for their low power consumption and high receiver sensitivity. These modules stand out for their autonomous power supply, typically powered by batteries and capable of lasting up to 10 years without needing a new battery.

The NGU employs the LoRA 433MHz SX1278 module. It is a cost-effective RF front-end transceiver that excels in long-range and low-data-rate applications. Its exceptional sensitivity (-136dB/m in LoRa modulation) and 20dB/m power output ensure reliable connectivity over long distances. The module operates on a 433MHz frequency, supports multiple modulation formats, and can function in extreme temperature ranges from -40° to +80°, making it versatile for varied environmental conditions. Additionally, it includes features such as a built-in temperature sensor with ultra-low standby current that operates within a 1.8 to 3.6V supply voltage range. The following table describes the LoRA module.

3.3. NDU. The NDU is fitted with a vital device called the Heuristic Random Forest (HRF), which is mainly deployed to determine whether or not different agricultural areas require fertilizer replenishment. This DMP is supported by a series of HAs that evaluate various agricultural factors. The following lists contain detailed descriptions and equations for each technique:

1. **Estimation of Evapotranspiration (ET):** For the goal of measuring the amount of water that evaporates as a result of evaporation and crop water loss, the estimated amount of ET is an important metric. Conducting a review of the water level is particularly significant in the SF sector because it has an indirect impact on the rate at which plants absorb fertilizers.

$$ET = ET_o \times K_c \quad (3.1)$$

Here, ET_o represents the reference ET, calculated based on climate data, including temperature, humidity, and solar radiation. K_c is the crop coefficient that varies according to different growth stages, signifying the crop's ET under specific conditions as compared to the reference.

2. **Water Retention Ratio (WRR):** The WRR is the most crucial measure of the soil's capacity to store water. If the WRR is higher, it means the soil can store additional moisture, which in turn impacts its capacity to maintain nutrient solutions and determine how to apply fertilizer plans. EQU (3.2) for WRR is:

$$WRR = \frac{FC - PWP}{AWC} \quad (3.2)$$

The field capacity (FC), the permanent wilting point (PWP), and the available water content (AWC) of the soil are represented in this equation.

3. **Nutritional Deficiency Predictor (NDP):** This study compares the required nutrient levels for a particular crop cycle with the readily accessible nutrient levels in the soil. The goal of this heuristic is to make a prediction about the possibility of nutritional deficiencies. Obtaining this data is essential for determining the necessity of supplemental nutrition. The NDP's EQU (3.3) stands for:

$$NDP = \sum (N_{\text{req}} - N_{\text{avail}}) \times F_{\text{factor}} \quad (3.3)$$

Where N_{req} is the nutrient requirement for a specific crop stage, N_{avail} represents the available nutrient level in the soil and F_{factor} is a crop-specific adjustment factor.

4. **Soil pH Influence (SPI):** Using the SPI heuristic, one can identify how the pH of the soil affects the availability of nutrients. This heuristic assists with changing the pH of the soil in order to enhance the absorption of nutrients, which is essential because numerous crops have distinct optimal pH levels for production. The is calculated as EQU (3.4).

$$SPI = pH_{\text{opt}} - pH_{\text{soil}} \quad (3.4)$$

where pH_{opt} is the optimal pH level for the crop, and pH_{soil} is the current soil pH level.

This experiment set up an HRF model using advanced algorithms and machine learning (ML) methods to determine the nutrient requirements of SF regions. Factors such as ET, WRR, NDP, and SPI ensure that an inclusive and accurate method for adding nutrients across multiple SF zones is provided.

This work begins by computing the RF model using equations 1 through 4. Next, this study will collect and normalize past and present information based on the heuristics EE, WRR, NDP, and SPI. This helps to guarantee that the input is reliable and scalable. Subsequently, we fine-tune the model's set-up, including the number of trees, feature splitting method, and DMP at each node. Every tree in the RF conducts a separate independent evaluation based on the heuristics, which contributes to the final decision about the value of added nutrients across multiple SF areas.

When training the model, it is crucial to use labelled data, which shows if additional nutrition would have been needed under scenarios that were similar in the past. The model can collect information from these data points, allowing it to make intelligent decisions about adding nutrients. After the training phase, we verify the model's precision and reliability using a distinct data set. After verification, we set up the HRF model for practical application. The HRF model uses real-time data associated with ET, WRR, NDP, and SPI to predict the need for additional nutrients in various zones of the SF field. We use the predictions to produce a complete prescription map, making them highly significant. Figure 3.3 below displays the sequence of the NDU module's functions.

After the testing and installation of the HRF model, its essential task will be to analyse real-time data points, such as ET, WRR, NDP, and SPI. This will allow it to calculate nutritional supplementation needs across all SF field zones. The creation of a complex prescription map, which is a vital tool for the NUAV-VFSS, is an effective end to this process. The nutritional map is a digital data file that houses a vast amount of information about the GPS coordinates of the SF field and the determined rates of fertilizer application. A meticulously coded program in MATLAB® is responsible for the development of this map (Fig. 3.4), which

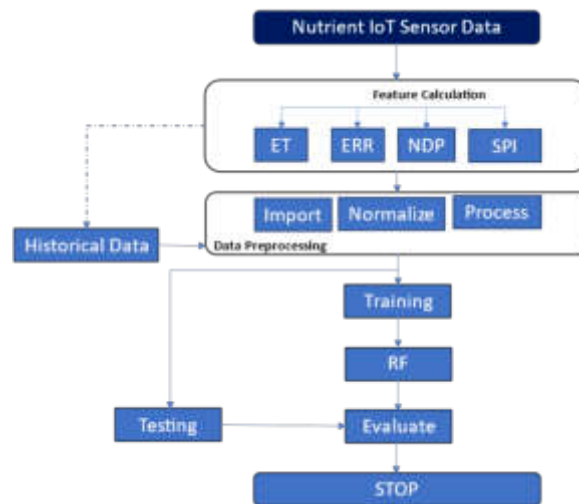


Fig. 3.3: Flowchart of the fertilizer DMP.

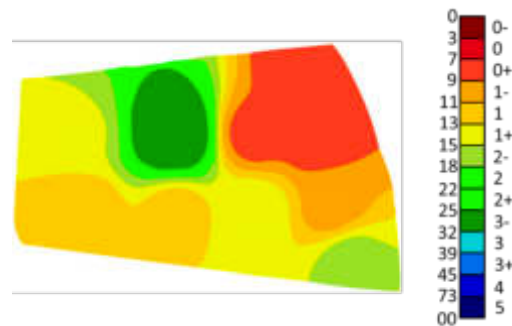


Fig. 3.4: Prescription Map.

combines the data on the range of vital nutrients with the geographical coordinates in order to provide an in-depth representation of the supply of nutrients.

This study edges a nutrient map onto a 10×10 -meter fishnet grid, enabling a more precise approach to nutrient management. This grid divides the map into manageable segments, extracting centroid coordinates and nutrient data. This work uses these data to calculate fertilizer application rates for each grid cell, ensuring they align with the target nutrient requirements. This investigation transmits the resulting digital nutrient application map, complete with precise latitude and longitude coordinates, to the NUAUV-VFSS. This data equips the UAV to experiment with targeted fertilizer applications and guarantees a customized nutrient measure for each field zone.

3.4. NUAUV-VFSS. The goal of NUAUV-VFSS is to deliver focused fertilizer applications with unprecedented precision. The system aims to deliver these applications. Figure 3.6 provides a conceptual model of the UAV-VFSS system, showcasing its complex design and functionality. This assignment utilized the DJI Agras T30 drone, a product of SZ DJI Technology (Shenzhen) Co., Ltd. SZ DJI Technology (Shenzhen) Co., Ltd. developed this specific UAV by integrating specific modules for prescription map conversion and spray control. The HRF model generates a prescription map, which the UAV uses to follow. This guarantees the precise placement of fertilizers where they are required.

Table 3.5: UAV-VFSS Description

Component	Description	Manufacturer/Supplier
DJI Agras T30	UAV for the VFSS	SZ DJI Technology (Shenzhen) Co., Ltd.
KLP02-E KA	Micro diaphragm Pump to Control the flow of fertilizer.	Kamoer Fluid Technology (Shanghai) Co., Ltd.
F110-015	Spray the Nozzle to ensure the fertilizer is evenly distributed.	Mid-South Ag. Equipment, USA
ATSAMD21G18	Microcontroller for the spray control subsystem.	—
YF-S201	The magnetic hall flow sensor measures the system's flow rate, which ranges from 1–30l/min.	DATAQ Instruments, Inc, Ohio, USA

A micro diaphragm pump (KLP02-E KA, Kamoer Fluid Technology (Shanghai) Co., Ltd.) is critical to the performance of the VFSS because it ensures that the fertilizer flow is adequately controlled within the system framework. This pump, when used with a spray nozzle (F110-015 FanTip Nozzle 110, produced by Mid-South Ag. Equipment in the United States), makes it possible to promote the even distribution of fertilizer across the zones that were previously specified. The ATSAMD21G18 microcontroller is the most significant element of the spray control subsystem. It is responsible for regulating the spraying system's performance. Therefore, we require the accuracy and dependability of this microcontroller to ensure accurate fertilizer distribution. The VFSS incorporates a magnetic hall flow sensor (YF-S201, DATAQ Instruments, Inc., Ohio, USA) with a measuring range of 1–30l/min. This sensor plays a crucial role in monitoring and adjusting the fertilizer flow rate, ensuring it aligns with the prescribed demands on the prescription map. Table 3.5 breaks down each element of the UAV-VFSS into smaller sections and provides more detailed information.

The NDU unit transmits the prescription MAP, received through the Wi-Fi module and uploaded into the ATSAMD21G18 microcontroller.

The built-in GPS of the UAV-VFSS continuously transmits real-time positional data to the microcontroller on the UAV as it traverses the landscape. The microcontroller initiates a key-matching function as soon as it receives the current coordinates of the UAV. The microcontroller accomplishes this by comparing the UAV's position with the encrypted GPS data in the preloaded prescription map. This map divides the fertilizer application process into distinct areas across the field, providing complete guidance on successful application.

The UAV moves across the field, aligning with a prescription map. A microcontroller identifies the current unit area and retrieves the corresponding prescription value, which is the specific quantity of fertilizer for that area. A variable spray controller then receives this information and adjusts the UAV's spray mechanism to apply the prescribed amount of fertilizer, ensuring the exact amount matches the prescription map for that specific area.

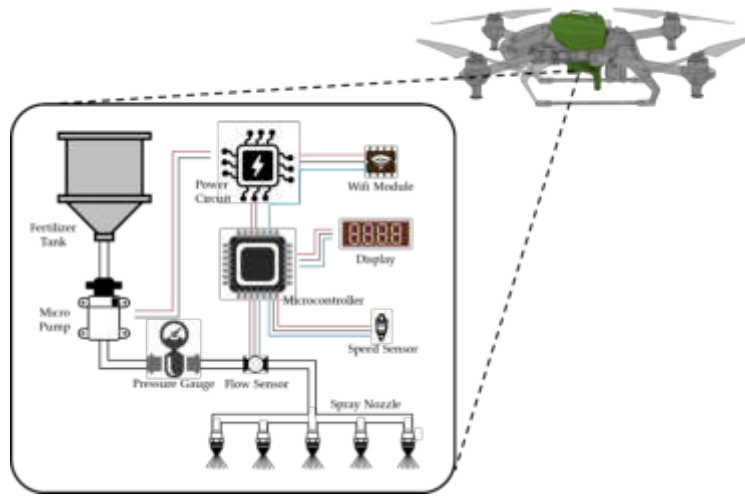


Fig. 3.6: UAV-VFSS Structure.

Table 4.1: Data from three sites showing the results of the Nutrient sensor data

Variable	Unit	Site	Mean	Min	Max	Median	S.D	Variance	Skewness	Kurtosis
Nitrogen	Kg/ha	12	80	70	95	82	7.5	56.25	0.2	-0.5
		42	120	110	135	122	8.0	64	0.1	-0.4
		63	100	90	110	102	6.5	42.25	-0.2	-0.3
Phosphorus	Kg/ha	12	40	35	50	41	4.8	23.04	-0.1	0.2
		42	55	50	65	56	5.0	25	0.3	-0.1
		63	48	43	55	49	4.0	16	-0.4	0.5
Potassium	Kg/ha	12	85	80	95	87	5.0	25	0.2	-0.2
		42	75	70	85	76	4.5	20.25	-0.3	0.3
		63	90	85	100	91	4.8	23.04	0.1	-0.4

4. Experimental Study. This work tested the “SA-IoT-MS” system in a practical field experiment on a 14-hectare maize field in Maharashtra, India. This method divided the 35,000-square-foot area into smaller zones for detailed analysis and precise intervention, ensuring the system’s controlled and measurable application in the region’s agronomic conditions.

This simulation advantageously placed the NSUs to collect vital SF data such as SM, temperature, pH, and nutrient content from three sites. The data collection spanned four months (15th March 2022 to 30th July 2022), covering the entire maize growing season and capturing agronomic variables and changes throughout different stages of crop development. The results are crucial in assessing the effectiveness of the “SA-IoT-MS” system in a practical SF setting, as shown in 4.1 and 4.2.

- Kg/ha: Kilograms per hectare.
- Mean: The average value of nutrient content.
- Min: The minimum recorded value of nutrient content.
- Max: The maximum recorded value of nutrient content.
- Median: The middle value in the range of nutrient content.
- S.D (Standard Deviation): Measures the variation in the nutrient content.
- Variance: The square of the standard deviation.
- Skewness: A measure of the asymmetry of the distribution of nutrient content.
- Kurtosis: A measure of whether the data are heavy-tailed or light-tailed relative to a normal distribu-

Table 4.2: Data from three sites showing the results of the environmental sensor data

Variable	Unit	Site	Mean	Min	Max	Median	S.D	Variance	Skewness	Kurtosis
Temperature	°	12	26	22	30	26.5	2.5	6.25	0.1	-0.2
		42	27	24	31	28.5	2.2	4.84	-0.1	0.3
		63	27	24	31	27.2	2.1	4.41	0.2	-0.3
Humidity	%	12	65	60	70	65.5	3.2	10.24	-0.2	0.1
		42	68	65	72	68.4	2.8	7.84	0.0	-0.1
		63	70	68	73	70.3	1.6	2.56	-0.1	0.2
SM	%	12	30	25	35	30.5	3.1	9.61	0.1	-0.3
		42	28	26	31	28.7	1.5	2.25	-0.2	0.4
		63	32	28	36	32.2	2.5	6.25	0.0	-0.2
pH		12	6.5	6.2	6.8	6.5	0.2	0.04	0.0	0.0
		42	6.7	6.5	6.9	6.7	0.1	0.01	-0.1	0.1
		63	6.6	6.4	6.8	6.6	0.1	0.01	0.1	-0.1

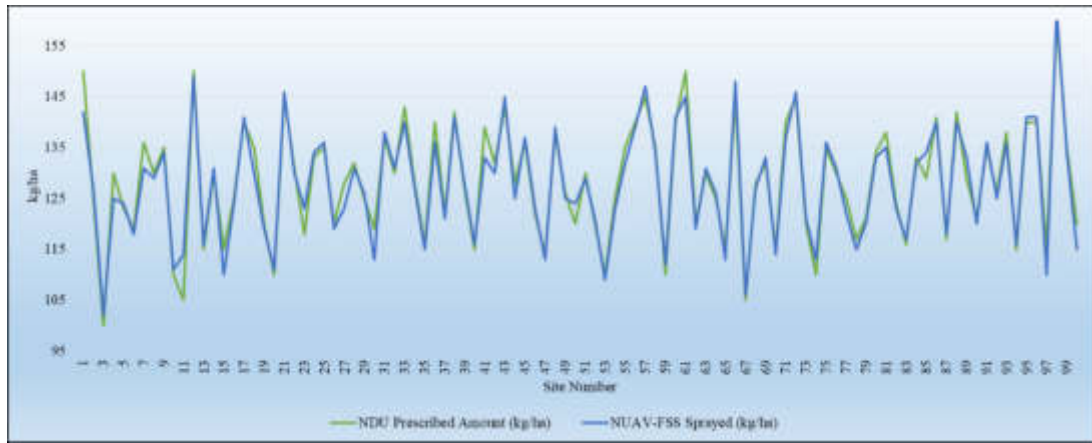


Fig. 4.1: NDU prescribed fertilizer amount vs NUAU-VFSS Sprayed amount.

tion.

- °: Degrees Celsius.

The NDU uses sensor data to determine the precise amount of fertilizer (measured in kilograms per hectare, kg/ha) needed at each site in the field. This system uses the heuristic RF algorithm to determine nutrient needs based on unique conditions. The NDU calculates the required levels and relays them to the NUAU-VFSS, which uses a control module to apply the prescribed fertilizer levels. This method is more efficient and environmentally sustainable, minimizing waste and runoff while making fertilizer application more efficient and effective. Figure 4.1 showcases the effectiveness of the integrated system, comparing the prescribed fertilizer levels with the actual quantities applied by the NUAU-VFSS. The data demonstrates the system’s high efficiency and precision, closely mirroring the NDU’s prescriptions and proving its value as an integral part of SF practices.

Figure 4.1 highlights that the NUAU-VFSS is a fast system, which is an accomplishment worthy of recognition. This demonstrates the system’s ability to precisely adhere to the NDU’s rules and adapt to the unique standards imposed by various field regions. SM places a significant value on precision and adaptability, significantly contributing to higher CY, reduced resource waste, and environmentally friendly SF practices.

5. Conclusion. The “SA-IoT-MS” system, combined with a variable-rate UAV sprayer, significantly advances precision agriculture. It improves fertilizer management efficiency and effectiveness by combining IoT technology and UAV capabilities. The deployment of NSUs collects environmental and soil parameters data,

enabling data-driven agriculture. The NGU and NDU process this data to make precise fertilizer requirement decisions. The NUAV-VFSS executes these decisions, catering to different agricultural zones. A field experiment in Maharashtra, India, validated the system's functionality and highlighted its potential for revolutionizing SF practices. The results demonstrated crop yields and resource optimization improvements, promoting sustainable agriculture.

This model sets a precedent for future developments in SF, showcasing the benefits of integrating advanced technologies like IoT and UAVs into agricultural operations.

REFERENCES

- [1] A. ALI, T. HUSSAIN, N. TANTASHUTIKUN, N. HUSSAIN, AND G. COCETTA, *Application of Smart Techniques, Internet of Things and Data Mining for Resource Use Efficient and Sustainable Crop Production*, Agriculture, 13 (2023), p. 397.
- [2] G. LAVANYA, C. RANI, AND P. GANESHKUMAR, *An automated low cost IoT based Fertilizer Intimation System for smart agriculture*, Sustainable Computing: Informatics and Systems, 28 (2020), p. 100300.
- [3] N. LIN, X. WANG, Y. ZHANG, X. HU, AND J. RUAN, *Fertigation management for sustainable precision agriculture based on Internet of Things*, Journal of Cleaner Production, 277 (2020), p. 124119.
- [4] Y. LU, M. LIU, C. LI, X. LIU, C. CAO, X. LI, AND Z. KAN, *Precision Fertilization and Irrigation: Progress and Applications*, AgriEngineering, 4 (2022), pp. 626–655.
- [5] S. MISHRA, *Internet of things enabled deep learning methods using unmanned aerial vehicles enabled integrated farm management*, Heliyon, 9 (2023).
- [6] A. MONTEIRO, S. SANTOS, AND P. GONÇALVES, *Precision agriculture for crop and livestock farming – Brief review*, Animals, 11 (2021), p. 2345.
- [7] V. SAIZ-RUBIO AND F. ROVIRA-MÁS, *From smart farming towards agriculture 5.0: A review on crop data management*, Agronomy, 10 (2020), p. 207.
- [8] S. R. SALEEM, Q. U. ZAMAN, A. W. SCHUMANN, AND S. M. Z. A. NAQVI, *Variable rate technologies: development, adaptation, and opportunities in agriculture*, in Precision Agriculture, Elsevier, 2023, pp. 103–122.
- [9] D. SU, W. YAO, F. YU, Y. LIU, Z. ZHENG, Y. WANG, T. XU, AND C. CHEN, *Single-neuron PID UAV variable fertilizer application control system based on a weighted coefficient learning correction*, Agriculture, 12 (2022), p. 1019.
- [10] A. SUBEESH AND C. MEHTA, *Automation and digitization of agriculture using artificial intelligence and internet of things*, Artificial Intelligence in Agriculture, 5 (2021), pp. 278–291.
- [11] S. SUDHAKAR, V. VIJAYAKUMAR, C. S. KUMAR, V. PRIYA, L. RAVI, AND V. SUBRAMANIASWAMY, *Unmanned Aerial Vehicle (UAV) based Forest Fire Detection and monitoring for reducing false alarms in forest-fires*, Computer Communications, 149 (2020), pp. 1–16.
- [12] J. SUN, A. M. ABDULGHANI, M. A. IMRAN, AND Q. H. ABBASI, *IoT enabled smart fertilization and irrigation aid for agricultural purposes*, in Proceedings of the 2020 International Conference on Computing, Networks and Internet of Things, 2020, pp. 71–75.
- [13] B. SWAMINATHAN, S. PALANI, S. VAIRAVASUNDARAM, K. KOTECHEA, AND V. KUMAR, *IoT-driven artificial intelligence technique for fertilizer recommendation model*, IEEE Consumer Electronics Magazine, 12 (2022), pp. 109–117.
- [14] S. WEN, Q. ZHANG, J. DENG, Y. LAN, X. YIN, AND J. SHAN, *Design and experiment of a variable spray system for unmanned aerial vehicles based on PID and PWM control*, Applied Sciences, 8 (2018), p. 2482.
- [15] J. YANG AND M. GU, *Design of the Auto-Variable Spraying System Based on ARM9&Linux*, in 2018 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), IEEE, 2018, pp. 1–2487.
- [16] R. ZHANG AND L. SONG, *Study of variable spray control system based on machine vision*, in 2014 IEEE 13th International Conference on Cognitive Informatics and Cognitive Computing, IEEE, 2014, pp. 455–458.

Edited by: Vadivel Ayyasamy

Special issue on: Internet of Things and Autonomous Unmanned Aerial Vehicle Technologies for Smart Agriculture Research and Practice

Received: Jan 3, 2024

Accepted: Jun 20, 2024



RECURRENT NEURAL NETWORK BASED INCREMENTAL MODEL FOR INTRUSION DETECTION SYSTEM IN IOT

HIMANSHU SHARMA, PRABHAT KUMAR* AND KAVITA SHARMA[†]

Abstract. The security of Internet of Things (IoT) networks has become an integral problem in view of the exponential growth of IoT devices. Intrusion detection and prevention is an approach used to identify, analyze, and block cyber threats to protect IoT from unauthorized access or attacks. This paper introduces an adaptive and incremental intrusion detection and prevention system based on RNNs, to the ever-changing field of IoT security. IoT networks require advanced intrusion detection systems that can identify emerging threats because of their various and dynamic data sources. The complexity of IoT network data makes it difficult for traditional intrusion detection techniques to detect potential threats. Using the capabilities of RNNs, a model for creating and deploying an intrusion detection and prevention system (IDPS) is proposed in this paper. RNNs work particularly well for sequential data processing, which makes them an appropriate choice for IoT network traffic monitoring. NSL-KDD dataset is taken, pre-processed, features are extracted, and RNN-based model is built as a part of the proposed work. The experimental findings illustrate how effective the suggested approach is at identifying and blocking intrusions in Internet of Things networks. This paper not only demonstrates the effectiveness of RNNs in enhancing IoT network security but also opens avenues for further exploration in this burgeoning field. It presents a scalable, adaptive intrusion detection and prevention solution, responding to the evolving landscape of IoT security. As IoT networks continue to expand, the research enriches the discourse on developing resilient security strategies to combat emerging threats in scalable computing environments.

Key words: IoT, IDS, Machine Learning, Deep Learning, RNN

1. Introduction. Internet of Things is a network that allows everyday electronic devices to exchange data and coordinate their actions. The level of interconnection holds out the possibility of greater ease and efficiency in our day-to-day activities. Nevertheless, just as there are two sides to every coin, there are considerable worries associated with the Internet of Things (IoT), notably in regard to its security. The significance of IoT networks cannot be overstated; that has the potential to transform industries, improve the quality of life, and drive economic growth.

IoT networks are driving innovation in industrial automation, making manufacturing processes more efficient and reducing downtime [1]. Industries are using connected machines to streamline their operations and reduce downtime. In transportation, they are paving the way for autonomous vehicles, which have the potential to revolutionize mobility and reduce accidents. In the realm of energy, IoT enables the smart grid, optimizing energy distribution and promoting energy efficiency. In smart cities, IoT facilitates the creation of urban environments where infrastructure, transportation, and utilities are interconnected [2]. Cities are becoming smarter by embedding sensors that can help manage traffic in real-time, turn off streetlights when no one's around, or even alert about potential infrastructure issues. This promises sustainability, reduced traffic congestion, and an improved quality of life for urban residents. In agriculture, precision farming driven by IoT allows farmers to optimize crop yields, conserve resources, and promote sustainable practices, addressing the global challenge of food security. In the realm of healthcare, IoT devices enable remote patient monitoring, personalized treatment plans, and timely interventions. Patients can receive better care, and healthcare providers can operate more efficiently.

IoT devices often have limited computational resources and may need robust security mechanisms [3]. This makes them vulnerable to a wide range of cyber threats. Attackers can exploit vulnerabilities in IoT devices to gain unauthorized access, compromise data integrity, and disrupt critical services. Data privacy

*Computer Science and Engineering Department, National Institute of Technology Patna, India (himanshugbpuat@gmail.com, prabhat@nitp.ac.in).

[†]Computer Science and Engineering Department, Galgotias College of Engineering Technology, Greater Noida, India (kavitasharma06@yahoo.co.in).

is another critical concern. Many IoT devices collect sensitive information, including personal and location data. Unauthorized access to this data can lead to privacy breaches, identity theft, and legal consequences. Furthermore, the evolving cyber threat landscape poses a continuous challenge. Malicious actors are becoming increasingly sophisticated, using techniques like zero-day exploits and ransomware to target IoT vulnerabilities. To address these security challenges, effective solutions are essential and robust security measures need to be implemented.

Intrusion Detection Systems, also known as IDS, have traditionally been a primary line of defence against various types of cyberattacks [4]. The role of IDS in network security has been crystal clear: act as vigilant watchdogs, constantly monitoring traffic, detecting anomalies, and triggering alerts for potential threats. Traditional IDSs, built upon signature-based or rule-based mechanisms, have served well within the constraints of their design. However, the dynamism and complexity of IoT demand a more nuanced approach. Simple pattern matching or static rule sets are often ineffectual against sophisticated or zero-day attacks on IoT networks [6].

Over the past few years, machine learning and deep learning paradigms have emerged at the forefront of technological innovation and helped to cope up with such Security Concerns. Among the various architectures within deep learning, Recurrent Neural Networks (RNNs) hold particular promise for time-sequence data, which is intrinsic to network traffic in IoT. Unlike traditional feed-forward neural networks, RNNs possess the ability to 'remember' past inputs through their internal memory. This capability allows them to discern patterns in sequential data, making them particularly suited for IDS in IoT, where understanding temporal data sequences is crucial.

However, the mere existence of RNNs only sometimes translates to their effective implementation in IDS for IoT. Several challenges need to be addressed—the high dimensionality of network data, the real-time processing requirements of IoT, and the scalability concerns posed by billions of interconnected devices, to name a few. Moreover, while the application of RNNs in various domains like natural language processing or stock market prediction is well-documented, their tailored application for IoT intrusion detection is still nascent. This research seeks to bridge this knowledge gap, offering a comprehensive exploration of the design, implementation, and efficacy of an RNN-based IDS for IoT. As the narrative unfolds, the intricacies of the IoT landscape, highlighting its unique challenges, has been explained. Subsequently, an in-depth exploration of the RNN architecture will set the stage for understanding its applicability in the IDS domain. Through rigorous experimentation and evaluation, this research will not only propose but also validate the superiority of the RNN-based IDS for IoT, especially when compared against traditional models like J48, Random Forest (RF), Support Vector Machines (SVM), Multilayer Perceptron (MLP), and Naive Bayes(NB).

The core problem addressed in this research revolves around the inadequacy of existing intrusion detection and prevention systems to effectively safeguard IoT networks. Specifically, the research questions guiding this study include:

1. How can Recurrent Neural Networks (RNNs) be employed to detect and prevent intrusions in IoT networks?
2. What are the challenges and opportunities associated with implementing RNN-based techniques in the context of IoT network security?
3. How does the performance of RNN-based intrusion detection and prevention compare with traditional methods in terms of accuracy, adaptability, and real-time responsiveness?

The significance of this study lies in its potential to transform the landscape of IoT network security. By introducing a novel approach that leverages RNNs for intrusion detection and prevention, this research contributes to the development of adaptive and resilient security mechanisms for IoT networks. These mechanisms are vital to ensuring the continued growth and adoption of IoT technologies across various sectors, as security concerns have been a major impediment to realizing the full potential of IoT.

2. Literature Review. This section examines the existing research on security and intrusion detection approaches for the Internet of Things. It will provide a comprehensive overview of the current state of IoT security, highlighting the vulnerabilities specific to IoT networks. Additionally, it will explore the various intrusion detection and prevention methods employed in traditional networks and IoT environments. This review will serve as the foundation for identifying gaps in the literature and setting the stage for the proposed RNN-based approach.

IoT applications are growing to smart grids, retail, residences, cities, and healthcare despite forecasts. Security is needed to avoid service disruption, illegal access, and cyberattacks like tampering and others to assure data accuracy and process efficiency. ML/DL is common in IDS development. IDSs protect IoT devices and systems from security and operation threats. Intrusion detection systems (IDS) are essential in IoT networks, which increasingly encompass critical infrastructure including healthcare, transportation, and energy. With the help of intrusion detection systems, network managers may quickly identify, address, and collect vital information needed to stop and lessen security risks.

Traditional machine learning methods like SVM [7], [8], K-Nearest Neighbor (KNN) [9], ANN [10], Random Forest (RF) [11], [12], and others [13] have been successful for intrusion detection systems. On the other hand, the DL method has outperformed ML in terms of accuracy, particularly for large datasets. Because picking features takes time and they won't know which characteristics are valuable until the model is trained and evaluated, researchers creating machine learning algorithms must exercise caution and only extract features that can improve the model. Machine learning is challenged when dealing with datasets of different sizes since it is not always easy to extract the most predictive features [14]. Furthermore, because deep learning models can independently extract properties from massive data sets, they outperform traditional machine learning techniques and are more accurate [15].

Kumari et al. proposed a semi-supervised intrusion detection system [16] using a hybrid SVM-FCM clustering platform for classification. This was an extra semi-supervised intrusion detection system. Active SVM uses a modest amount of labeled input and a lot of unlabeled data. This was done to prove that active learning SVM can identify like a typical support vector machine after N iterations. For multi-class classification, the FCM classifier was used on data items around support vectors. This model used SVM and FCM classifier engines for intrusion detection. If both classifiers regarded an input instance normal, we may confidently call it normal. If the SVM engine classified the input instance as an outlier and the FCM engine identified its sub-category, the instance is considered abnormal and the sub-class is selected by selecting the circle with the highest fuzzy membership and geographical proximity to the support vectors.

In order to identify malicious attacks in IoT contexts, Otoum et al. [17] introduced a novel DL-based intrusion detection system to resolve the challenges associated with protecting IoT nodes. In their proposed model, the spider monkey optimization (SMO) algorithm and the stacked deep polynomial network (SDPN) are combined to achieve the highest detection and recognition rates. SMO selects the most relevant attributes from the datasets, while SDPN classifies the output as normal or aberrant. Using DL to identify intrusions with recurrent neural networks (RNN-IDS) was recommended by Yang et al. [18]. They show through their experimental results that RNN-IDS is ideally adapted for producing IDS with good accuracy and that it outperforms conventional ML both binary and multi-class techniques. Dawoud et al. [19] presented a deep learning-based intrusion detection system for SDN-based IoT architecture. SDN modeling was used for the IoT security, scalability, and resilience enhancing purposes, whereas Restricted Boltzman Machine (RBM) was used as the engine for intrusion detection. This serves as an example of the integration of SDN and IoT. The suggested model was tested, evaluated, and validated by utilizing the KDD Cup'99 dataset, on which it produced a competitive performance more than 94% in terms of precision and accuracy.

Khan et al. [20] employed an ensemble-based voting classifier, where the final prediction was derived by combining the conventional machine learning algorithm with voting on its predictions. Using a stacking-based ensemble model, IoT devices are better able to detect anomalies in IoT networks, according to Naz et al. [21]. To enhance the effectiveness and precision of ensemble-based IDS, Bhati et al. [22] implemented ensemble-based IDS with XGBoost, which improves the accuracy. In [23], Arko et al. presented an overview of several machine learning techniques that can be used to identify potentially harmful or out-of-the-ordinary data, as well as the most effective approach for two datasets: the first dataset was created from data exchanged between sensors, and the second dataset is UNSW-NB15.

The use of ensemble learning, in which many methods/models or experts are put to use in order to solve a specific artificial intelligence-based problem, was another approach that researchers took in order to ensure the strong security of the IoT. In the context of the problem of intrusion detection, ensemble learning fosters stronger generalization, and the voting amongst the various strategies of ensemble give higher detection accuracy than the individual models, according to the proposal made by Illy et al. [24].

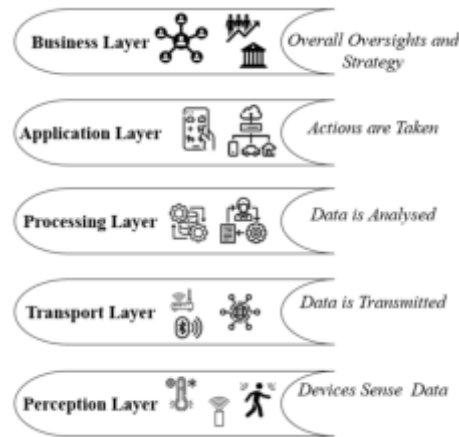


Fig. 3.1: IoT Architecture

In [25], Verma et al. investigated the viability of machine learning classification techniques for defending the IoT against DoS attacks. The Classifiers are evaluated using well-known datasets such as CIDD5-001, UNSWNB15, and NSL-KDD. Some cyber security experts have modified deep learning components to accomplish ML features for cyber-security, including IoT.

Yin et al. examined the structure of a deep learning-based intrusion detection system (IDS) and presents a novel RNN-IDS approach[26]. An comprehensive study examines the model's operationality in binary and multiclass classification scenarios and the effect of neuron count and learning rate changes on its efficacy. Using benchmark datasets, it is compared to J48, artificial neural network, random forest, and support vector machine. The authors suggest GPU acceleration to reduce training time, avoid exploding and vanishing gradients, and study the classification performance of LSTM, Bidirectional RNNs algorithms in intrusion detection.

Khan et al. proposed a deep learning-based intelligent IDS for IoT networks to address the security issues[27]. A Recurrent Neural Network with Gated Recurrent Units (RNN-GRU) can classify assaults across the physical, network, and application levels. This suggested model is trained and tested using the ToN-IoT dataset, which is unique for a three-layered IoT system and offers new attacks compared to other publicly available datasets. The proposed model's performance was analyzed using accuracy, precision, recall, and F1-measure, with Adam and Adamax optimization techniques. Adam was found to perform best.

3. IoT Architecture. The Internet of Things, or IoT for short, is a bit like a huge, worldwide web where computers and everyday objects are connected to each other. Think of it as a world where your fridge, watch, car, and even your shoes can 'talk' to each other through the internet. For all these things to work smoothly, we need a plan or structure, just like building a house. This plan is called the IoT architecture. At its core, the IoT architecture can be described as multi-layered, each serving a specific purpose, working together to deliver an interconnected, intelligent ecosystem. Let's break down this architectural framework in Figure 3.1.

3.1. Perception Layer (Device Layer). Imagine stepping into a dense forest, with every rustling leaf, chirping bird, or distant animal footstep communicating a piece of information. That's precisely the role of the Perception Layer. Often termed the physical or device layer, it's the frontline where real-world data is gathered. Comprising sensors, actuators, and other IoT devices, this layer perceives or senses the environment. Whether it's a smart thermostat sensing room temperature or an agricultural sensor gauging soil moisture, data collection begins here.

3.2. Transport Layer. Having collected the data, the next step is its relay to central hubs for further action. Enter the Transport Layer. Acting as the communication bridge, this layer ensures data moves from devices to data centers using a myriad of transmission mediums[30]. This could be via satellite, cellular networks, Wi-Fi, or even more niche protocols like Zigbee. The fundamental task here is secure, swift, and efficient data

transmission.

3.3. Processing Layer (Middleware Layer). All of the analysis that is done on data takes place at the Processing Layer. One could compare it to a location where information is stored, worked on, and interpreted. After being entered into databases, the raw data is subsequently transformed into information that can be utilized by specialized tools and computers. For example, by analyzing the data from a smart thermostat, it is possible to determine how to regulate the heating in order to save water and energy.

3.4. Application Layer. Application Layer is responsible to put the information received from Middle layer for practical use. Here, specific applications tailored for end-users interpret the processed data to offer tangible services. In a smart home setting, based on data from various sensors, the application layer might adjust lighting, heating, or even play your favourite song once you walk in. Essentially, this layer personalizes the IoT experience, translating processed data into relatable user actions[29].

3.5. Business Layer. Finally, at the top of all resides the Business Layer. Beyond the complexities of devices and data, this layer aligns the entire IoT architecture with overall business objectives. By analyzing data patterns, consumer behaviours, and device performance, strategic business decisions emerge. Whether it's launching a new product, optimizing an existing one, or even exploring uncharted market territories, this layer ensures the IoT system remains profitable, scalable, and aligned with overarching business objectives.

IoT architecture can be thought as a well-organized city where every part has a role. Every layer is crucial, from the devices that sense things to the pathways that transport data, the brains that process information, the hands that act on it, and the wise tree overseeing it all. This amazing plan lets our world of connected devices work together, making our lives easier and smarter. As more and more things around us start 'talking' to each other, knowing a bit about this architecture helps us appreciate the magic of the IoT world.

4. IDS for IoT. An Intrusion Detection System (IDS) for the Internet of Things (IoT) is paramount due to the inherent vulnerabilities associated with IoT devices and their increasing pervasiveness. Given the unique characteristics of IoT environments, traditional IDS might not be directly suitable, necessitating specialized approaches[7]. Here's a classification of Intrusion Detection Systems for IoT:

4.1. Based on Placement. Intrusion Detection Systems (IDS) can be fundamentally classified by their placement within the system they monitor. Host-based IDS (HIDS) operate on individual IoT devices. They focus on the internals of the device, such as system logs, processes, and system calls. Their primary advantage is their ability to effectively detect insider attacks and anomalies that manifest within a specific device. In contrast, Network-based IDS (NIDS) are centered on monitoring network traffic. By capturing and analyzing packets transmitted across the network, they are particularly apt at detecting unauthorized access or Distributed Denial of Service (DDoS) attacks that exploit network vulnerabilities[31].

4.2. Based on Detection Method. The detection methodology behind an IDS plays a critical role in its efficacy. Signature-based IDS operate using predefined patterns or signatures of known threats, making them adept at identifying recognized threats, but they are inherently limited when it comes to zero-day attacks. Anomaly-based IDS, on the other hand, rely on historical data to build a profile of what is considered "normal" behavior. When the current behavior deviates significantly from this profile, an alert is triggered. While this approach can detect previously unknown attacks, it might also lead to false positives. Specification-based IDS take a slightly different approach by using well-defined specifications that describe correct operation, and they raise alerts when there are deviations from these specifications. These are especially suitable for environments where correct behaviors can be meticulously defined. Lastly, Hybrid IDS merge the techniques of signature and anomaly-based detection to strike a balance between detection rates and false positives.

4.3. Based on the Type of IoT Environment. The type of IoT environment can dictate the design and priorities of an IDS. For instance, Home IoT IDS are designed specifically for smart home devices, such as thermostats, cameras, and smart appliances. They prioritize the privacy of users while ensuring usability. In industrial settings, the Industrial IoT (IIoT) IDS focus on ensuring system uptime, safety, and resilience in places like manufacturing plants and power grids. Healthcare IoT IDS, intended for medical devices and systems, place an unsurpassed emphasis on patient safety and data integrity. And in the realm of transportation,

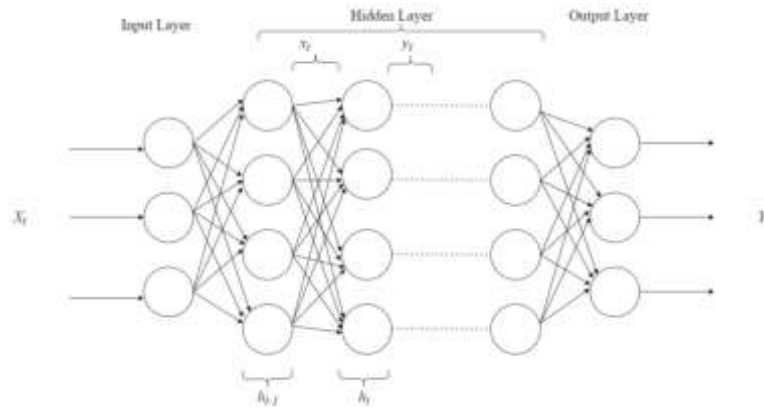


Fig. 5.1: RNN model

Vehicle IoT IDS cater to connected cars and vehicular networks, where passenger safety and real-time response are paramount.

4.4. Based on Operational Capability. In terms of operational capability, IDS can be collaborative or standalone in nature. Collaborative IDS are designed with multiple IDS nodes that collaborate and share information, offering a holistic view of the network and the ability to correlate events across diverse devices. This collaborative approach often leads to more robust detection and mitigation strategies. In contrast, Standalone IDS operate independently without the need for collaborative data. While they might be simpler and easier to deploy, they may lack the comprehensive view that collaborative systems offer. In essence, given the multifaceted nature of IoT devices and networks, an optimal IDS often necessitates a blend of these categorizations, each tailored to the unique requirements and threat landscapes of the environment.

5. Proposed Method. Proposed RNN-based IDS framework leverages the sequential processing capabilities of RNNs to analyze patterns in network traffic and detect intrusions. By using the NSL-KDD dataset, which is a benchmark in the IDS domain, the model can be trained to recognize a wide variety of intrusion patterns relevant to IoT environments. The combination of real-time processing, alert systems, and feedback loops ensures the IDS remains dynamic, effective, and up-to-date in the ever-evolving landscape of IoT security threats. Given the sequential nature of network traffic data, RNNs can potentially excel in identifying patterns and anomalies.

RNN has a looped or recurrent hidden layer, which allows it to maintain a 'memory' of previous inputs in its internal structure. This is what enables it to process sequences of data rather than single data points. As shown in FIG. 5.1, Inside Input layer, at each time step t , the RNN receives an input vector x_t . This vector will usually be an encoded form of the data for that time step, such as a word in a sentence or a feature in a time series. In Hidden Layer, The recurrent layers computes the hidden state h_t at time step t . This hidden state is a function of the input x_t at the current time step and the hidden state h_{t-1} from the previous time step. In output layer, at each time step t , the RNN produces an output vector y_t . This output can be computed based on the hidden state h_t and, if necessary, the input x_t .

Training an RNN involves adjusting its weights based on the difference between its predicted outputs and the actual outputs for a sequence. This is done using a variant of the backpropagation algorithm called Backpropagation Through Time (BPTT). BPTT works by unrolling the entire network for a sequence and applying the standard backpropagation algorithm, considering the temporal depth introduced by the recurrent layer.

As shown in FIG. 5.2 the process of an RNN-based Intrusion Detection System begins with the preprocessing of the Dataset. This includes label encoding, feature scaling, and feature selection to make sure the data is compatible. After the data has been pre processed, it is split into two sets: training (80%) and testing (20%).

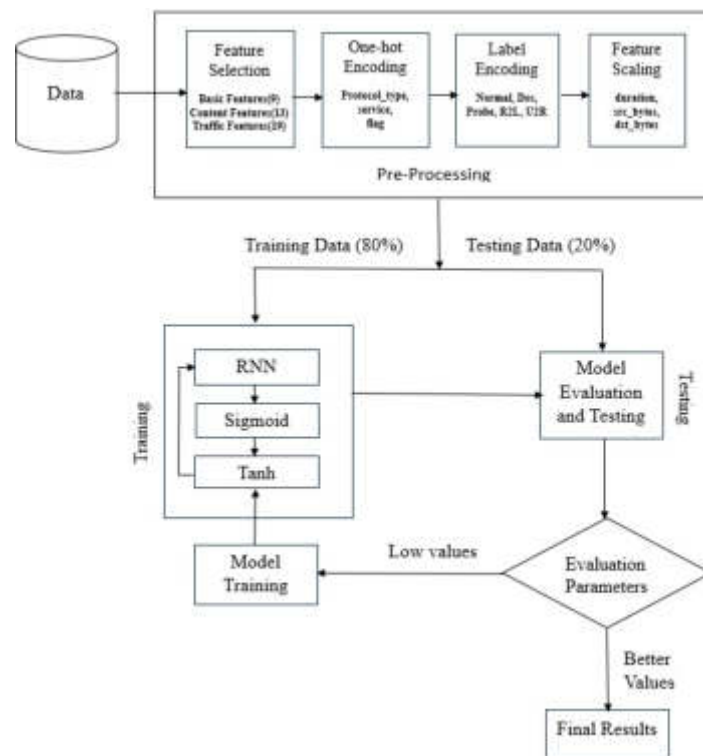


Fig. 5.2: Proposed RNN IDS

The training sample is used to build and train the RNN model. During training the model fits the data for 100 epochs with a batch size of 32. This is how it learns to spot patterns that are linked to intrusions. After that, the model is put to the test using the testing dataset to see how well it works using accuracy, precision, recall, and F1-score as measures. If the model's original evaluation metrics show that it isn't working as well as it could, it goes through incremental training, which involves fine-tuning its parameters even more until it gets good results. This iterative review process makes sure that the model's ability to find and stop cyber threats is always getting better. Once the model works well as it should, it can be used for real-time attack detection in live systems, which is a strong way to protect them.

5.1. Preprocessing of Dataset. Intrusion Detection Systems (IDS) often deal with large and complex datasets that require preprocessing to be effectively used for detecting malicious activities. Proper preprocessing is crucial as the quality and relevance of the data directly impact the model's performance. In this section various steps involved in Data Preprocessing are discussed followed by dataset description.

5.1.1. Dataset Description. NSL-KDD[28] is an improved version of the famous KDD Cup '99 dataset. The NSL-KDD dataset has made a name for itself in cybersecurity study, especially in the area of Intrusion Detection Systems (IDS). In this digital age, where network breaches and cyberattacks are getting smarter and happening more often, it is very important to have effective and accurate IDS. As a result, the NSL-KDD dataset has become an important tool for study into creating, testing, and improving different IDS models by providing a standard against which to measure and contrast their effectiveness. To fully understand what the NSL-KDD dataset is and how it can be used, it is important to go back to where it came from: the KDD Cup '99 dataset, which was created in 1999 as part of the Third International Knowledge Discovery and Data Mining Tools Competition. Even though it has problems like a huge number of duplicate records and built-in biases, the KDD'99 dataset quickly became the standard for IDS study. To address such limitations, the NSL-KDD was created to eliminate redundancies and give a more balanced dataset for constructing and assessing IDS

models. NSL-KDD is notable for its comprehensive and diverse composition, encapsulating various aspects of network interactions and potential intrusions, this includes:

A. Variety of Features. It includes a large group of 41 features that cover a wide range of topics, such as basic features of each TCP connection, content features that show what's inside the packets, and traffic features that are estimated using a two-second time window. The features can be broadly divided into three groups namely Basic features, Content features and Traffic features. Basic Features encompass attributes derived directly from the connection. Examples include the duration of the connection, the type of protocol used (e.g., TCP, UDP, ICMP), and other foundational data attributes. Content Features are derived from the content of the connections, such as the number of failed login attempts. These attributes provide insights into the suspicious behavior exhibited within the connection. Traffic Features are Computed with respect to a temporal window, these features capture network traffic statistics, analyzing patterns over a specified interval. These are further split into "time-based" and "connection-based" traffic features.

B. Multi-class Labels. Instances are divided into "normal" and several "attack" kinds. These are further broken down into four main types of attacks: DoS (Denial of Service), R2L (Remote to Local), U2R (User to Root), and Probing. Binary and Multi-class Classification: The attack types allow for both binary classification (normal vs. attack) and multi-class classification, which opens up a lot of theoretical and practical options.

C. Training and Test Sets. The dataset is split into "KDDTrain+" and "KDDTest+" sections, which make it easier to train, test, and validate models while keeping the lines between them clear to stop data leaks.

5.1.2. Feature Selection. In the realm of IDS for IoT, the relevance of features might differ from traditional network environments. For example, IoT devices often have resource constraints and unique patterns of network traffic. Therefore, selecting features that best characterize the IoT device behaviour is crucial. By focusing on the most relevant features, models can be more interpretable, faster, and potentially yield better performance The KDD'99 dataset initially has 41 features, categorized into basic features, content features, and traffic features.

Basic Features (9). These are derived from the packet headers without inspecting the payload, e.g., duration, protocol type, and service.

Content Features (13). These include features extracted from the payload like the number of failed login attempts.

Traffic Features (19). These are computed with respect to a window interval and are either time-based or connection-based. Using methods like Pearson's correlation coefficient can help determine if some features are highly correlated. If two features have high correlation, it might be beneficial to keep only one of them to avoid redundancy.

5.1.3. One-Hot-encoding:. One-Hot-Encoding is used to convert all categorical properties to binary properties. One-Hot-Endcoding requirement, the input to this transformer must be an integer matrix expressing values taken with categorical (discrete) properties. The output will be a sparse matrix in which each column corresponds to a possible value. It is assumed that the input properties have values in the range $[0, n_values]$. Therefore, to convert each category to a number, properties must first be converted with LabelEncoder.

There are 3 categorical attributes in this dataset are "Protocol_type", "service", and "flag" excluding "label" attribute. These features, although packed with essential information, are represented as text or categorical values, which are not inherently quantifiable and thus not directly compatible with RNN algorithms.

One-hot encoding is a favored technique for converting categorical data into a format that can be provided to RNN model. The process essentially creates a binary column for each category and indicates the presence of the category with a "1" or "0". Let's break down the one-hot encoding process for each of these features.

Protocol type. This feature indicates the type of protocol used for the connection, such as "tcp", "udp", or "icmp". Instead of these textual values, one-hot encoding would result in three new binary columns named "protocol_tcp", "protocol_udp", and "protocol_icmp". For a specific record in the dataset, if the protocol type is "tcp", the "protocol_tcp" column would have a value of "1" while the other two columns would be "0".

Service. The "service" attribute is a bit more complex as it delineates the network service on the destination, e.g., "http", "ftp", "telnet", and so on. Given the diverse range of services in the NSL-KDD dataset, one-hot encoding would result in multiple new binary columns, one for each service type. For instance, if a specific

record has the service type "http", then the "service_http" column would be "1", while all other 'service_*' columns would be "0".

Flag. Representing the status of the connection, typical values might include "SF", "S0", "REJ", etc. Just like the earlier attributes, each unique flag value would get its binary column. So, if a specific connection record had its flag set as "REJ", the corresponding "flag_REJ" column would hold a "1", with all other 'flag_*' columns set to "0".

Post one-hot encoding, the NSL-KDD dataset will have an expanded feature set with new binary columns replacing the original categorical ones. This transformation ensures that the data is in a numerical format, making it suitable for RNN without losing the categorical information's granularity. It's crucial, however, to note that this process can increase the dimensionality of the dataset significantly, especially if categorical features have numerous unique values. As such, after one-hot encoding, dimensionality reduction techniques might be considered to optimize the dataset's size without compromising the integrity of the information.

5.1.4. Label Encoding. The NSL-KDD dataset, a cornerstone in network intrusion detection research, underwent a transformation to simplify the representation of its diverse range of attacks. A large number of attacks were categorized into five broad categories, and to make these categories machine-friendly and facilitate easier computation, label encoding was applied. Initially, the dataset had various textual tags indicative of different kinds of attacks. To standardize and streamline this, the tags were remapped as follows:

Normal Activities: Previously labeled with various tags indicating normal behavior, these were consolidated and encoded with the value 0.

DoS (Denial of Service) Attacks: All tags specific to different types of DoS attacks(neptune, back, land, pod, smurf, teardrop, mailbomb, apache2, processtable, udpstorm, worm) were unified under the umbrella term "DoS" and were encoded with the value 1.

Probe Attacks: These are attacks where the malicious actor scans the network to gather information or find known vulnerabilities. All such attacks(ipsweep, nmap, portsweep, satan, mscan, saint) were labeled as "Probe" and assigned the encoded value 2.

R2L (Remote to Local) Attacks: In these attacks(ftp_write, guess_passwd, imap, multihop, phf, spy, warez-client, warezmaster, sendmail, named, snmpgetattack, snmpguess, xlock, xsnoop, httptunnel), an attacker who does not have an account on the target machine tries to gain access. Such attempts, previously labelled with various specific tags, were brought together under "R2L" and encoded with the value 3.

U2R (User to Root) Attacks: In U2R attacks, the attacker starts with access to a normal user account on the system and tries to exploit some vulnerability to gain root privileges. All such tags(buffer_overflow, loadmodule, perl, rootkit, ps, sqlattack, xterm) were encoded with the value 4.

The transformation process ensured the dataset became more streamlined. Instead of dealing with a multitude of tags that can make data processing and analysis cumbersome, especially for Deep learning models, we now have a standard set of five encoded labels. This not only helps in reducing the complexity but also in improving the efficiency of subsequent computations. To achieve this encoding, a straightforward mapping mechanism was used. A typical process would involve iterating over the dataset, examining the existing attack tag, and then replacing it with the new encoded value. This encoding, though seemingly simple, is a crucial step in data preprocessing, especially when the data is meant to be fed into machine learning or deep learning models. Properly encoded labels ensure models train effectively and provide meaningful results. Given the critical importance of network intrusion detection in today's hyper-connected world, such streamlined data representations play a pivotal role in advancing cybersecurity research and solutions.

5.1.5. Feature Scaling. In the domain of data preprocessing for deep learning models, feature scaling stands as a pivotal step to standardize the range of independent variables or features of the data. This process is paramount, especially in datasets with features that have different scales, as it can drastically impact the performance of certain algorithms. The KDD dataset, renowned in the realm of network intrusion detection, is no exception to this rule. For the NSL-KDD dataset, taking into account the varying magnitudes, units, and range of the features, the decision was made to apply logarithmic scaling, a specialized scaling method. This method is particularly useful when dealing with data that spans several orders of magnitude. By applying logarithmic transformations, we can diminish the effects of outlier values and compress the scale on which the

data lies, rendering it more manageable and interpretable. The features duration, src_bytes, and dst_bytes are taken into account. Their original ranges were considerably broad and spanned multiple magnitudes. However, by applying the logarithmic scaling method, these features were transformed to more condensed ranges. Specifically:

For the duration feature, post-logarithmic scaling, the range was condensed to [0, 4.77].

The src_bytes feature, after the application of logarithmic scaling, had its values fall within the range [0, 9.11].

Similarly, the dst_bytes feature was scaled such that its values now lie in the [0, 9.11] range.

It's noteworthy to mention that before applying the logarithmic scaling, a small constant might be added to the feature values to handle instances of values being zero, since the logarithm of zero is undefined. In essence, the logarithmic scaling of the NSL-KDD dataset's features ensures that the variances in the data's magnitude do not negatively influence the performance of machine learning algorithms. This transformation not only promotes better convergence during model training but also contributes to a more accurate and insightful representation of the underlying patterns and structures within the dataset.

5.2. Methodology. The purpose of this study was to use the NSL-KDD dataset to construct and assess an intrusion detection system (IDS) based on RNNs. Two different forms of configuration were used for proposed RNN-IDS: multiclass classification and binary classification. The goal of using both classification techniques was to evaluate proposed RNN model's adaptability and effectiveness in differentiating between different types of attacks and normal traffic.

An RNN-based intrusion detection system (IDS) is built and evaluated using the NSL-KDD dataset. The proposed RNN-IDS made use of two distinct configuration types: multiclass and binary classification. The purpose of combining the two classification methods was to test how well the suggested RNN model could distinguish between malicious and normal traffic.

Given an IoT network traffic dataset(NSL-KDD dataset), the task is to classify sequences of network data into one of N categories, such that "normal" and "Attack" in case of Binary Classification & "normal", "Dos", "Probe", "R2L", and "U2R" in case of Multiclass Classification

Let:

$X = \{x_1, x_2, \dots, x_T\}$: A sequence of feature vectors, where T is the length of the sequence.

$Y = \{y_1, y_2, \dots, y_T\}$: The corresponding labels or categories.

The objective is to Model a Function f using RNN such that $f(X) \approx Y$

RNNs are designed to recognize patterns in sequences of data by utilizing memory elements. The primary component of the RNN is its hidden state, which gets updated at each time step of the sequence as depicted by Equation 5.1.

$$h_t = \sigma(Wx_t + Uh_{t-1} + b) \tag{5.1}$$

$$E = mc^2 \tag{5.2}$$

$$\int_a^b f(t) \left(\sum_i E_i B_{i,k,x}(t) \right) dt \tag{5.3}$$

where:

h_t is the hidden state at time t .

W and U are weight matrices

b is the bias vector.

σ is a non-linear activation function, often the hyperbolic tangent (tanh).

Output would be as follows:

$$y_t = \phi(Vh_t + c) \tag{5.4}$$

where:

V is the weight matrix for the output.

c is the output bias.

ϕ is a softmax function when the task is multi-class classification, providing a probability distribution over the N .

For training the IDS, a suitable loss function, L as depicted by equation 5.5, is categorical cross-entropy for classification tasks:

$$L(Y, \hat{Y}) = - \sum_{t=1}^T \sum_{n=1}^N y_{t,n} \log(\hat{y}_{t,n}) \quad (5.5)$$

where $\hat{y}_{t,n}$ is the predicted probability of the n^{th} class at time t , and $y_{t,n}$ is a binary indicator (0 or 1) if label n is the correct classification for observation t .

For binary classification using RNN-IDS, the hidden layer uses 80 neurons and the activation function is hard sigmoid, while the output layer uses tanh. For instance, sensor updates are common examples of how IoT devices commonly broadcast data in a specified pattern. Anomaly detection may rely heavily on the temporal dependencies present in this data. On the other hand, a model based on RNNs could do better. A comparison is conducted to ascertain the RNN-IDS's effectiveness in relation to more conventional machine learning models. 'J48', 'RF', 'SVM', 'MLP', and 'NB' were used to assess RNN-IDS's performance by way of the evaluation metrics. In order to determine the F1 score, Precision, Accuracy, and Recall, these contrasting Machine Learning Models are constructed and implemented on the NSL-KDD dataset.

5.3. Evaluation. Using the NSL-KDD dataset, Proposed work used an RNN model to identify and classify four attack types: Dos, Probe, R2L, and U2R, along with normal traffic labels. The model's performance was carefully compared against five popular Machine Learning (ML) methods: J48, Random Forest (RF), Support Vector Machines (SVM), Multilayer Perceptron (MLP), and Naive Bayes(NB).

To allow for thorough comparisons, the model was trained on a stratified dataset split and evaluated using the following metrics.

Accuracy: The percentage of true predictions in our model compared to total predictions.

Precision: It measures positive prediction accuracy, calculated as the ratio of true positive outcomes to the sum of true positives and false positives.

Recall (Sensitivity): Rate of true positive predictions compared to true positives and false negatives.

F1 Score: It is a balanced measure of precision and recall, particularly in skewed datasets. Can be calculated as the harmonic mean of Precision and Recall.

The performance of the RNN-IDS model for binary classification (Normal, anomaly) and multi class classification (such as Normal, DoS, R2L, U2R, and Probe) has been the subject of investigation in two separate studies that have been created specifically for this purpose. These experiments are designed at the same time as standard experiments and results are compared other machine learning strategies J48, NB, RF, MLP, SVM.

A variety of errors occurred during the development of the RNN-based model for the binary and multiclass categorization of different attacks, each providing information about a different set of difficulties. Overly sensitive models or data noise can cause False Positives (FP), in which the model incorrectly predicts an attack. This is avoided by modifying the threshold of the model and using post-processing. Conversely, attacks that the model was unable to identify as False Negatives (FN) were caused by over-regularization and inadequate modeling of specific attack patterns. Reducing FN errors required improving feature representation, taking into account more intricate models, or modifying regularization. Accurate forecasts of attacks are indicated by True Positives (TP) and True Negatives (TN), respectively. Constant fine-tuning, feature engineering, and striking a balance between sensitivity and specificity is used to improve model accuracy. Errors resulted from missing and irrelevant features, hence feature representation is regularly assessed and improved.

5.3.1. Binary Classification. In the binary classification scenario, the NSL-KDD dataset was framed to classify network activities into two broad categories: 'Normal' and 'Attack'. The 'Attack' category encompassed all four varieties of attack labels (Dos, Probe, R2L, and U2R), consolidating them into a single overarching class and making a binary classification setup possible. LSTM input layer must be 3D the meaning of the 3 input dimensions are: samples, time steps, and features. The number of samples is assumed to be 1 or more.

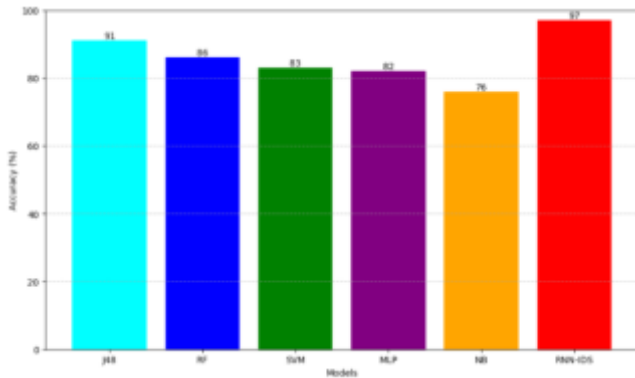


Fig. 5.3: Comparison of Accuracy between Proposed(RNN-IDS) and other(j48,RF,SVM,MLP,NB) methods applied on NSL-KDD dataset.

reshape() function takes a tuple as an argument that defines the new shape. To obtain strong comparative analysis with other models, the model was trained on a NSL-KDD dataset and evaluated using the following metrics.

Accuracy. Accuracy is calculated as the proportion of correctly predicted instances to the total number of instances in the dataset. It is a metric used to evaluate the IDS model's ability to correctly classify network traffic as either normal or malicious. Mathematically, accuracy can be expressed as follows:

$$Accuracy = \frac{\text{True Positives (TP)} + \text{True Negatives (TN)}}{\text{Total Instances (P + N)}} \quad (5.6)$$

True Positives (TP) is the count of attack instances correctly identified as attacks.

True Negatives (TN) True Negatives (TN) is the count of normal instances correctly identified as normal.

P is the total actual positive instances (actual attacks).

N is the total actual negative instances (actual normal activities).

The performance of RNN-IDS model in terms of Accuracy is superior to other classification algorithms in binary classification as shown in FIG. 5.3.

For Binary classification Accuracy of RNN-IDS came out to be 97% which is 6% more as compared to the accuracy of best model among as compared with other.

Precision. Precision, also known as the positive predictive value, is an essential evaluation metric, where the cost of false positives (incorrectly identified as an attack) may be significant. Precision attempts to assess the accuracy of the IDS, i.e., how many instances classified as positive (attack) are in fact positive.

Mathematically, precision is calculated using the following formula:

$$Precision = \frac{\text{TruePositives(TP)}}{\text{TruePositives (TP) +FalsePositives(FP)}} \quad (5.7)$$

where:

True Positives (TP) is the number of attack instances that were correctly identified as attacks.

False Positives (FP) is the number of normal instances that were incorrectly identified as attacks.

From FIG. 5.4 it can be observed that the Proposed RNN-IDS achieved a precision score of 0.95 which is a way better than other compared models during Binary Classification. This highlights the system's capacity to reduce false positives and provides the percentage of true positive predictions among all positive predictions.

Recall. Recall, also known as Sensitivity or True Positive Rate, is a crucial evaluation metric, revealing the model's ability to correctly identify and classify positive (attack) instances. It resolves the question: "How many true positive instances did the model successfully identify as being positive?".

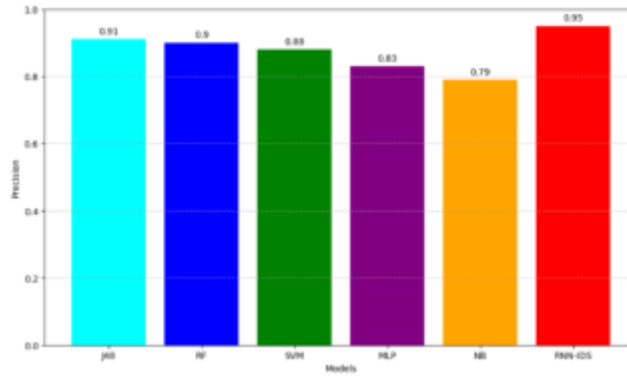


Fig. 5.4: Comparison of Precision between Proposed(RNN-IDS) and other(j48,RF,SVM,MLP,NB) methods applied on NSL-KDD dataset.

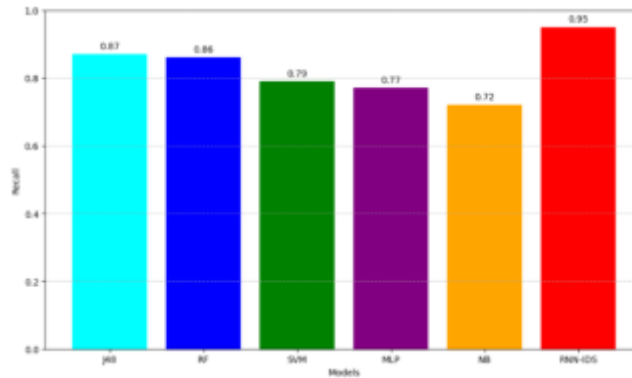


Fig. 5.5: Comparison of Recall between Proposed(RNN-IDS) and other(j48,RF,SVM,MLP,NB) methods applied on NSL-KDD dataset.

Mathematically, recall is computed as follows:

$$Recall = \frac{\text{True Positives(TP)}}{\text{True Positives (TP) + False Negatives(FN)}} \quad (5.8)$$

where:

- True Positives (TP): Represent the instances which were attacks and were correctly identified as attacks by the IDS.
- False Negatives (FN): Represent the instances which were attacks but were incorrectly identified as normal by the IDS.

Proposed RNN-IDS exhibited a recall score of 0.95, which represents the ratio of true positive predictions to all actual positive instances, underlining the system's capacity to identify actual intrusions effectively. It can be clearly depicted from the FIG. 5.5 that Proposed RNN-IDS surpassed all the other compared models in terms of Recall.

Score. The F1 Score is the harmonic mean of precision and recall, and it offers a balance between the two factors whenever there is an imbalance in the class distribution. It takes into consideration both false positives and false negatives, and it is especially useful in circumstances in which one form of error is more substantial than the other.

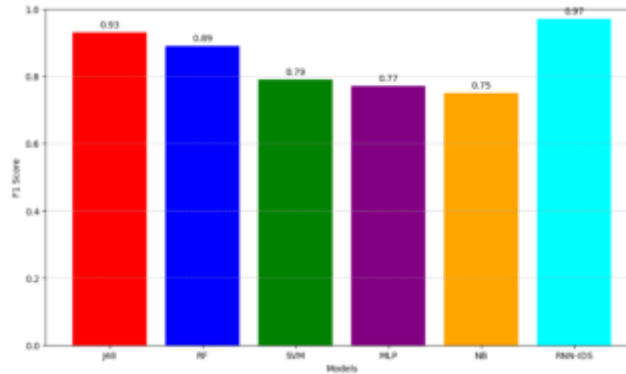


Fig. 5.6: Comparison of F1 Score between Proposed(RNN-IDS) and other(j48,RF,SVM,MLP,NB) methods applied on NSL-KDD dataset.

Mathematically, the F1 Score is defined as:

$$F1 - Score = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5.9)$$

where:

Precision (Positive Predictive Value) is defined as:

$$\text{Precision} = \frac{\text{True Positives(TP)}}{\text{True Positives(TP)} + \text{False Positives(FP)}}$$

Recall (Sensitivity or True Positive Rate) is defined as:

$$\text{Recall} = \frac{\text{True Positives(TP)}}{\text{True Positives(TP)} + \text{False Negatives(FN)}}$$

The F1 score, harmonizing precision and recall, for proposed RNN-IDS was 0.97. FIG. 5.6 reflects the model's performance in binary classification in terms of F1-score as compared to other models.

5.3.2. Multi Class Classification. The dataset was organized using the multiclass classification paradigm to categorize network activity into five different labels: "Normal" and four classes of attacks (Dos, Probe, R2L, and U2R). The RNN model was put through a rigorous training program before being put to the test to see how well it could classify data across these many classifications.

Accuracy. In a multi-class classification problem with more than two classes, like four types of attacks and one normal label, the formulation might get a little more complicated because we have to figure out the True Positives and True Negatives for each class separately and then add them all up. If this is the case and C is the number of classes, the accuracy may be represented as:

$$\text{Accuracy} = \frac{\sum_{i=1}^C \text{TP}_i + \text{TN}_i}{\text{Total Instances}} \quad (5.10)$$

TP_i and TN_i refer to the True Positives and True Negatives for the i th class respectively.

Proposed RNN-IDS recorded an accuracy of 95%. It can be observed in FIG. 5.7, when compared to the machine learning models, the RNN's performance was superior, indicating its adeptness in correctly classifying instances.

Precision. When dealing with multiple attack types (classes) in an IDS for IoT multi-class classification scenario, the precision for each class is calculated separately and then the macro-average precision is derived across all classes. This gives a general idea of the IDS model's precision across various attack types.

In a multi-class context, the following formula can be modified to compute class-wise precision as:

$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \quad (5.11)$$

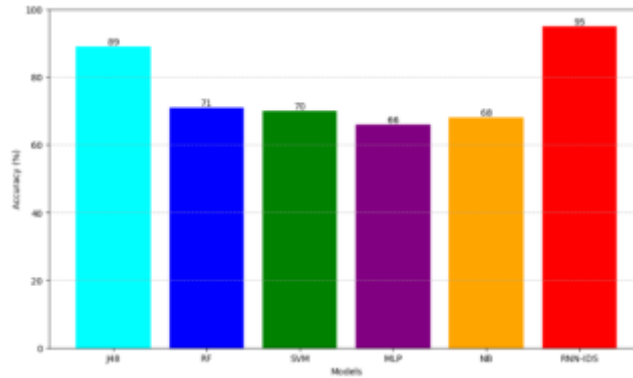


Fig. 5.7: Comparison of Accuracy between Proposed(RNN-IDS) and other(j48,RF,SVM,MLP,NB) methods applied on NSL-KDD dataset for Multiclass Classification.

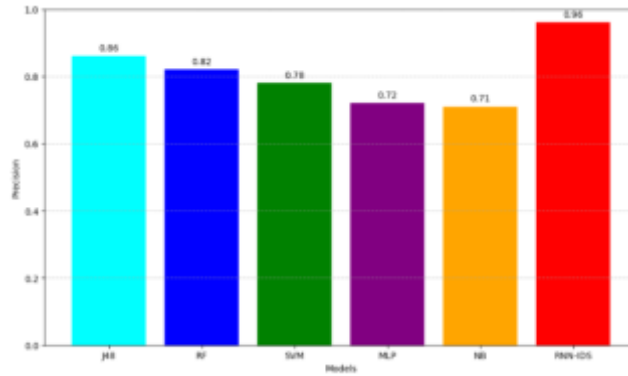


Fig. 5.8: Comparison of Precision between Proposed(RNN-IDS) and other(j48,RF,SVM,MLP,NB) methods applied on NSL-KDD dataset for Multiclass Classification.

where:

- Precision_i is the precision for the ith class (type of attack).
- TP_i and FP_i represent the True Positives and False Positives for the ith class respectively.

After calculating the precision for each class, the macro-average precision across all classes is computed as follows:

$$Macro - AveragePrecision = \frac{\sum_{i=1}^C Precision_i}{C} \quad (5.12)$$

Here C is the total number of classes.

Precision indicates the system's capacity to reduce false alarms, which is essential for IoT IDS usability and reliability. This metric is examined alongside recall and F1 score to evaluate the proposed model.

Precision is vital as it tells us about the model's capability to correctly identify positive instances. With a precision score of 0.96, the RNN-IDS model edged out most ML-based models, showcasing its reliability in positive identifications in FIG. 5.8.

Recall. In multi-class classification for IDS in IoT, there are several sorts of attacks (classes), recall has been computed for each class and then the macro-average recall is calculated over all classes to provide a generalized model recall measure. Class-wise recall for multi-class classification can be computed as:

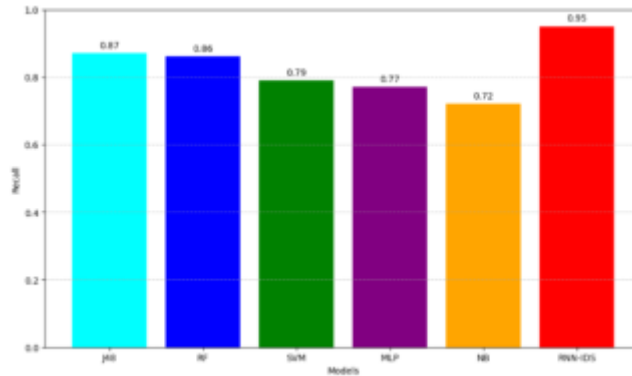


Fig. 5.9: Comparison of Recall between Proposed(RNN-IDS) and other(j48,RF,SVM,MLP,NB) methods applied on NSL-KDD dataset for Multiclass Classification.

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{5.13}$$

where:

- Recall_i is the recall for the ith class (type of attack).
- TP_i and FN_i are the True Positives and False Negatives for the ith class, respectively. Macro-average recall across all classes in a multi-class classification scenario can be derived as follows:

$$Macro - Average Recall = \frac{\sum_{i=1}^C Recall_i}{C} \tag{5.14}$$

where C is the total number of classes.

IDS for IoT relies on recall because missing an attack can be disastrous. This comprehensive evaluation helps fine-tune the model for reliable IoT IDS.

Recall focuses on the model’s ability to identify all potential positive instances. The RNN-IDS model achieved a commendable recall score of 0.95, which was notably higher than some ML models as can be seen in Fig. 5.9, emphasizing its proficiency in identifying actual attack instances.

F1 Score. In a multiclass classification scenario, such as categorizing various types of network intrusions in IDS for IoT, the F1 Score can be calculated for each class separately, and then an average can be calculated to evaluate the classifier’s overall performance. Micro and macro F1 Scores are two ways to calculate the average F1 Score for multiclass classification problems.

Micro F1 Score: Calculated by aggregating the contributions of all classes to find the average.

$$F_{1micro} = 2x \frac{\sum_{i=1}^C TP_i}{\sum_{i=1}^C TP_i + \sum_{i=1}^C FP_i + \sum_{i=1}^C FN_i} \tag{5.15}$$

where C represents the number of classes, and TP_i,FP_i, and FN_i denote the true positives, false positives, and false negatives for the i-th class, respectively.

Macro F1 Score: The arithmetic mean of the per-class F1 Scores.

$$F_{1macro} = \frac{1}{C} \sum_{i=1}^C F1_i \tag{5.16}$$

where F1_i represents the F1 Score for the i-th class.

The F1-Score serves as a balanced measure, taking into account both precision and recall. Proposed RNN-IDS model’s score of 0.94 was demonstrably superior, revealing its balanced performance in precision and sensitivity,as can be depicted in FIG. 5.10.

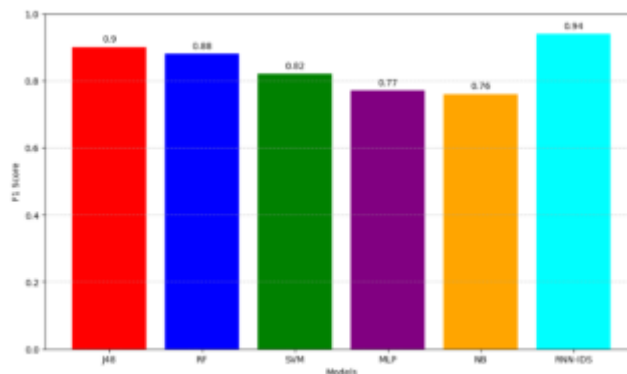


Fig. 5.10: Comparison of F1 Score between Proposed(RNN-IDS) and other(j48,RF,SVM,MLP,NB) methods applied on NSL-KDD dataset for Multiclass Classification.

6. Conclusion. The results of experiment demonstrate the effectiveness of proposed RNN-IDS for intrusion detection in both binary and multiclass classification scenarios. In binary classification, proposed RNN-IDS system excelled with high accuracy, precision, recall, and F1 score, indicating its ability to accurately identify both normal and intrusive network activities while minimizing false alarms. Furthermore, in the multiclass classification setting, proposed RNN-IDS showcased its adaptability by accurately classifying various intrusion types. This capability is crucial for network administrators and security professionals, as it enables them to pinpoint specific attack categories for prompt mitigation.

Comparing proposed RNN-IDS with renowned machine learning models: J48, RF (Random Forest), SVM (Support Vector Machine), MLP (Multi-Layer Perceptron), and NB (Naive Bayes), deduced that it consistently outperformed them in terms of accuracy, precision, recall, and F1 score. This suggests that the use of recurrent neural networks offers substantial advantages over conventional techniques when it comes to intrusion detection on the NSL-KDD dataset.

Proposed research proves the potential of RNN-based IDS systems in enhancing network security. The results indicate that proposed RNN-IDS is a promising approach for accurately detecting network intrusions, and its superior performance over traditional models makes it a valuable asset for real-world cyber security applications. While the current research demonstrates the effectiveness of RNN based intrusion detection systems, certain limitations highlight avenues for future exploration. Expanding the research to include larger and more diverse datasets may improve the model's resilience and generalization. In addition, the current work focuses on simulated environments, therefore deploying the model in real-world IoT settings would provide useful insights into its practical efficacy.

REFERENCES

- [1] N. Chaabouni, M. Mosbah, A. Zemmari, C. Sauvignac, and P. Faruki, "Network intrusion detection for IoT security based on learning techniques," *IEEE Communications Surveys & Tutorials*, 21(3), (2019), pp. 2671-2701.
- [2] S. Bajpai and K.S.B.K. Chaurasia, "Intrusion detection system in IoT network using ML," *NeuroQuantology*, 20(13), (2022), pp. 3597.
- [3] A. Sinha, G. Shrivastava, and P. Kumar, "Architecting user-centric internet of things for smart agriculture," *Sustainable Computing: Informatics and Systems*, 23, (2019), pp. 88-102.
- [4] M.J.S. Aneja, T. Bhatia, G. Sharma, and G. Shrivastava, "Artificial intelligence based intrusion detection system to detect flooding attack in VANETs," In *Handbook of Research on Network Forensics and Analysis Techniques*, IGI Global, (2018), pp. 87-100.
- [5] Y. Otoum and A. Nayak, "As-ids: Anomaly and signature based ids for the internet of things," *Journal of Network and Systems Management*, 29(3), (2021), pp. 23.
- [6] D. Musleh, M. Alotaibi, F. Alhaidari, A. Rahman, and R.M. Mohammad, "Intrusion detection system using feature extraction with machine learning algorithms in IoT," *Journal of Sensor and Actuator Networks*, 12(2), (2023), pp. 29.
- [7] H. Sharma, P. Kumar, and K. Sharma, "Identification of Device Type Using Transformers in Heterogeneous Internet of

- Things Traffic," International Conference On Innovative Computing And Communication, Singapore: Springer Nature Singapore.,(2023), pp. 471-481.
- [8] R.R. Reddy, Y. Ramadevi, and K.N. Sunitha, "Effective discriminant function for intrusion detection using SVM," 2016 International conference on advances in computing, communications and informatics (ICACCI), (2016), pp. 1148-1153. IEEE.
- [9] W. Li, P. Yi, Y. Wu, L. Pan, and J. Li, "A new intrusion detection system based on KNN classification algorithm in wireless sensor network," *Journal of Electrical and Computer Engineering*, (2014).
- [10] B. Ingre and A. Yadav, "Performance analysis of NSL-KDD dataset using ANN," 2015 international conference on signal processing and communication engineering systems, (2015), pp. 92-96. IEEE.
- [11] N. Farnaaz and M.A. Jabbar, "Random forest modeling for network intrusion detection system," *Procedia Computer Science*, 89, (2016), pp. 213-217.
- [12] J. Zhang, M. Zulkernine, and A. Haque, "Random-forests-based network intrusion detection systems," *IEEE Trans. Syst., Man, Cybern. C, Appl.Rev.*, vol. 38, no. 5, (2008), pp. 649-659.
- [13] J.A. Khan and N. Jain, "A survey on intrusion detection systems and classification techniques," *Int. J. Sci. Res. Sci., Eng. Technol.*, vol. 2, no. 5, (2016), pp. 202-208.
- [14] S.A. Bini, "Artificial intelligence, machine learning, deep learning, and cognitive computing: what do these terms mean and how will they impact health care?," *The Journal of arthroplasty*, 33(8), (2018), pp. 2358-2361.
- [15] X. Wang and X. Lu, "A host-based anomaly detection framework using XGBoost and LSTM for IoT devices," *Wireless Communications and Mobile Computing*, (2020), pp. 1-13.
- [16] V.V. Kumari and P.R.K. Varma, "A semi-supervised intrusion detection system using active learning SVM and fuzzy c-means clustering," 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), (2017), pp. 481-485. IEEE.
- [17] Y. Otoum, D. Liu, and A. Nayak, "DLIDS: a deep learningbased intrusion detection framework for securing IoT," *Transactions on Emerging Telecommunications Technologies*, 33(3), (2022), p. e3803.
- [18] Y. Yang, L. Wu, G. Yin, L. Li, and H. Zhao, "A survey on security and privacy issues in Internet-of-Things," *IEEE Internet of things Journal*, 4(5), (2017), pp. 1250-1258.
- [19] A. Dawoud, S. Shahristani, C. Raun, "Deep learning and software-defined networks: towards secure iot architecture," *In. Things.*, 34, (2018), pp. 82-89.
- [20] M. Khan, M. Khatkhatk, S. Latif, A. Shah, M. Ur Rehman, W. Boulila, et al., "Voting classifier-based intrusion detection for IoT networks," in *Advances on Smart and Soft Computing*, Springer, (2022), pp. 313328.
- [21] N. Naz, M. Khan, S. Alsuhibany, M. Diyan, Z. Tan, M. Khan, et al., "Ensemble learning-based IDS for sensors telemetry data in IoT networks," *Math. Biosci. Eng.*, 19, (2022), pp. 1055010580.
- [22] B.S. Bhati, G. Chugh, F. AlTurjman, and N.S. Bhati, "An improved ensemble based intrusion detection technique using XGBoost," *Transactions on emerging telecommunications technologies*, 32(6), (2021), pp. e4076.
- [23] A. Arko, S. Khan, A. Preeti, M. Biswas, "Anomaly Detection In IoT Using Machine Learning Algorithms," Brac University, (2019).
- [24] P. Illy, G. Kaddoum, C.M. Moreira, K. Kaur, S. Garg, "Securing fog-to-things environment using intrusion detection system based on ensemble learning," (2019), pp. 1518.
- [25] A. Verma, V. Ranga, "Machine Learning intrusion detection systems for IoT applications," *Wireless Pers. Commun.*, 111, (2020), pp. 22872310.
- [26] C. Yin, Y. Zhu, J. Fei, and X. He, "A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks," *IEEE Access*, 5, (2017).
- [27] Noor Wali Khan, Mohammed S. Alshehri, Muazzam A. Khan, Sultan Almakdi, Naghmeh Moradpoor, Abdulwahab Alazeb, Safi Ullah, Naila Naz, and Jawad Ahmad, "A hybrid deep learning-based intrusion detection system for IoT networks," *Mathematical Biosciences and Engineering*, 20(8), (2023), pp. 13491-13520.
- [28] R. Zhao, "NSL-KDD," *IEEE Dataport*, (2022).
- [29] A. Sinha, P. Kumar, N.P. Rana, R. Islam, and Y.K. Dwivedi, "Impact of internet of things (IoT) in disaster management: a task-technology fit perspective," *Annals of Operations Research*, 283, (2019), pp. 759-794.
- [30] A. Arora, S.K. Yadav, and K. Sharma, "Denial-of-service (dos) attack and botnet: Network analysis, research tactics, and mitigation," In *Research Anthology on Combating Denial-of-Service Attacks*, (2021), pp. 49-73. IGI Global.
- [31] S.K. Yadav, K. Sharma, and A. Arora, "Security integration in ddos attack mitigation using access control lists," *International Journal of Information System Modeling and Design (IJISMD)*, 9(1), (2018), pp. 56-76.

Edited by: Himani Bansal

Special issue on: Recent Advance Secure Solutions for Network in Scalable Computing

Received: Dec 8, 2024

Accepted: Apr 28, 2024



DIFFCRNN: A NOVEL APPROACH FOR DETECTING SOUND EVENTS IN SMART HOME SYSTEMS USING DIFFUSION-BASED CONVOLUTIONAL RECURRENT NEURAL NETWORK

MARYAM M. AL DABEL *

Abstract. This paper presents a latent diffusion model and convolutional recurrent neural network for detecting sound event, fusing advantages of different networks together to advance security applications and smart home systems. The proposed approach underwent initial training using extensive datasets and subsequently applied transfer learning to adapt to the desired task to effectively mitigate the challenge of limited data availability. It employs the latent diffusion model to get a discrete representation that is compressed from the mel-spectrogram of audio. Subsequently a convolutional neural network (CNN) is linked as the front-end of recurrent neural network (RNN) which produces a feature map. After that, an attention module predicts attention maps in temporal-spectral dimensions level, from the feature map. The input spectrogram is subsequently multiplied with the generated attention maps for adaptive feature refinement. Finally, trainable scalar weights aggregate the fine-tuned features from the back-end RNN. The experimental findings show that the proposed method performs better compared to the state-of-art using three datasets: the DCASE2016-SED, DCASE2017-SED and URBAN-SED. In experiments on the first dataset, DCASE2016-SED, the performance of the approach reached a peak in $F1$ of 66.2% and ER of 0.42. Using the second dataset, DCASE2017-SED, the results indicate that the $F1$ and ER achieved 68.1% and 0.40, respectively. Further investigation with the third dataset, URBAN-SED, demonstrates that our proposed approach significantly outperforms existing alternatives as 74.3% and 0.44 for the $F1$ and ER .

Key words: Sound event detection, latent diffusion model, spectrogram, deep neural network.

1. Introduction. The objective of sound event (SE) detection is to provide devices with the capability to identify and classify acoustic environments. It can be characterized as the process of discerning the presence of both overlapping and non-overlapping sound events, as well as determining their respective initiation and duration intervals [44]. A distinct auditory occurrence that may be recognized as a distinct notion is referred to as a sound event [18]. In our everyday lives, we often encounter many forms of sound events as an example bird cries, dog barking, and human speech. In a real-world acoustic environment, the occurrence of these sound events may not be sequential but rather exhibit a tendency to regularly overlap. The SE detection systems may enhance the capabilities of current security applications, smart home systems and surveillance systems when applied jointly. In addition, they can be used in industrial environments to detect deficiencies in equipment and machinery.

Different approaches have been used to perform the SE detection task. There are two fundamentals to boost the overall classification performance of SE models: i) the extraction of acoustic features with robust characterization abilities, and ii) efficient classification techniques. The widely used features are linear predictive coding [32], linear predictive cepstral coefficients, discrete wavelet transform, mel frequency cepstral coefficients [32] and log-mel spectrograms. Turning to conventional classifiers, examples include support vector machines [13], Gaussian mixture models [15], hidden Markov models [11], multi-layer perceptron [42]. Such conventional models, however, are only useful to single acoustic events and small datasets [31]. These conventional classification models are less likely to satisfy the classification needs due to the large dataset size and audio complexity. The advances of machine learning has made it possible for neural network classification models to outperform more conventional classifiers, such as feedforward neural networks, recurrent neural networks [36], convolutional neural networks [22] and convolutional recurrent neural networks [2, 12, 21, 29]. Most SE research in recent years has employed deep learning-based classification models [4, 1]. While neural network-

*Department of Computer Science and Engineering, College of Computer Science and Engineering, University of Hafr Al Batin, Saudi Arabia (maldabel@uhb.edu.sa).

based classification models have been widely used in the field of acoustics, difficulties with sound detection still exist include the following: i) the SE model has more parameters, more feature space dimensions, and larger datasets; ii) the temporal-frequency structure of sounds is very complex and may be continuous, abrupt, or periodic; and iii) inconsistency and ambiguous duration of sounds impact model classification performance. The main contributions of this paper are summarized as follows.

- Instead of choosing a random combination, as in earlier efforts, we take into account the pressure levels of the audio pairings when mixing them for data augmentation by applying the Latent Diffusion Model. This makes sure that the combined audio accurately represents both of the source audio.
- Combining the convolutional recurrent neural networks and an attention module in a unified framework that connect both the convolutional neural network layer and recurrent neural network layer.
- Conducting a series of comparative experiments to evaluate the performance of the proposed models.

The rest of this paper is set up as follows. Section 2 discusses and reviews previous related work. Section 3 introduces the proposed framework. Sections 4, 5 and 6 report and analyze the experimental results. Section 8 summarizes the work.

2. Related Work. Early work in SE detection typically aims at identifying only the dominating sound event among the overlapping sound events and their associated onset-offset periods. However, this strategy is less appropriate for applications that need the simultaneous detection of several sound events.

Widely known classifiers were used for such task including the combined Gaussian mixture model-hidden Markov model [16], non-negative matrix factorization [17], convolutional neural networks [48, 38], and recurrent neural networks [37, 47] networks. In [16], for instance, the combined Gaussian mixture model-hidden Markov model was employed to detect the overlapping sound events based on multiple restricted Viterbi passes. Whereas in [17], the combined Gaussian mixture model-hidden Markov model was designed to better identified the overlapping sound events by a preprocessing stage, in which a non-negative matrix factorization method was implemented as a stage to get multiple streams of source separated audio.

As deep learning methods advanced, many deep neural network-based solutions for the SE challenges were proposed. A multi-class multi-label feed-forward deep neural networks was applied in [6] such that each input frame was produced by concatenating multiple temporal-frames of the feature. This technique outperformed the best SE technique previously reported in [17]. Individual Gaussian mixture models are trained for each sound class when using generative classifiers like Gaussian mixture model. The sound class is determined during inference based on the greatest probable outcomes of the Gaussian mixture model. In [5], for each sound class in the dataset, several feed-forward deep neural networks classifiers were similarly trained. The cumulative outcomes of the various single-class feed-forward deep neural networks classifiers were used for the SE task during inference. The findings indicated that the multiple single-class technique performed slightly poor when compared to the multi-class multiple label approach.

Recently, in an attempt to enhance classification performance, a study based on the attention mechanism has also been conducted in the area of SE research. For instance, the Convolutional Long Short-Term Memory and Deep Neural Networks model incorporates the temporal attention mechanism that was first presented in [14]. The system can look at every time step and try to identify the high impact one so that it can be given more weight. Another model was suggested in [27] using an attention-based multi-stream network model. The attention weight is calculated based on the degree of energy change in the spectrogram. The authors in [49] noted that not all frame-level characteristics can affect environmental sound performance equally. In particular, there are other time frames, such as silent frames, noisy frames, can cause the robustness of the classification model to degrade and will also result in errors in the classification. Based on this assumption, It is crucial to record the primary temporal segment of the sound stream. While the aforementioned techniques do help with classification performance, they did not take into account the variation of the frequency bands and their effect on the process. In addition, the method in [46] was developed to stack multiple attention network to get robust features. A temporal attention mechanism was suggested in [28] for convolutional layers to boost the representative ability of convolutional neural networks by re-weighting the convolutional neural networks feature maps using dot-product operation along the time dimension from input spectrogram.

Deep learning models have the ability to acquire effective representations from raw data without the need for manual intervention. Convolutional neural networks (CNNs) can automatically extract feature maps through

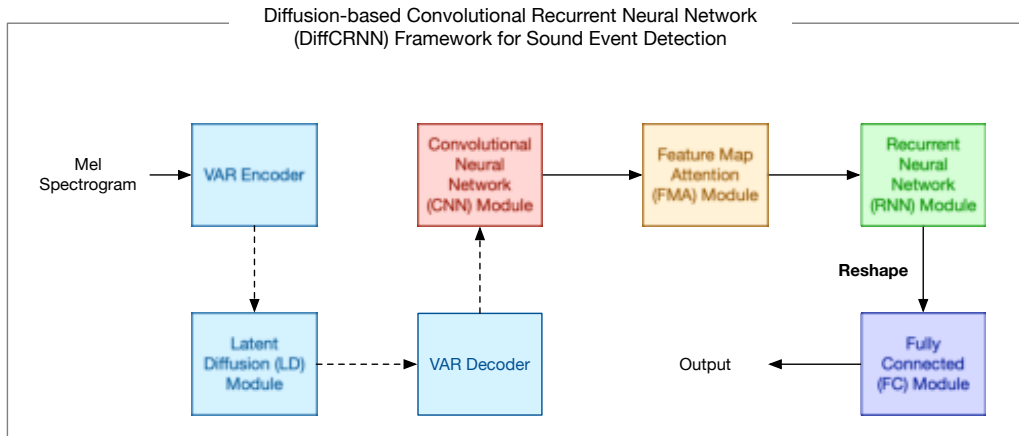


Fig. 3.1: The architecture of the proposed Diffusion-based Convolutional Recurrent Neural Network (DiffCRNN) system.

the convolution process, enabling them to capture the spatial features of input data [22, 39]. Furthermore, weight sharing significantly reduces the number of parameters in a convolutional neural network (CNN), thereby facilitating the training process of a CNN model compared to an equivalent dense neural network. Nevertheless, CNN-based models encounter challenges in capturing temporal dependencies when the input consists of time series data [20, 3]. Recurrent neural networks (RNNs) are extensively employed in various tasks, including text classification and speech recognition [10]. However, RNN-based models are limited in their ability to effectively extract features from raw data and face challenges with gradient vanishing and exploding when processing long time sequences [4]. Thus, this paper utilizes deep convolutional recurrent neural networks (namely DiffCRNN) to detect DiffCRNN by combining CNNs and RNNs. The DiffCRNN model utilizes convolutional layers to extract spatial features from raw data, while the recurrent layers are responsible for capturing the sequence information.

3. DiffCRNN: Framework Design. The architecture of the proposed Diffusion-based Convolutional Recurrent Neural Network (DiffCRNN) framework is illustrated in Figure 3.1. The framework has five main modules which are the latent diffusion based module, the convolutional neural networks (CNN) based module, the feature map attention based module, the recurrent neural network (RNN) based module and, finally, the fully connected layer based module.

In particular, the latent diffusion based module has three primary sub-modules: (i) encoder, (ii) latent diffusion model, and (iii) audio variational auto-encoder. The encoder is responsible for encoding the input description of the audio. Next, the process of reverse diffusion is used to construct a latent representation of the audio or audio prior from Gaussian noise, utilizing the textual representation. The audio variational auto-encoder subsequently employs the latent audio representation to yield a mel-spectrogram. The primary objective of the CNN is to extract a multi-dimensional and higher-order features from the input spectrogram. Further, the FM-attention module learns the importance of each dimensions in a dynamic way, in which important feature map information is extracted and unimportant dimensions are discounted. The RNN module then attempts to acquire contextual information and anticipate both the start and offset times of sound events in a precise way. Finally, the output characteristics of the RNN serve as the input for the fully connected layer in order to get the classification score of the DiffCRNN system.

This section described the architecture in more detail. The latent diffusion based module is described in Section 3.1. The CNN module is reviewed in Section 3.2. Then, Section 3.3 explains the feature map attention based module. Finally, in Section 3.4, the RNN module is represented.

3.1. Latent Diffusion Based Module. The latent diffusion based module (LD) consists of three primary parts: the encoder, latent diffusion model, and audio variational auto-encoder.

3.1.1. The encoder sub-module:. The encoder (E_τ) is the pre-trained large language models using FLAN-T5 [9] to obtain text encoding τ . The token count and token-embedding size are L and d_τ , respectively. The use of gradient descent, which emulates the process of imitating characteristics, is of significant importance in the task of learning the relationship between textual and auditory concepts, without the need for fine-tuning the E_τ , by treating each input sample as a distinct job. Enhanced pretraining techniques have the potential to enable the E_τ , however, to prioritize essential information with less interference and enhanced contextual understanding. Therefore, the E_τ is held constant, on the assumption that the reverse diffusion process may acquire knowledge of the audio inter-modality mapping prior to its generation.

3.1.2. The latent diffusion sub-module. The purpose of this sub-module is motivated by [40, 30] with the aim to produce the audio prior s_0 under the direction of text encoding τ . This basically comes down to parameterized $p_0(s_0|\tau)$ via approximating the correct prior $q(s_0|\tau)$.

The mechanisms of forward and reverse diffusion allow the sub-module to accomplish the aforementioned. The forward diffusion consists of a series of Markov of Gaussians with predetermined noise parameters $0 < \delta_1 < \delta_2 < \dots < \delta_N < 1$ to get more distorted iterations of the samples, s_0 as follows;

$$q(s_n|s_{n-1}) = \mathcal{N}(\sqrt{1 - \delta_n}s_{n-1}, \delta_n\mathbf{I}), \quad (3.1)$$

$$q(s_n|s_0) = \mathcal{N}(\sqrt{\bar{\kappa}_n}s_0, (1 - \bar{\kappa}_n)\mathbf{I}), \quad (3.2)$$

such that N denotes the quantity of forward diffusion iterations, $\kappa_n = 1 - \delta_n$, and $\bar{\kappa}_n = \prod_{i=1}^n \kappa_i$.

A more direct sampling of s_n from sample noisier versions can be applied through re-parametrization using as follows;

$$s_n = \sqrt{\bar{\kappa}_n}s_0 + (1 - \bar{\kappa}_n)\epsilon, \quad (3.3)$$

such that the noise sample $\epsilon \in \mathcal{N}(\mathbf{0}, \mathbf{I})$. The last stage of the forward procedure yields $s_N \in \mathcal{N}(\mathbf{0}, \mathbf{I})$.

The reverse method uses noise estimation ($\hat{\epsilon}_\theta$) to denoise and recover s_0 using loss function as follows;

$$\Omega = \sum_{n=1}^N \lambda_n \mathbb{E}_{\epsilon_n \in \mathcal{N}(\mathbf{0}, \mathbf{I})} \|\epsilon_n - \hat{\epsilon}_\theta^{(n)}(s_n, \tau)\|_2^2. \quad (3.4)$$

such that s_n is sampled from equation. 3.3 based on ϵ_n and λ_n which are the standard normal noise and the weight of reverse step n , respectively. The n is considered to be a measure of signal-to-noise ratio in respect to $\kappa_{1:N}$.

3.1.3. The augmentation sub-module. In this sub-module, we synthesis more text-audio pairings by superimposing existing audio pairs and concatenating their captions. To avoid overpowering low-pressure samples, the pressure level of audio R is considered. Audio sample (x_1) weight is determined as a relative pressure level:

$$p = (1 + 10^{\frac{R_1 - R_2}{20}})^{-1}, \quad (3.5)$$

such that R_1 and R_2 denotes the pressure levels of two used audio samples y_1 and y_2 . This guarantees accurate depiction of the two audio samples after mixing.

In addition, the square of a sound wave's amplitude determines how much energy it has [45]. As a results, y_1 and y_2 were mixed as follows;

$$\text{mix}(y_1, y_2) = \frac{py_1 + (1 - p)y_2}{\sqrt{p^2 + (1 - p)^2}}. \quad (3.6)$$

3.1.4. The free guidance sub-module. This sub-module is a classifier-free in which the input τ is used to rebuild the s_0 by directing the reverse diffusion. The contribution of text guidance to the noise level $\hat{\epsilon}_\theta$ is managed by a guidance scale v with respect to unguided estimation throughout inference:

$$\hat{\epsilon}_\theta^{(n)}(s_n, \tau) = v\epsilon_\theta^{(n)}(s_n, \tau) + (1 - v)\epsilon_\theta^{(n)}(s_n). \quad (3.7)$$

3.1.5. The decoder sub-module. In this sub-module, we implement the audio variational auto-encoder to convert the mel-spectrogram of an audio sample into a s_0 . The latent diffusion sub-module re-builds the \hat{s}_0 based on the τ . The encoder and decoder are formulated of ResUNet blocks and are trained via maximizing evidence lower-bound and minimizing adversarial loss [25].

3.2. CNN Module. Assuming that \mathbf{h}^{n-1} is the feature map of size $C^{n-1} \times P^{n-1} \times Q^{n-1}$ from the $(n-1)$ -th layer, such that C^{n-1} is the channel number and $P^{n-1} \times Q^{n-1}$ is the size of the feature map at the time and frequency axes, the result of the n -th convolutional layer is defined as

$$\mathbf{h}_j^n = \sum_{i=1}^{C^{n-1}} \mathbf{w}_{ij}^n * \mathbf{h}_i^{n-1} + b_j^n, \quad (3.8)$$

where \mathbf{h}_j^n denotes the j -th channel of \mathbf{h}^n , \mathbf{w}_{ij} denotes the (i, j) -th convolutional kernel, $*$ is the convolutional operation, and b_j^n represents the bias at the j -th channel. In order to accelerate convergence, convolutional layers are typically followed by batch normalization and a ReLU activation function. Batch normalization can also increase the stability of CNN [8].

In order for the CNN model to function properly, the three-dimensional feature map that includes the channel, time frame, and feature vector must be transformed into a classification vector. It is possible, as mentioned in the previous section, to immediately flatten the feature map into a vector in order to reduce the number of dimensions. Flattening, on the other hand, could result in a sub-optimization due to the fact that it might preserve duplicate information. As a result, the time-frequency attention pooling will be covered here to produce a vector that is more compact and has less information that is redundant than the one that is generated by flattening.

The temporal-frequency global attention (TFGA) pooling in CNNs decreases the dimensionality of a feature map through measuring the contribution of each temporal-frequency unit. It is composed of two sub-modules: an attention sub-module, and a classification sub-module, which come typically after a set of convolutional layers and local average pooling layers. The attention sub-module has a two-dimensional convolutional layer with an output channel number equal to the number of classes K , and a kernel size of 1×1 , which results in an attention tensor A . An activation function (softmax or sigmoid) is applied after the convolutional layer to yield a tensor A^* with values in the range $[0, 1]$. Next, the tensor A^* is normalized using

$$P_{kpq} = \frac{A_{kpq}^*}{\sum_{p=1}^{P_w} \sum_{q=1}^{Q_w} A_{kpq}^*}, \quad (3.9)$$

such that P denotes the probability tensor. Moving to the classification sub-module, the feature map is transformed into a new one C with the channel number of K using an additional two-dimensional convolutional layer with a kernel size of 1×1 . After that, the resultant classification tensor C is multiplied by P to determine the probability of each class by applying the following

$$p_k = \sum_{p=1}^{P_w} \sum_{q=1}^{Q_w} C_{kpq} \odot P_{kpq}, \quad (3.10)$$

Additionally, to complete a classification task, a softmax or log-softmax function is employed to operate on C or p . In order to make more accurate predictions, the time-frequency attention pooling can assess the contribution of each time-frequency bin to classification [20].

3.3. Feature-map Attention Module. In the Feature Map (FM-attention) algorithm, the multi-dimensional feature map \mathbf{h} is acquired from CNN module, such that C is the channel number and $T \times F$ represents at the time and frequency axes the size of the feature map. Then the high-order feature map \mathbf{h} was input to the FM-attention model. The FM-attention has a Sigmoid activation layer and fully connected feedforward layer in order to compute the high impact weight of each feature dimension of \mathbf{h} of size $C \times T \times F$. The high impact weight U is the outputs of the Sigmoid layer, which is assigned to different feature dimensions. First, \mathbf{h} is permuted into 3-dimensional tensor \mathbf{h}' of size $T \times C \times F$. Subsequently, \mathbf{h}' is flattened as a 2-dimensional tensor \mathbf{h}'' by fixing the dimension T .

Next, the input to the feedforward layer is \mathbf{h}'' . The number of hidden units in this layer is set to CF . The dimension of weights U is $M = CF$, which can be written as:

$$U = \{U_1, U_2, \dots, U_d, \dots, U_M\}, \quad (3.11)$$

where U_m influences the m th dimensional feature of \mathbf{h}'' , the expression of U_m is:

$$U_m = \frac{\exp(O_m)}{\sum_{j=1}^{j=m} \exp(O_j)}, \quad (3.12)$$

The dimension of \mathbf{h}'' is M . The j th dimensional output of the Sigmoid activation layer is O_j . The high impact weight U is repeated T times, and its dimension U results in $T \times C \times F$. The U is reshaped to form U' , FM-attention vector, of size $T \times C \times F$. The outputs of the FM-attention module can be written as:

$$\mathbf{h}_{att} = U' \odot \mathbf{h}', \quad (3.13)$$

where “ \odot ” denotes the Hadamard product. Also, the outputs \mathbf{h}_{att} of FM-attention module are fed into the RNN module.

3.4. RNN Module. The hidden state h_t at the time step t , $t = 1, \dots, T$, can be represented as

$$h_t = \sigma_h(\mathbf{w}_h x_t + \mathbf{u}_h h_{t-1} + b_h), \quad (3.14)$$

such that \mathbf{w}_h and \mathbf{u}_h denote the weights, T represents the total number of time steps, b_h denotes the bias, h_{t-1} represents the previous hidden state at the time step $t - 1$, x_t denotes the input vector at the time step t , and σ_h represents an activation function. In classification tasks, the final recurrent layer’s hidden states are often merged into a single vector and sent on to a fully connected layer. Typically, a vector can be generated as the fully connected layer’s input by either computing the average of the hidden states or extracting the hidden state at the most recent time step.

This simple RNN, however, is unable to process long-term context information owing to the exploding and vanishing gradient problem. For this reason, the Long Short-Term Memory (LSTM) RNN structure [19] and Gated Recurrent Units (GRU) RNN structure [50] were suggested to address such problem. The neurons in the simple RNN model is changed to memory blocks in the LSTM-RNN model, such that the memory blocks are connected recurrently. The LSTM, [19], is employed by replacing Equation 3.14 with the following steps: At the t -th time step, an LSTM unit comprises of an input gate i_t , an output gate o_t , a forget gate f_t , and a cell state c_t . The procedure of an LSTM unit is implemented as follow;

$$i_t = \sigma(\mathbf{w}_i x_t + \mathbf{u}_i h_{t-1} + b_i), \quad (3.15)$$

$$f_t = \sigma(\mathbf{w}_f x_t + \mathbf{u}_f h_{t-1} + b_f), \quad (3.16)$$

$$o_t = \sigma(\mathbf{w}_o x_t + \mathbf{u}_o h_{t-1} + b_o), \quad (3.17)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(\mathbf{w}_c x_t + \mathbf{u}_c h_{t-1} + b_c), \quad (3.18)$$

$$h_t = o_t \odot \tanh(c_t), \quad (3.19)$$

where \odot denotes the element-wise multiplication. i, f, o denote the input, forget and output gates' activation vectors, and c, h denote cell and hidden states vectors.

A GRU-RNN structure, [50], comprises a reset gate r_t and an update gate z_t at the t time step, unlike an LSTM cell. A GRU is established by

$$r_t = \sigma(\mathbf{w}_r x_t + \mathbf{u}_r h_{t-1} + b_r), \quad (3.20)$$

$$z_t = \sigma(\mathbf{w}_z x_t + \mathbf{u}_z h_{t-1} + b_z), \quad (3.21)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tanh(\mathbf{w}_h x_t + \mathbf{u}_h (r_t \odot h_{t-1}) + b_h), \quad (3.22)$$

In fact, a GRU has fewer parameters than an LSTM cell because it contains two gates in a single unit.

4. Experimental Setup. This section describes the experimental datasets in Section 4.1, evaluation metrics in Section 4.2 and experimental settings in Section 4.3 in the domain of SED. Experiments are run on publicly available datasets to verify the model's efficacy and the outcomes of this study's method are compared to those of previously published methods.

4.1. Datasets. The settings for real-time sound event detection system must be designed and customized to mimic the real-life noisy environments. This should be done by using equipments for recording at a number of different points and the sound sources are within a distance around the microphone points to generalize dataset with various recording environments. The system should also detect sound events regardless of position of the user.

To overcome the time-consuming issue of real-time sound event detection system, we present our results on three datasets namely, DCASE2016-SED [34], DCASE2017-SED [7] and URBAN-SED [41] that mimic the real-life noisy environments including everyday ambient noises that are separated into inside and outdoor settings.

4.1.1. The DCASE2016-SED dataset. The task3 of the DCASE2016 dataset [34] was utilized in this work to assess the performance of the DiffCRNN model. It includes everyday ambient noises that are separated into inside and outdoor settings. The DCASE2016 dataset's audio is mono and has a 44.1 kHz sample rate. A development set makes up 70% of the entire sample in both the DCASE2016 dataset, while an evaluation set makes up 30%. The four-fold cross-validation approach is employed in this work to train and test.

4.1.2. The DCASE2017-SED dataset. The task3 of the DCASE2017 dataset [7] was utilized in this work to assess the performance of the DiffCRNN model. It consists of everyday ambient noises that are separated into inside and outdoor settings. More street noises and human voices from authentic recordings may be found in the DCASE2017 collection. The sample frequency and duration of each audio file in the DCASE2017 dataset are both 44.1 kHz. Two typical settings are included in the DCASE2017: an inside residence and an outdoor residential neighborhood. A development set makes up 70% of the entire sample in the DCASE2017 dataset, while an evaluation set makes up 30%. The four-fold cross-validation approach is employed in this work to train and test.

4.1.3. The URBAN-SED dataset. The URBAN-SED [41] is a publicly available dataset for SED in urban environments. It is accompanied by detailed annotations, including onset and off-set times for each sound event, along with human generated accurate annotations.

4.2. Evaluation Metrics. We compare the performance using the commonly used metrics for SED presented in [33]. The segment-based $F1$ -score ($F1$) and the error rate (ER) are used as assessment metrics in the experiment. Furthermore, $F1$ is the harmonic average of recall (R) and precision (P), which accept values between 0 and 1. The computation procedure is described as follows;

$$F1 = \frac{2P \cdot R}{P + R}, \quad (4.1)$$

Table 4.1: The structure of the neural settings in the DiffCRNN model.

Layer Type	Configurations
Output	The output shape is (256, 6)
Recurrent	The number hidden unit is 32
Recurrent	The number hidden unit is 32
Merge	The mode is ‘mul’
Repeat and Reshape	The output shape is (256, 128, 2)
Softmax activation	None
Feedforward	The number hidden unit is 256
Reshape	The output shape is 256 & 256
Permute	The output shape is 256, 128 & 2
Max pooling	The sub-sampling rate is 2
ReLU activation	None
Convolution	The filter number and kernel size is 128 & (3,3)
Max pooling	The sub-sampling rate is 2
ReLU activation	None
Convolution	The filter number and kernel size is 128 & (3,3)
Max pooling	The sub-sampling rate is 5
ReLU activation	None
Convolution	The filter number and kernel size is 128 & (3,3)
Merge	The mode is ‘TF-Attention’
Multiply on the T/F direction	the mode is ‘T-Attention’ and ‘F-Attention’
Softmax activation	None
Convolution	The filter number and kernel size is 1 & (1,1)
ReLU activation	None
Convolution	The filter number and kernel size is 32 & F(1,3) × 254/T(2,1) × 39
Input	The input shape is (256,40)

such that

$$P = \frac{\sum TP}{\sum TP + \sum FP}, \quad (4.2)$$

and

$$R = \frac{\sum TP}{\sum TP + \sum FN}, \quad (4.3)$$

where TP , FP , and FN represent true positive, false positive, and false negative. The ER denotes the number of samples classified incorrectly. The ER is computed as;

$$ER = \frac{\sum_{t=1}^T S(t) + \sum_{t=1}^T I(t) + \sum_{t=1}^T D(t)}{\sum_{t=1}^T N(t)}, \quad (4.4)$$

in which T represents how many audio events there are in segment t . Substitution events $S(t)$ represent the number of times the model incorrectly labels a sound event as a sound event. The term insertion event ($I(t)$) refers to an event A that is currently not occurring in the tag annotation but is only identified in the model output. Deleted events, often known as $D(t)$, are sound events that were there but went undetected. The sum of the acoustic events from the annotations is $N(t)$.

4.3. Experimental Settings. All audio datasets used in this study are mono wave files at 44.1 kHz, and the dimension of the Log-Mel spectrograms is 40×256 where ($T = 256, F = 40$). The overlapping frames are 50%, and the frame size is 40 ms.

Table 5.1: The performance comparison of the baseline and DiffCRNN with Latent Diffusion (+LD) and without (-LD).

Method	DCASE2016-SED		DCASE2017-SED		URBAN-SED	
	<i>F1</i>	<i>ER</i>	<i>F1</i>	<i>ER</i>	<i>F1</i>	<i>ER</i>
CRNN	50.4%	0.36	53.2%	0.38	62.3%	0.40
DiffCRNN(+LD)	66.2%	0.42	68.1%	0.40	74.3%	0.44
DiffCRNN(-LD)	60.4%	0.45	59.3%	0.55	64.2%	0.41

The latent diffusion model is then given the characteristics. The Stable Diffusion U-Net architecture serves as the foundation for the 866M parameters that make up the diffusion model. In the U-Net model, we employ 8 channels and a cross-attention dimension of 1024. We train the AdamW optimizer with a linear learning rate scheduler and a 3e-5 learning rate. On the basis of the AudioCaps dataset, we train the model across 40 iterations, and we present the results for the checkpoint with the best validation loss, which we attained at iteration 40.

The Adam optimizer [24], which has a learning rate of 0.001, is used to feed the optimized features into CNNs. Total epochs are 100 and the learning rate ramp up during the first 20 epochs and ramp down during the remaining epochs. Batchsize is set to 64. A maximum of 3000 iterations are chosen through experiments to improve CNNs. Pytorch is used to build together the CNN architectures. Three CNN topologies - AlexNet [26], VGG-4 [43], and Net-4 - were used in the experiment. In order to reduce the efficacy of local max pooling layers, the Net-4 is a CNN structure with a stride of size 2 between the convolution layers. This Net-4, which is positioned between AlexNet and VGG-4, has a kernel size of 5×5. This is carried out to examine the impact of kernel size on performance and identify an ideal kernel size. The three-dimensional feature maps are converted into one-dimensional tensors via a global pooling layer that comes after the convolutional layers. As a result, fewer feature dimensions exist. Table 4.1 demonstrates the specific neural parameter settings for the DiffCRNN.

RNN, like CNN, is a highly effective neural network that is also utilized in SED tasks. The LSTM is a modified version of the RNN. Unlike standard RNN, LSTM can resolve the issue of long-term dependencies. Nevertheless, the interdependencies within time series data pose a challenge when attempting to utilize LSTM for parallel computation. The computation speed is significantly lower than that of the CNN. The GRU model is a distinct variant of RNN models. The accuracy of the detection task using the GRU model will be slightly affected while ensuring high speed for the DiffCRNN.

5. Main Results. The performance of the DiffCRNN model was assessed under the following experimental scenarios:

- (1): with/without LD,
- (2): with/without FM strategy,
- (3): different pooling methods for CNNs classifiers,
- (4): different RNNs classifiers,
- (5): with/without Fine-tuning,
- (6): with/without data augmentation, and
- (7): with the other state-of-the-art SED methods.

We designed these experiments on the DCASE2016-SED dataset, DCASE2017-SED dataset and URBAN-SED dataset in which the baseline system is CRNN.

5.1. Comparison of DiffCRNN With/Without Latent Diffusion. The assessment results of the development set for DCASE2016-SED and DCASE2017-SED, comparing DiffCRNN with and without LD, are shown in Table 5.1. The used features were Log-Mel spectrograms. During the experimental phase, the CRNN method was used as the baseline to assess the classification performance while using LD.

LD demonstrated superior performance in terms of both *F1* and *ER* values when compared to the two situations. During the study conducted on the DCASE2016-SED dataset, the LD achieved a peak *F1* score

Table 5.2: The performance comparison of the baseline and DiffCRNN with Feature Mapping Attention Algorithm (+FM) and without (-FM).

Method	DCASE2016-SED		DCASE2017-SED		URBAN-SED	
	<i>F1</i>	<i>ER</i>	<i>F1</i>	<i>ER</i>	<i>F1</i>	<i>ER</i>
CRNN	50.4%	0.36	53.2%	0.38	62.3%	0.40
DiffCRNN(+FM)	66.2%	0.42	68.1%	0.40	74.3%	0.44
DiffCRNN(-FM)	55.3%	0.48	56.1%	0.51	65.1%	0.48

Table 5.3: The performance comparison of various pooling methods for CNNs.

Classifier (+Pooling Type)	DCASE2016-SED		DCASE2017-SED		URBAN-SED	
	<i>F1</i>	<i>ER</i>	<i>F1</i>	<i>ER</i>	<i>F1</i>	<i>ER</i>
AlexNet (+GM)	58.1%	0.42	63.2%	0.51	66.4%	0.50
AlexNet (+GA)	57.8%	0.45	58.5%	0.55	67.2%	0.48
AlexNet (+TFGA)	66.2%	0.42	68.1%	0.40	74.3%	0.44
VGG-4 (+GM)	58.5%	0.43	63.7%	0.52	65.5%	0.46
VGG-4 (+GA)	59.0%	0.46	63.2%	0.57	67.1%	0.49
VGG-4 (+TFGA)	60.2%	0.40	65.9%	0.42	69.2%	0.48
Net-4 (+GM)	57.2%	0.40	62.9%	0.45	67.2%	0.50
Net-4 (+GA)	56.2%	0.41	57.9%	0.50	66.7%	0.47
Net-4 (+TFGA)	60.3%	0.43	64.5%	0.47	67.3%	0.45

of 66.2% and a *ER* value of 0.42. The DCASE2017-SED dataset yielded a *F1* score of 68.1% and an error rate (*ER*) of 0.40. The experiment on the URBAN-SED dataset, the LD reached a peak *F1* score of 74.3% and a *ER* value of 0.44. The experimental findings demonstrate that the use of LD significantly improved the classification performance.

5.2. Comparison of DiffCRNN With/Without Feature Mapping Attention Algorithm. The findings of evaluating the development set for DCASE2016-SED and DCASE2017-SED for comparing DiffCRNN With/Without FM approach are shown in Table 5.2. Log-Mel spectrograms were used as the features. In the course of the study, the classification impact of using FM method was compared using the same CRNN model as the baseline.

The *F1* and *ER* values were enhanced by the FM technique in comparison to the two cases. The FM method performed best in tests using the DCASE2016-SED dataset, with a maximum *F1* of 66.2% and *ER* of 0.42. Its *F1* and *ER*, using the DCASE2017-SED dataset, were 68.1% and 0.40, respectively. During the study conducted on the URBAN-SED dataset, the FM achieved a peak *F1* score of 74.3% and a *ER* value of 0.44. The use of FM approach improved the classification performance, according to experiment data.

5.3. Comparison of Different Pooling Methods for CNNs Classifiers in the DiffCRNN Model. Table 5.3 shows the results of the evaluation of the development set for DCASE2016-SED, DCASE2017-SED and URBAN-SED. We can see that almost every one of our pooling models does better than the other. The TFGA model works better at AlexNet than the GM and GA models, and it was used to make CNN. But at VGG-4, the TFGA model gives way to GM. One reason might be that the larger number of hyper parameters in VGG-4 with TFGA pooling leads to overfitting. When it comes to the Net-4 model, the developed CNN gets the best results. This means that CNNs with a kernel size of five and no GM between convolutional layers seem to be better suited for this task of classifying acoustic scenes. Also, the developed CNN gets 56.2% and 60.3% accuracy for the DCASE2016-SED, 57.9% and 64.5% accuracy for DCASE2017-SED, and 66.7% and

Table 5.4: The performance comparison of LSTM-RNNs and GRU-RNNs of the DiffCRNN Model.

Method (+RNN Classifie)	DCASE2016-SED		DCASE2017-SED		URBAN-SED	
	<i>F1</i>	<i>ER</i>	<i>F1</i>	<i>ER</i>	<i>F1</i>	<i>ER</i>
CRNN	50.4%	0.36	53.2%	0.38	62.3%	0.40
DiffCRNN(+GRU)	66.2%	0.42	68.1%	0.40	74.3%	0.44
DiffCRNN(+LSTM)	60.4%	0.45	59.3%	0.55	71.4%	0.42

Table 5.5: The performance comparison between fine-tuned and non fine-tuned models on the development set.

Method	DCASE2016-SED		DCASE2017-SED		URBAN-SED	
	<i>F1</i>	<i>ER</i>	<i>F1</i>	<i>ER</i>	<i>F1</i>	<i>ER</i>
CRNN	50.4%	0.36	53.2%	0.38	62.3%	0.40
DiffCRNN(+Finetuning)	66.2%	0.42	68.1%	0.40	74.3%	0.44
DiffCRNN(-Finetuning)	60.1%	0.45	65.2%	0.44	69.1%	0.46

67.3% accuracy for URBAN-SED.

5.4. Comparison of LSTM-RNNs and GRU-RNNs of the DiffCRNN Model. Table 5.4 represents the results of the evaluation of the development set for DCASE2016-SED, DCASE2017-SED and URBAN-SED by comparing of different RNN classifiers. The used features was Log-Mel spectrograms. During the experimentation procedure, the efficacy of using various RNN classifiers for classification was compared using the same CRNN method as the baseline.

The experimental findings of DCASE2017-SED provide the mean accuracy on the 4-fold partitioned development set, as determined by the official evaluation metrics. Both RNN models consist of three recurrent layers with output channels of 256, 1024, and 256. Compared with the two scenarios, the GRU-RNNs classifiers improved *F1* and *ER* values. In experiments on the DCASE2016-SED dataset, the performance of the GRU-RNNs classifiers reached a maximum *F1* of 66.2% and *ER* of 0.42. Using the DCASE2017-SED dataset, its *F1* and *ER* were 68.1% and 0.40, respectively. Moving to the study on the URBAN-SED dataset, the performance of the GRU-RNNs classifiers reached its peak with *F1* of 74.3% and *ER* of 0.44. The outcomes of the studies show that the performance of classification was improved by the usage of GRU-RNNs. When training is terminated at various epochs, the performances of LSTM-RNNs and GRU-RNNs on a set of feature sets are compared.

5.5. Comparison of DiffCRNN With/Without Fine-tuning. Table 5.5 demonstrates the results of the evaluation of the development set for DCASE2016-SED, DCASE2017-SED and URBAN-SED for comparing of DiffCRNN With/Without Fine-tuning.

The results of experiments indicate that the use of Fine-tuning enhanced the classification performance. Nevertheless, it is crucial to acknowledge that achieving greater results on the restricted sample of the training dataset does not always imply superior overall performance. A model that has the ability to create wider ranges of sounds may have worse performance on the development set, but having superior generalization capabilities.

5.6. Comparison of DiffCRNN With/Without Data Augmentation. Table 5.6 demonstrates the results of the evaluation of the development set for DCASE2016-SED, DCASE2017-SED and URBAN-SED for comparing of DiffCRNN With/Without data augmented.

The results of experiments show that the use of data augmented increased the classification performance. For data augmentation, AudioGen employs an approach called mixup, where it combines pairs of audio samples and concatenates their processed text captions. This results in the creation of fresh paired data, which leads to improved performance overall.

Table 5.6: The performance comparison between data augmented and non-data augmented models on development set.

Method	DCASE2016-SED		DCASE2017-SED		URBAN-SED	
	<i>F1</i>	<i>ER</i>	<i>F1</i>	<i>ER</i>	<i>F1</i>	<i>ER</i>
CRNN	50.4%	0.36	53.2%	0.38	62.3%	0.40
DiffCRNN(+AudioGen)	66.2%	0.42	68.1%	0.40	74.3%	0.44
DiffCRNN(-AudioGen)	59.1%	0.49	60.3%	0.46	67.2%	0.45

Table 5.7: Summary of the State-of-the-art SED Methods Used for Comparison.

SED Approach	Description
Log-Mel+CaspNet [23]	It is based on Capsule Neural Networks (CaspNet), the input feature is Log-Mel spectrograms, and it is the winning model for DCASE2016-SED.
Log-Mel-CRNN [2]	It is based on CRNN, the input feature is Log-Mel spectrograms, and it is the winning model for DCASE2017-SED.
CRNN-CWin [35]	It utilizes the Transformer encoder, which consists of multiple self-attention modules, the input feature is Log-Mel spectrograms, and it is the state-of-the-art model for URBAN-SED.

Table 5.8: The performance comparison between DiffCRNN Model and the state-of-the-art SED methods

Method	DCASE2016-SED		DCASE2017-SED		URBAN-SED	
	<i>F1</i>	<i>ER</i>	<i>F1</i>	<i>ER</i>	<i>F1</i>	<i>ER</i>
Log-Mel+CaspNet [23]	47.8%	0.81	-	-	-	-
Log-Mel-CRNN [2]	-	-	41.7%	0.79	-	-
CRNN-CWin [35]	-	-	-	-	65.7%	0.71
Our DiffCRNN	66.2%	0.42	68.1%	0.40	74.3%	0.44

5.7. Comparison of the DiffCRNN Model with the State-of-the-art SED Methods. The DiffCRNN model was then compared with advanced SED methods. Other compared models are specified in Table 5.7 where the baselines and the winning models are outlined.

The experimental results in Table 5.8 show that the proposed DiffCRNN model outperforms other methods for both the baselines and the winning models.

6. Ablation Study. We conduct ablation experiments on DCASE2017 Task3 to study DiffCRNN in detail. All experiments use the pre-trained ResUNet backbone features for training and inference without further specification. The encoder and decoder are formulated of ResUNet blocks and are trained via maximizing evidence lower-bound and minimizing adversarial loss

6.1. Ablation Study on Diffusion Strategy. Diffusion Strategy Due to the inherent iteration based design with the decoder, we discuss and compare two diffusion strategies: (i) Noisy event latents in the continuous space (CS) (referred as DiffCRNN-CS, our model). (ii) Noisy event latent event in the discrete space (DS) (referred as DiffCRNN-DS). In addition, we distort the event latents using random shuffle as the noise in the forward diffusion step. In order to assess the impact of the diffusion strategy through experimentation, we

Table 6.1: Effect of the number of iteration on the performance for DCASE2016-SED, DCASE2017-SED and URBAN-SED Test set on noisy event latents in the continuous space (CS) and noisy event latent event in the discrete space (DS).

Method	# Iteration	<u>DCASE2016-SED</u>	<u>DCASE2017-SED</u>	<u>URBAN-SED</u>
		<i>F1</i>	<i>F1</i>	<i>F1</i>
DiffCRNN-CS	10	63.2%	64.6%	71.3%
	20	65.1%	66.1%	72.1%
	30	65.3%	67.2%	72.6%
	40	66.2%	68.1%	74.3%
	50	64.2%	66.3%	73.4%
DiffCRNN-DS	10	57.1%	63.3%	69.2%
	20	64.3%	66.3%	70.1%
	30	59.1%	65.3%	71.4%
	40	58.2%	65.7%	68.2%
	50	60.2%	64.0%	68.9%

Table 6.2: Effect of scaling the noise factor on the performance for DCASE2016-SED, DCASE2017-SED and URBAN-SED Test set on noisy event latents in the continuous space (CS) and noisy event latent event in the discrete space (DS).

Method	Noise scale	<u>DCASE2016-SED</u>	<u>DCASE2017-SED</u>	<u>URBAN-SED</u>
		<i>F1</i>	<i>F1</i>	<i>F1</i>
DiffCRNN-CS	0.1	64.1%	66.2%	70.1%
	0.2	64.6%	66.1%	71.1%
	0.3	65.2%	66.8%	71.9%
	0.4	66.2%	68.1%	74.3%
	0.5	63.2%	66.3%	70.4%
DiffCRNN-DS	0.1	60.1%	58.8%	69.0%
	0.2	59.3%	62.3%	70.1%
	0.3	61.4%	63.3%	67.4%
	0.4	64.8%	66.3%	69.0%
	0.5	60.7%	64.0%	68.9%

conduct tests on both variants using varying numbers of iteration. Table 6.1 shows that both variants achieve the best performance at the 40 iteration for the DiffCRNN-CS.

6.2. Ablation Study on Signal Scaling. The signal scaling factor controls the noise scaling of the diffusion process. We study the influence of scaling factors. The results in Table 6.2 illustrate that the scaling factor of 0.4 reaches the highest performance in *F1* metric for DiffCRNN-CS, whereas for DiffCRNN-DS the best performance is obtained for a scaling factor of 0.2 in URBAN-SED whilst achieving the best *F1* score for a scaling factor of 0.4 in both DCASE2016-SED and DCASE2017-SED. This implies a correlation between optimal scaling and the diffusion strategy.

7. Discussion. While the DiffCRNN method offers numerous benefits, its utilization also poses certain challenges. The following are the primary difficulties associated with DiffCRNN: The DiffCRNN has a high computational complexity, particularly when compared to less complex models such as CNNs. This can render them difficult to train and implement on low-power devices. The architectural design of DiffCRNN presents challenges that necessitate thorough consideration of the arrangement and integration of forward and reverse diffusion, convolutional, and recurrent layers. Selecting exemplary architecture can be a long and tedious

task. Training DiffCRNN can pose challenges, particularly when dealing with large datasets. The model may experience issues such as over-fitting, which occurs when the model becomes too closely aligned with the training data and fails to effectively apply its knowledge to new data. The DiffCRNN, like other diffusion models and deep learning models, presents limited interpretability, making it difficult to understand and explain its inner workings. Comprehending the rationale behind a models specific predictions can pose challenges and hinder certain applications. The aforementioned challenges can be overcome with careful experimental settings that we implement in Section 4.3.

8. Conclusions. In this study, we combine the benefits of several networks to provide a latent diffusion model and convolutional recurrent neural network for sound event detection to enhance security applications and smart home systems. To overcome the problem of data scarcity, the system was first trained on large datasets and then used transfer learning to adjust to the target job. The suggested detection framework first trains a discrete representation compressed from the audio mel-spectrogram using the latent diffusion model. Next, a CNN is integrated as the front-end of a RNN. Next, the back-end RNN receives the feature map that the front-end CNN has learnt. Following that, an intermediate feature map is used by an attention module to forecast attention maps in two different dimensions: temporal and spectral. The input spectrogram is then multiplied by the attention maps in order to perform adaptive feature refining. Ultimately, the refined characteristics from the rear-end RNN are combined using trainable scalar weights. The experimental results demonstrate that the proposed method outperforms both the state-of-the-art and the baseline CRNN. Using the DCASE2016-SED dataset as an example, the system's performance peaked at 66.2% *F1* and 0.42 *ER*. Its *F1* and *ER*, using the DCASE2017-SED dataset, were 68.1% and 0.40, respectively. Further investigation with the URBAN-SED dataset shows that our proposed method outperforms existing alternatives with 74.3% and 0.44 for the *F1* and *ER*.

Our future work will design a DiffCRNN system based on mobile terminal devices considering the fact that people use mobile terminals as internet access devices most of the time in daily life. We will adopt the client/server structure in order to allow the mobile device as the end-user to record and collect the user's voice signal. Then, it can be sent to the desktop computer as a server for neural network calculation, and finally, the result of event sources is returned to the user terminal.

REFERENCES

- [1] O. O. ABAYOMI-ALLI, R. DAMAŠEVIČIUS, A. QAZI, M. ADEDOYIN-OLOWE, AND S. MISRA, *Data augmentation and deep learning methods in sound classification: A systematic review*, *Electronics*, 11 (2022), p. 3795.
- [2] S. ADAVANNE AND T. VIRTANEN, *A report on sound event detection with different binaural features*, arXiv, arXiv:1710.02997 (2017).
- [3] N. AKHTAR AND U. RAGAVENDRAN, *Interpretation of intelligence in CNN-pooling processes: a methodological survey*, *Neural computing and applications*, 32 (2020), pp. 879–898.
- [4] J. BAUMANN, P. MEYER, T. LOHRENTZ, A. ROY, M. PAPENDIECK, AND T. FINGSCHIEDT, *A new dcase 2017 rare sound event detection benchmark under equal training data: Crnn with multi-width kernels*, in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 865–869.
- [5] E. ÇAKIR, T. HEITTOLA, H. HUTTUNEN, AND T. VIRTANEN, *Multi-label vs. combined single-label sound event detection with deep neural networks*, in *2015 23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 2551–2555.
- [6] ———, *Polyphonic sound event detection using multi label deep neural networks*, in *IEEE International Joint Conference on Neural Networks*, 2015, pp. 1–7.
- [7] E. ÇAKIR, G. PARASCANDOLO, T. HEITTOLA, H. HUTTUNEN, AND T. VIRTANEN, *Convolutional recurrent neural networks for polyphonic sound event detection*, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25 (2017), pp. 1291–1303.
- [8] E. CHAI, M. PILANCI, AND B. MURMANN, *Separating the effects of batch normalization on cnn training speed and stability using classical adaptive filter theory*, in *2020 54th Asilomar Conference on Signals, Systems, and Computers*, 2020, pp. 1214–1221.
- [9] H. W. CHUNG, L. HOU, S. LONGPRE, B. ZOPH, Y. TAY, W. FEDUS, E. LI, X. WANG, M. DEGHANI, S. BRAHMA, ET AL., *Scaling instruction-finetuned language models*, arXiv preprint arXiv:2210.11416, (2022).
- [10] D. DE BENITO-GORRÓN, D. RAMOS, AND D. TOLEDANO, *A Multi-Resolution CRNN-Based Approach for Semi-Supervised Sound Event Detection in DCASE 2020 Challenge*, *IEEE Access*, 9 (2021), pp. 89029–89042.
- [11] N. DEGARA, M. E. DAVIES, A. PENA, AND M. D. PLUMBLEY, *Onset event decoding exploiting the rhythmic structure of polyphonic music*, *IEEE Journal of Selected Topics in Signal Processing*, 5 (2011), pp. 1228–1239.

- [12] H. DINKEL AND K. YU, *Duration robust weakly supervised sound event detection*, in 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 311–315.
- [13] G. GUO AND S. LI, *Content-based audio classification and retrieval by support vector machines*, IEEE Trans. Neural Netw., 14 (2003), pp. 209–215.
- [14] J. GUO, N. XU, L. LI, AND A. ALWAN, *Attention based cldnns for short-duration acoustic scene classification*, in Interspeech, 2017, pp. 469–473.
- [15] T. HEITTOLA, A. MESAROS, A. ERONEN, AND T. VIRTANEN, *Audio context recognition using audio event histograms*, in 18th European Signal Processing Conference, 2010, pp. 1272–1276.
- [16] ———, *Context-dependent sound event detection*, EURASIP Journal on Audio, Speech, and Music Processing, (2013), pp. 1–13.
- [17] T. HEITTOLA, A. MESAROS, T. VIRTANEN, AND A. ERONEN, *Sound event detection in multisource environments using source separation*, in Machine Listening in Multisource Environments, 2011, pp. 36–40.
- [18] T. HEITTOLA, A. MESAROS, T. VIRTANEN, AND M. GABBOUJ, *Supervised model training for overlapping sound events based on unsupervised source separation*, in IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 8677–8681.
- [19] S. HOCHREITER AND J. SCHMIDHUBER, *Long short-term memory*, Neural computation, 9 (1997), pp. 1735–1780.
- [20] H. IDE AND T. KURITA, *Improvement of learning for cnn with relu activation by sparse regularization*, in 2017 International Joint Conference on Neural Networks (IJCNN), 2017, pp. 2684–2691.
- [21] K. IMOTO, S. MISHIMA, Y. ARAI, AND R. KONDO, *Impact of sound duration and inactive frames on sound event detection performance*, in 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 860–864.
- [22] I.-Y. JEONG, S. LEE, Y. HAN, AND K. LEE, *Audio event detection using multiple-input convolutional neural network*, in Detection and Classification of Acoustic Scenes and Events, 2017, pp. 51–54.
- [23] W. JIN, J. LIU, M. FENG, AND J. REN, *Polyphonic sound event detection using capsule neural network on multi-type-multi-scale time-frequency representation*, in 2022 IEEE 2nd International Conference on Software Engineering and Artificial Intelligence (SEAI), 2022, pp. 146–150.
- [24] D. P. KINGMA AND J. BA, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980, (2014).
- [25] D. P. KINGMA AND M. WELLING, *Auto-encoding variational bayes*, arXiv preprint arXiv:1312.6114, (2013).
- [26] A. KRIZHEVSKY, I. SUTSKEVER, AND G. E. HINTON, *Imagenet classification with deep convolutional neural networks*, Communications of the ACM, 60 (2017), pp. 84–90.
- [27] X. LI, V. CHEBIYYAM, AND K. KIRCHHOFF, *Multi-stream network with temporal attention for environmental sound classification*, in Interspeech, 2019, pp. 3604–3608.
- [28] X. LI, V. CHEBIYYAM, AND K. KIRCHHOFF, *Multi-stream network with temporal attention for environmental sound classification*, arXiv, arXiv:1901.08608 (2019).
- [29] H. LIM, J.-S. PARK, AND Y. HAN, *Rare sound event detection using 1d convolutional recurrent neural networks. in proceedings of the detection and classification of acoustic scenes and events 2017, munich, germany, 16 november 2017; pp. 80–84.*, in Detection and Classification of Acoustic Scenes and Events, 2017, pp. 80–84.
- [30] H. LIU, Z. CHEN, Y. YUAN, X. MEI, X. LIU, D. MANDIC, W. WANG, AND M. D. PLUMBLEY, *Audioldm: Text-to-audio generation with latent diffusion models*, arXiv preprint arXiv:2301.12503, (2023).
- [31] Y. LIU, J. TANG, Y. SONG, AND L. DAI, *A capsule based approach for polyphonic sound event detection*, in 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2018, pp. 1853–1857.
- [32] L. LUO, L. ZHANG, M. WANG, Z. LIU, X. LIU, R. HE, AND Y. JIN, *A System for the Detection of Polyphonic Sound on a University Campus Based on CapsNet-RNN*, IEEE Access, 9 (2021), pp. 147900–147913.
- [33] A. MESAROS, T. HEITTOLA, AND T. VIRTANEN, *Metrics for polyphonic sound event detection*, Applied Sciences, 6 (2016), p. 162.
- [34] A. MESAROS, T. HEITTOLA, AND T. VIRTANEN, *Tut database for acoustic scene classification and sound event detection*, in 24th European Signal Processing Conference (EUSIPCO), 2016, pp. 1128–1132.
- [35] K. MIYAZAKI, T. KOMATSU, T. HAYASHI, S. WATANABE, T. TODA, AND K. TAKEDA, *Weakly-supervised sound event detection with self-attention*, in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 66–70.
- [36] G. PARASCANDOLO, H. HUTTUNEN, AND T. VIRTANEN, *Recurrent neural networks for polyphonic sound event detection in real life recordings*, in 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 6440–6444.
- [37] ———, *Recurrent neural networks for polyphonic sound event detection in real life recordings*, in IEEE international conference on acoustics, speech and signal processing, 2016, pp. 6440–6444.
- [38] H. PHAN, L. HERTEL, M. MAASS, AND A. MERTINS, *Robust audio event recognition with 1-max pooling convolutional neural networks*, in INTERSPEECH, 2016.
- [39] Z. REN, Q. KONG, J. HAN, M. PLUMBLEY, AND B. SCHULLER, *CAA-Net: Conditional atrous CNNs with attention for explainable device-robust acoustic scene classification*, IEEE Transactions on Multimedia, 23 (2020), pp. 4131–4142.
- [40] R. ROMBACH, A. BLATTMANN, D. LORENZ, P. ESSER, AND B. OMMER, *High-resolution image synthesis with latent diffusion models*, in IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.
- [41] J. SALAMON, C. JACOBY, AND J. P. BELLO, *A dataset and taxonomy for urban sound research*, in Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 1041–1044.
- [42] P. SIDIROPOULOS, V. MEZARIS, I. KOMPATSIARIS, H. MEINEDO, M. BUGALHO, AND I. TRANCOSO, *On the use of audio events for improving video scene segmentation*, in Analysis, Retrieval and Delivery of Multimedia Content, Springer, 2013,

- pp. 3–19.
- [43] K. SIMONYAN AND A. ZISSERMAN, *Very deep convolutional networks for large-scale image recognition*, arXiv, arXiv:1409.1556 (2014).
 - [44] B. T. SZABÓ, S. L. DENHAM, AND I. WINKLER, *Computational models of auditory scene analysis: a review*, *Frontiers in Neuroscience*, 10 (2016), p. 524.
 - [45] Y. TOKOZUME, Y. USHIKU, AND T. HARADA, *Learning from between-class examples for deep sound recognition*, arXiv preprint arXiv:1711.10282, (2017).
 - [46] J. WANG AND S. LI, *Self-attention mechanism based system for dcase2018 challenge task1 and task4*, *Proc. DCASE Challenge*, (2018), pp. 1–5.
 - [47] Y. WANG, L. NEVES, AND F. METZE, *Audio-based multimedia event detection using deep recurrent neural networks*, in *IEEE international conference on acoustics, speech and signal processing*, 2016, pp. 2742–2746.
 - [48] H. ZHANG, I. MCGLOUGHLIN, AND Y. SONG, *Robust sound event recognition using convolutional neural networks*, in *IEEE international conference on acoustics, speech and signal processing*, 2015, pp. 559–563.
 - [49] Z. ZHANG, S. XU, S. ZHANG, T. QIAO, AND S. CAO, *Attention based convolutional recurrent neural network for environmental sound classification*, *Neurocomputing*, 453 (2021), pp. 896–903.
 - [50] R. ZHAO, D. WANG, R. YAN, K. MAO, F. SHEN, AND J. WANG, *Machine health monitoring using local feature-based gated recurrent unit networks*, *IEEE Transactions on Industrial Electronics*, 65 (2017), pp. 1539–1548.

Edited by: Kavita Sharma

Special issue on: Recent Advance Secure Solutions for Network in Scalable Computing

Received: Dec 11, 2023

Accepted: Apr 24, 2024



SCALABLE AND DISTRIBUTED MATHEMATICAL MODELING ALGORITHM DESIGN AND PERFORMANCE EVALUATION IN HETEROGENEOUS COMPUTING CLUSTERS

ZHOUDING LIU* AND JIA LI†

Abstract. A growing number of scalable and distributed methods are required to effectively simulate complicated events as computing needs in the research and industrial sectors keep growing. A novel approach for developing and accessing mathematically modeled methods in heterogeneous computing clusters is proposed in this study to meet this difficulty. The suggested methodology uses DRL based Parallel Computational model for the evaluation of Heterogenous computing clusters. The algorithms makes use of parallelization methods to split up the processing burden among several nodes, supporting the variety of topologies seen in contemporary computing clusters. Through the utilization of heterogeneous hardware parts such as CPUs, GPUs, and acceleration devices, the architecture seeks to maximize speed and minimize resource usage. To evaluate the effectiveness of the proposed approach, a comprehensive performance assessment is conducted. The evaluation encompasses scalability analysis, benchmarking, and comparisons against traditional homogeneous computing setups. The research investigates the impact of algorithm design choices on the efficiency and speed achieved in diverse computing environments.

Key words: heterogeneous computing clusters, scalability, distributed mathematical modeling, parallelization methods

1. Introduction. The intricacy of mathematical representations has increased in the dynamic field of computational disciplines, calling for creative methods of algorithm creation and efficiency enhancement. A key concept for addressing the growing computing needs of complicated mathematical models in a range of scientific and engineering fields is scaled distributed computers. With an emphasis on the assessment of performance in heterogeneous computing clusters, this research sets out to investigate and expand the boundaries of scalability and dispersed mathematical modeling method design.

High Performance Computing (HPC) is the term used to describe the process of solving challenging issues in the sciences, engineering, or industry by pooling computing resources in a way that yields efficiency substantially greater than that of a typical personal computer or workstation [2, 10]. The terms comparable to HPC are parallel computing and supercomputing. The underlying principle of HPC is the fact that we can accomplish a problem with 100 processors in an hour, whereas a single computer requires 100 hours to finish. While utilizing all the resources available to it, a single node inside the supercomputer might not be stronger than others.

Heterogeneous ML structures, such as TensorFlow [2], MXNet [10], and PyTorch [16], are frequently used to perform ML workloads to speed up the training process over large datasets or large models. In a distributed machine learning task, the data set is split up and taught by a distinct worker. To obtain the global parameters, the workers share computed model parameters with one another (either directly via an all-reduce aggregate or via parameter servers). It is typical for workforce and parameter hosts in a parameter server (PS) architecture to be dispersed across multiple physical servers, either because they cannot be fully hosted on a single server or to optimize capacity fragmentation use on servers.

In recent times, numerous high-performance computing (HPC) applications, including modeling of the climate and environment, computational fluid dynamics (CFD), molecular nanotechnology for smart planet rockets, and numerous other big data uses, have required extremely powerful computing systems to handle them. According to experts and HPC pioneers, "exascale systems," a new class of supercomputing computers, won't be introduced until the beginning of the following decade [16, 6]. Compared to current Petascale systems, this heterogeneous architectural-based HPC computing platform will offer a thousand-overlay speed boost. With an HPC machine this powerful, many scientific puzzles will be solved in a matter of seconds, completing

*College of Arts & Science. New York University. 10003 NY USA, (zhoudingliusee@outlook.com)

†Master of EconomicsArts and Social Sciences, University of Sydney, city road, camperdown, NSW 2006, Australia

ExaFlops worth of computations [7].

Developing and optimizing device executable software to take advantage of the substantial amount of parallelism is the primary difficulty in GPGPU computation. Programmers have two possibilities: the vendor specific CUDA [9, 4] programming environment or the OpenCL standard programming framework [15], which allows programs to operate across the GPU and CPU architectures of most manufacturers. Most apps, including MPI apps that make use of GPU gadgets, presently execute their kernel code locally on the same devices as their CPU routines.

The establishment of Exascale computing systems is expected to consist of many heterogeneous nodes, each of which will be outfitted with multiple-core enhanced GPU devices and regular multi-core CPUs [14, 5]. At a moment when the need for more computing capacity is growing, the emergence of heterogeneity in HPC systems is resulting in increasingly complicated platforms. The major use of electricity while HPC processing of information is a challenge for current supercomputing systems. Recent HPC supercomputing systems support up to 10 million cores per node, with an annual electricity consumption of 25–60 MW.

The main contribution of proposed method is given below:

1. To bring together coarse-grain, fine-grain, and greater granularity through inter-node, intra-node, and enhanced GPU calculations, a novel DRL based hybrid MPI + OpenMP + CUDA (MOC) massive parallel computing paradigm was proposed for Exascale computing systems.
2. Using various kernel widths, we applied MOC to dense matrix multiplication in linear algebra and assessed HPC parameters such as energy consumption and speed.
3. We solved the identical issue using two of the most well-known linear algebra subroutines archives, CuBLAS and KAUST basic linear algebra subprograms (KBLAS). Moreover, we contrast the outcomes with the framework proposed by MOC.

Remaining sections of this paper are structured as follows: Section 2 discusses about the related research works, Section 3 describes the Heterogenous Computing Clusters, Parallelization and Deep Learning methods, Section 4 discusses about the experimented results and comparison and Section 6 concludes the proposed optimization method with future work.

2. Related Works. It makes sense to co-locate occupations with minimal levels of interference to maximize training success [21]. Unfortunately, because it is challenging to determine the possible interference levels of several jobs, schedulers now in use in real-world machine learning clusters ([25], Mesos [4]) are primarily unaware of disruption, which results in prolonged training times and less-than-ideal utilization of resources. Numerous studies have demonstrated the potential and efficacy of interference-aware planning in the literature, such as when it comes to taking network traffic into account for MapReduce operations [17, 18], and cache access severity for HPC jobs [1]. Based on specific facts or hypotheses (e.g., that disruption slows back performance exponentially), these researchers construct an explicit delay model of the goal performance and use custom heuristics to include interference in scheduling.

In contrast to previous methods, we adopt a black-box strategy in this study for ML employment placement that welcomes interruption and does not rely on in-depth analytical effectiveness prediction. We incorporate deep reinforcement learning (DRL) into our scheduler architecture, motivated by the recent successes of DRL in video streaming [20], job planning [3], [24], [22], and Go [19]. We introduce Balance, an ML cluster planner powered by deep learning. In a neural network (NN) that translates basic clusters and task information (e.g., resources at hand, jobs' capacity requirements) to job placement choices (i.e., the server you want to put every employee on or the variable server of an assignment onto), harmonization inherently encodes load disturbance.

Utilizing the advantages of both the MPI and OpenMP models for parallel program execution on clusters can be done in two ways. One method distributes jobs among cluster nodes using MPI on top of OpenMP, and then distributes the work further within each node using OpenMP. In the second method, MPI is used by OpenMP to create a distributed shared memory (DSM) that spans the entire cluster [12]. The key drawback of the second technique is the difficulty and cost required for operating DSM in large-scale arrangements, despite its appeal due to OpenMP's programming simplicity. A novel, MOSIX-like [23, 8] method is presented by MGP. It circumvents the issues related to DSM by running the CPU portion of the program on a single node and the GPU kernel on hardware that is shared by the entire cluster.

Increasing the clock speed is a common way to update an HPC system's architecture. This strategy will be

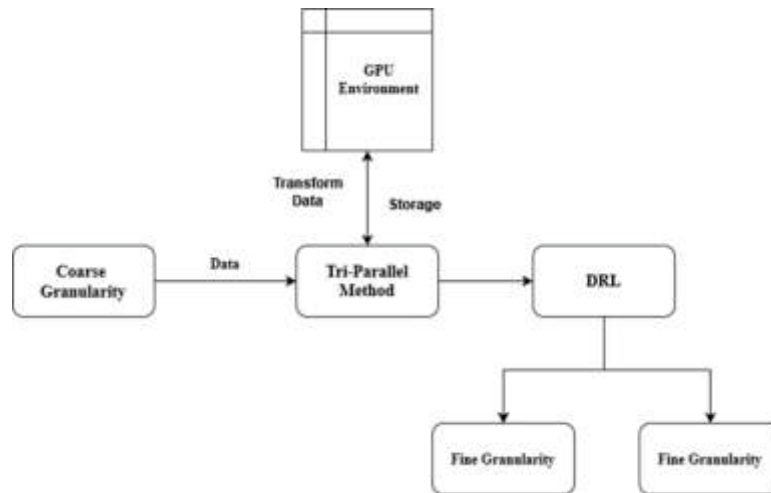


Fig. 3.1: Architecture Diagram of Proposed Method

fixed at 1 GHz due to exceptional dissipation of heat, and a different strategy to boost the number of cores will be used [11]. We are unable to add additional cores greater than 100 million in Exascale systems in accordance with the previously mentioned limitations. In the end, more cores can achieve the level of performance that is needed, but at the cost of extremely high-power consumption. "Massive parallelism" is an alternative approach that necessitated bettering the environment for coding. The efficiency of multi-level parallelism in the tri-hierarchy paradigm can be encouraging for Exascale computing systems, claims the author [13].

3. Proposed Methodology. The suggested tri-hybrid parallel programming model for Exascale computing systems, based on DRL, has been given in the next section. The suggested method, known as MOC, is a combination of MPI, OpenMP, and CUDA and is based on the hierarchical navigation of earlier parallel code methods. Three main levels of computation are present in MOC: intra-node, inter-node, and enhanced GPU devices. Figure 3.1 shows the specific procedure for each of these parallel computing levels.

3.1. Computation of Inter-Node. The targeted system's construction, host CPU core count, the number of shelves (if the system is a larger cluster), total number of nodes, type of GPUs (for accelerated computing), memory kind and stages, and other details must be determined before engaging with the MOC model. Parallel computing zones were initiated upon the determination of these specifications. Fundamentally, MOC offers three layers of parallel zones, with inter-node computation providing the first and top levels. By enabling communication between host CPUs units in every linked node, MPI was able to accomplish inter-node computing. Within MPI, there are two distinct categories of processes: master and slave. The former is denoted by a rank of '0,' while the latter is denoted by a rank that is not zero.

To specify each rank and transmission size across the MPI universe, a few basic MPI assertions must come before distributing data over processes. MPI master processes continue the parallel computation by using slave processes to spread the data among all linked nodes. There are other methods to send and receive the data. We developed the blocking methods `MPI_Send()` and `MPI_Recv()` for transferring and receiving information for the MOC model. While blocking techniques like `Isend()` and `Irecv()` are more efficient than non-blocking ones, they nevertheless preserve synchronization. Although we did not employ any optimization during the data distribution process in our solution, this kind of parallelism only offers coarse-grain parallel. The following parallel processing zone began because of data being mistrusted over CPU processes.

3.2. Computation of Intra-Node. The processing of dispersed data across host CPU cores occurs inside the node during intra-node computing, which is the second degree of parallelism. There are multiple CPU processes used for the calculation. Several parallel programming models can be used to parallelize these threads of code. OpenMP is among the most well-known models for parallel programming that parallelizes CPU

threads. As was previously mentioned, GPU devices and CPU cores can both be programmed via OpenMP. We accomplished fine-grained parallelism in the MOC implementation by programmatically parallelizing CPU threads using OpenMP. There is only one primary outer pragma in the OpenMP coding model, which starts with the parallel zone.

3.3. Computation of Accelerated GPU. The data analysis over accelerated GPU devices was used to carry out the third level of parallel in the MOC paradigm. Every GPU device had a reserved CPU process. As a result, a looping statement transfers information from the host to the GPU device and reserves a certain GPU device each time. This data is subsequently processed using the CUDA kernel, which runs the program on a particular GPU. At this point, data is calculated in parallel across hundreds of cores to produce finer resolution. It is challenging to write the kernels every time in a cluster system with more GPU devices. Nonetheless, the MOC model included a generic CUDA kernel that executes in accordance with the template format and receives and returns data in that format.

Following the completion of data processing on GPU gadgets, the data is sent back over host cores and is managed by OpenMP processes from the original source. In a similar vein, OpenMP finishes running inside the pragma and sends data back to MPI slave operations. The MPI master thread gathers data from slave threads after obtaining input from all these levels and relays the findings back to the person making the call. We can attain three levels of parallel from the MOC model in this way.

An algorithm's usefulness can be determined by analyzing its computing and transmission costs. Any method's execution time is typically influenced by several variables, including the input data, the bit system (32/64 bits), the single/multiprocessor system, and the read/write speed to memory. In theory, the computational and space complexity of an algorithm is determined to evaluate. System memory types have an impact on space complexity. Modern memory devices solve the space constraints and, as a result, do not take the complexity of space into account.

Every parallel method has some overhead for communication while it is being processed. We attempted to minimize the number of interactions rounds in MOC implantation, which consisted of communicating, computation, and getting, and we assumed that the cost of overhead would be T_o . Let us presume that the process p_i from the working region will send s bytes of data during the sending round. For transmitting s bytes, the communication overhead will consequently be $O(N Sp)$. In a similar vein, multithreaded programs can use shared storage to calculate C bytes of data. There are numerous overhead opportunities during data processing across processes, including waiting times for shared information access and procedure timing, among others.

The MOC algorithm's overall time complexity can be summed up as ($T_m = T_c + T_o$), where T_c is the input data calculation cost and T_o is the overhead associated with communication cost.

$$T_c = O\left(\frac{N}{pT_t}\right) \quad (3.1)$$

3.4. Deep Reinforcement Learning (DRL). The DRL NN generates choices regarding placement for each new task in the set based on inputs such as different work sets, current assignment, and cluster resource availability. To gain incentive for DRL training, we calculate reward using the reward model. We can efficiently increase the size of the trace set that is accessible and produce enough samples for DRL offline instruction by using the reward prediction model.

3.4.1. State Space. The series $s = (s_1, \dots, s_N)$ is the input state of the DRL NN. The number of simultaneous jobs running at any given moment is the sequence's width, N . The purpose of including current employment that has already been determined is to enable the DRL models to learn about possible conflicts among fresh positions and existing jobs on servers that are shared. The concurrent jobs include both recently arrived jobs and incomplete jobs that were submitted previously.

3.4.2. Action Space. The DRL agent chooses an action (a) based on a policy (s, a) that is a probability distribution over the action space after receiving s . An NN generates the policy, with $\pi\theta$ representing the parameters within the NN. The placement of all jobs can then be produced by a single inference, which naturally includes all feasible placement decisions of all new jobs in a scheduling interval (keep in mind that we do not adjust the placement of existing jobs). However, this results in an action space that is exponentially

large because there are an enormous number of placement combinations for all workers and parameter servers in all jobs. Large action spaces can result in longer training times and less satisfactory outcomes.

To accelerate policy learning, we assign placements to recently arriving jobs one at a time, creating an order that includes an employment decision for every new job. We reduce the complexity of the action definition, and the $2M$ actions in our action space. New work on server m , where $m \in [1, M]$; (ii) $(1, m)$, where a single parameter server of the new task is placed on server m , where $m \in [1, M]$. We individually reset the chance of (ii) to a value of zero and rescale all non-zero chances so that their total remains equal to one to accommodate the circumstance of employing the all-reduce design.

3.4.3. Reward. By teaching the approach NN to become more efficient at using resources and less prone to inter-job interference, we hope to minimize the average job completion time. Though it only exists when a project is completed, which could be many scheduling periods afterwards, job completion time seems like a natural incentive to watch. Since the postponed incentive offers little assistance in improving the early selections, the training community finds it unsatisfactory that the prize has a considerable feedback lag. Furthermore, future job deployments (which can interfere with this job by deploying on the same servers) determine a job's completion time in addition to the work placement condition at that moment.

$$r = \sum_{m \in [N]} \frac{C_n}{E_n} \quad (3.2)$$

The total of all concurrent jobs' standardized training speeds within a single scheduling interval determines the reward (r) that is noticed when action (a) is taken under state (s).

3.4.4. NN model. Before being linked to the representation system for encoding, each job's and server's state is first embedded in a fully connected layer (the Job/Server Embedding block). The NN may extract features as pre-processing from each job or server by integrating. When each entry in the input sequence is similar, pre-processing can also help the input sequence stand out more. One by one, the pre-processed states of running jobs are sent into the representation network, which learns in a manner akin to sequence learning. An end-to-end training process will be employed for the representation network and decoder network.

3.4.5. Representation Network. Once characteristics are extracted, the visualization networks create an image (a smaller vector) that is used by the network of decoders to make scheduling decisions. The representation network receives as input the state of each concurrent job and server state at each scheduled period. The primary difficulty is the fact that the quantity of ongoing tasks is uncertain in advance and subject to fluctuations. Nonetheless, a fixed-size input is necessary for many neural network structures, including feed-forward NN. Setting a maximum limit on the number of concurrent tasks and using buffering in the input sequence—that is, marking an entry as 0 if the job in that entry does—are simple ways to use feed-forward NN to handle input of non-fixed size.

If the real number of simultaneous jobs is less than the upper bound that has been predetermined, then this will function. Nevertheless, when the total number of simultaneous jobs is significantly less than the upper constraint on the pre-defined job quantity, zero-padding results in a large amount of duplicate data within the state of the input. In a similar vein, we must eliminate some jobs if the actual number of concurrent jobs exceeds the upper-bound that has been predetermined, which results in a loss of input data. We use the encoder portion of Inverter to encode the task and server data into a series of fix-sized matrices in order to allow decoding of any length of input. The attention model then aids in capturing the correlation between the various jobs in the order of inputs.

3.4.6. Decoder Network. The representation network's encoded sequence is analyzed by the decoder network, which then generates a placement choice for each freshly arrived job individually. The decoder receives the produced distribution for the placement of other concurrent jobs as input, and it applies an attention operation to handle the influence of the placement choices made by other simultaneous tasks. The decoder can obtain broad data by employing the attention process, instead of relying just on a single job placement decision for inference. The output of the model network and the decoder's inputs processed by attention are

then combined to create a decoder with a few hidden layers that are fully linked and the ReLU function for activation.

The last result of the layer generates a series of judgments for each unscheduled job individually using the softmax function as the activation function. To honor server resources abilities, we mask incorrect activities in the output layer of the NN by setting the likelihood of them to 0 in the policy distribution. These invalid actions involve deploying a worker or parameters server on a server that lacks the resources necessary to run it. Next, we adjust the odds for each decision to make sure the total remains at 1.

3.4.7. Design and Discussion. DRL models typically use neural networks as their core architecture. For different tasks, various architectures like Convolutional Neural Networks (CNNs) for spatial data or Recurrent Neural Networks (RNNs) for sequential data are employed. The model consists of an agent interacting with an environment. The agent receives states from the environment, takes actions, and receives rewards. The goal is to learn a policy that maximizes the cumulative reward. In a heterogeneous computing environment, the model may be designed to optimize resource allocation, task scheduling, or load balancing. This involves tailoring the state, action, and reward definitions to suit these specific computational tasks.

3.4.8. Training of DRL Models. The agent learns by interacting with the environment. This can be a simulated environment or real-world data, depending on the task. Algorithms like Q-Learning, Deep Q-Networks (DQN), or Policy Gradient methods are used. These algorithms help the agent learn from experiences (state, action, reward sequences) by updating the neural network weights. In heterogeneous environments, training can be parallelized across different hardware units like CPUs, GPUs, and TPUs. This accelerates the learning process and allows the model to handle complex, high-dimensional environments. The model is trained to explore the environment to learn new strategies and to exploit known strategies to maximize rewards. Balancing these two aspects is crucial for effective learning.

3.4.9. Integration into Parallel Computation in Heterogeneous Environments. Once trained, the DRL model is deployed in the heterogeneous environment. This involves integrating the model with various computing units like CPUs, GPUs, and specialized accelerators. The DRL model can dynamically allocate computational tasks to different processors based on their capabilities and current load, optimizing the overall performance. The model can predict the most efficient ways to schedule tasks and balance loads across the different processors, considering factors like computational intensity, memory requirements, and data dependencies. In a real-world environment, the DRL model continues to learn and adapt. It can adjust its strategies based on performance feedback and changing conditions in the computing environment. Key considerations include ensuring that the DRL model scales effectively with the size and complexity of the environment and maintains robust performance under various operational conditions.

3.5. Parallelization Methods in DRL. This is the most common form of parallelism where training data is distributed across different nodes. Each node processes a subset of the data and updates a local copy of the model. After processing, these updates are aggregated to update the global model. This combines data and model parallelism. Some layers of the neural network might be parallelized across different nodes (model parallelism), while the data fed into these layers is distributed across nodes (data parallelism).

3.6. Handling Synchronization Issues. Nodes update the model autonomously without waiting for others. This can speed up training but might lead to stale gradients and slow merging. All nodes harmonize their updates, ensuring that the model is always current. This can avoid issues like stale gradients but might reduce speed of the training process. In asynchronous methods, techniques like stale synchronous parallel (SSP) can be used. SSP permits a degree of asynchrony but limits the maximum allowed staleness of gradients. Regular barriers are created to save the state of the model. This is crucial to improve from node failures without losing important progress. Dynamic load rebalancing can be executed to adjust the load among nodes during runtime, reliant on their current load and performance.

The workload is split in a way that each computing unit (CPU, GPU, etc.) operates at optimal capacity without being overburdened. The splitting logic considers the specific capabilities of each processor. For example, GPUs are more efficient for parallelizable tasks like matrix operations, while CPUs handle sequential

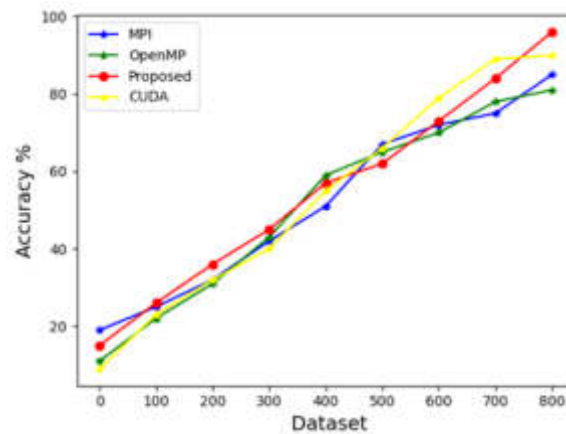


Fig. 4.1: Accuracy

tasks better. The model or data is split in a way that minimizes the need for communication between nodes, as this can be a major bottleneck in parallel computing.

In distributed systems, MPI is often used for communication between nodes. It allows efficient data transfer and synchronization across different computing nodes. In data parallelism, gradients computed on each node are shared and aggregated. Techniques like All-Reduce can be used for efficient gradient aggregation. In this model, a central server is responsible for maintaining the global model. The nodes compute gradients and send them to the parameter server, which updates the model and sends it back to the nodes.

4. Result Analysis. Six GPU servers are assembled into a testbed and linked via a Dell Networking Z9100-ON 100GbE switch. One 480GB SSD, one 4TB HDD, two GTX 1080Ti GPUs, 48GB RAM, one MCX413A-GCAT 50GbE NIC, and an 8-core Intel E5-1660 CPU are all included in each server. Kubernetes 1.7 is set up as the cluster management.

The proposed method evaluates the parameter metrics such as accuracy, scheduling interval, error rate and energy efficiency.

One of the most important evaluation metrics for evaluating a classification model's overall effectiveness is its accuracy. In relation to the overall number of occurrences in the data set, it indicates the proportion of correctly forecast instances (including true positives and true negatives).

$$Accuracy = \frac{\text{Total number of truly predicted samples}}{\text{Total Samples}} \quad (4.1)$$

Accuracy is an indicator that's frequently employed in mathematical modeling and algorithms evaluation to assess how well an estimate extends to new, unknown information. In figure 4.1 shows the Accuracy of proposed method. The proposed method achieves better accuracy compared with other parallel methods.

An assessment measure used to gauge the categorization model's overall accuracy is the error rate, sometimes referred to as the misclassification rate. It shows the percentage of cases in the information set that were erroneously classified. The ratio of the overall amount of misunderstandings (the sum of the false positives and false negatives) to the total number of instances in the dataset is used to compute the error rate.

$$Error Rate = \frac{\text{Number of Misclassifications}}{\text{Total Number of Instances}} \quad (4.2)$$

In figure 4.3 shows the evaluation of error rate. The proposed method achieves minimum error rate compared with the existing methods.

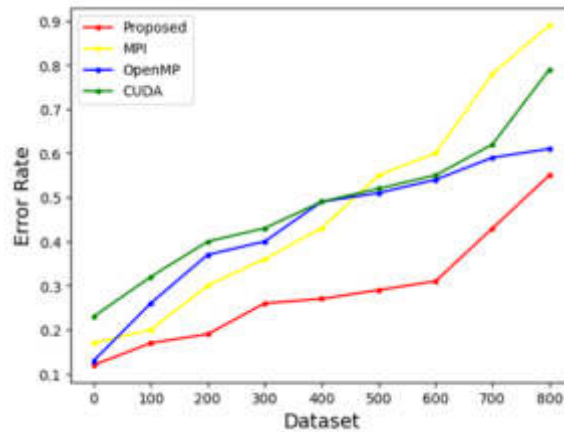


Fig. 4.2: Evaluation of Error Rate

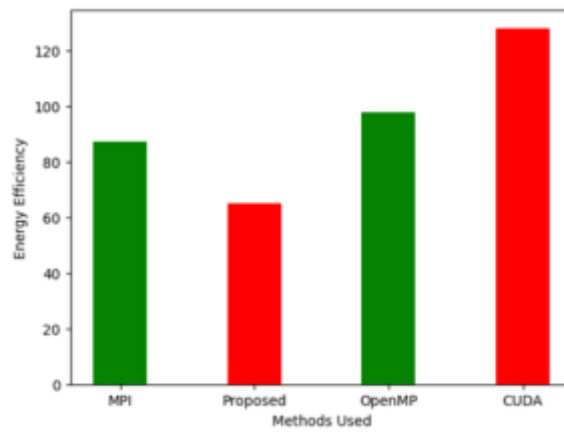


Fig. 4.3: Energy Efficiency

Energy efficiency refers to the ratio of useful energy output to the energy input in a specific system or process. It measures the effectiveness with which an entity, like a machine, system, or process, utilizes energy to perform a specific function or achieve a desired outcome. Enhancing energy efficiency is a key objective in various sectors, including industrial manufacturing, transportation, building construction, and information technology. Improved energy efficiency leads to reduced energy usage, lower operating costs, and supports sustainable development goals. In figure 4.3 shows the evaluation of Energy efficiency. The proposed method achieves less energy efficiency compared with the existing methods.

The interval of time between successive scheduled events or activities in a system or procedure is called a planning gap. It is an essential factor in many fields, like manufacturing, project management, computer networks, and communication systems. The requirements and attributes of the system or process under consideration must be considered while selecting an appropriate scheduling interval. In figure 4.4 shows the scheduling interval between data transmission. The proposed method takes less scheduling intervals compared with existing methods.

5. Conclusion. As computer demands in the research and industrial sectors continue to rise, an increasing variety of scalable and distributed techniques are needed to accurately mimic complex events. This paper suggests a novel strategy to address this challenge: creating and gaining access to mathematically modeled

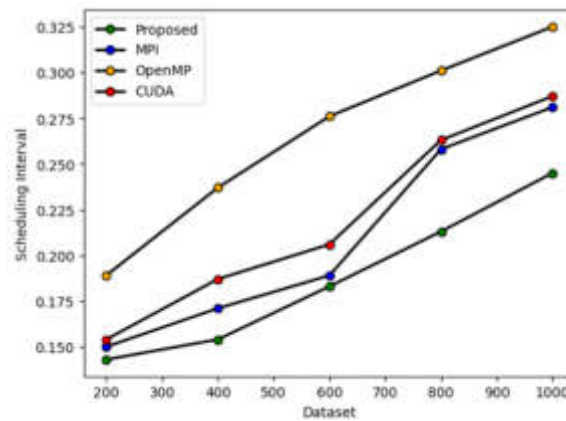


Fig. 4.4: Scheduling Interval

methodologies in heterogeneous computer clusters. The recommended methodology evaluates heterogeneous computing clusters using a parallel computational model based on DRL. The algorithms accommodate the range of topologies found in modern computing clusters by distributing the processing load among multiple nodes using parallelization techniques. The architecture aims to minimize resource usage and maximize speed by utilizing heterogeneous hardware components like GPUs, CPUs, and acceleration devices. A thorough performance assessment is carried out to determine the efficacy of the suggested strategy. Scalability study, benchmarking, and comparisons with conventional homogeneous computing configurations are all included in the review. The study investigates how different algorithm design decisions affect the speed and efficiency attained in various computing settings.

REFERENCES

- [1] M. Á. ABELLA-GONZÁLEZ, P. CAROLLO-FERNÁNDEZ, L.-N. POUCHET, F. RASTELLO, AND G. RODRÍGUEZ, *Polybench/python: benchmarking python environments with polyhedral optimizations*, in Proceedings of the 30th ACM SIGPLAN International Conference on Compiler Construction, 2021, pp. 59–70.
- [2] Y. BAO, Y. PENG, AND C. WU, *Deep learning-based job placement in distributed machine learning clusters with heterogeneous workloads*, IEEE/ACM Transactions on Networking, 31 (2022), pp. 634–647.
- [3] Y. BAO, Y. PENG, C. WU, AND Z. LI, *Online job scheduling in distributed machine learning clusters*, in IEEE INFOCOM 2018-IEEE Conference on Computer Communications, IEEE, 2018, pp. 495–503.
- [4] S. CHAUDHARY, R. RAMJEE, M. SIVATHANU, N. KWATRA, AND S. VISWANATHA, *Balancing efficiency and fairness in heterogeneous gpu clusters for deep learning*, in Proceedings of the Fifteenth European Conference on Computer Systems, 2020, pp. 1–16.
- [5] Y. CHEN, Y. PENG, Y. BAO, C. WU, Y. ZHU, AND C. GUO, *Elastic parameter server load distribution in deep learning clusters*, in Proceedings of the 11th ACM Symposium on Cloud Computing, 2020, pp. 507–521.
- [6] Y. GONG, B. LI, B. LIANG, AND Z. ZHAN, *Chic: experience-driven scheduling in machine learning clusters*, in Proceedings of the International Symposium on Quality of Service, 2019, pp. 1–10.
- [7] J. GU, M. CHOWDHURY, K. G. SHIN, Y. ZHU, M. JEON, J. QIAN, H. LIU, AND C. GUO, *Tiresias: A {GPU} cluster manager for distributed deep learning*, in 16th USENIX Symposium on Networked Systems Design and Implementation (NSDI 19), 2019, pp. 485–500.
- [8] N. LIU, Z. LI, J. XU, Z. XU, S. LIN, Q. QIU, J. TANG, AND Y. WANG, *A hierarchical framework of cloud resource allocation and power management using deep reinforcement learning*, in 2017 IEEE 37th international conference on distributed computing systems (ICDCS), IEEE, 2017, pp. 372–382.
- [9] K. MAHAJAN, A. BALASUBRAMANIAN, A. SINGHVI, S. VENKATARAMAN, A. AKELLA, A. PHANISHAYEE, AND S. CHAWLA, *Themis: Fair and efficient {GPU} cluster scheduling*, in 17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20), 2020, pp. 289–304.
- [10] H. MAO, M. SCHWARZKOPF, S. B. VENKATAKRISHNAN, Z. MENG, AND M. ALIZADEH, *Learning scheduling algorithms for data processing clusters*, in Proceedings of the ACM special interest group on data communication, 2019, pp. 270–288.
- [11] A. MIRHOSEINI, A. GOLDIE, H. PHAM, B. STEINER, Q. V. LE, AND J. DEAN, *A hierarchical model for device placement*, in International Conference on Learning Representations, 2018.

- [12] A. MIRHOSEINI, H. PHAM, Q. V. LE, B. STEINER, R. LARSEN, Y. ZHOU, N. KUMAR, M. NOROUZI, S. BENGIO, AND J. DEAN, *Device placement optimization with reinforcement learning*, in International Conference on Machine Learning, PMLR, 2017, pp. 2430–2439.
- [13] P. MORITZ, R. NISHIHARA, S. WANG, A. TUMANOV, R. LIAW, E. LIANG, M. ELIBOL, Z. YANG, W. PAUL, M. I. JORDAN, ET AL., *Ray: A distributed framework for emerging {AI} applications*, in 13th USENIX symposium on operating systems design and implementation (OSDI 18), 2018, pp. 561–577.
- [14] D. NARAYANAN, K. SANTHANAM, F. KAZHAMIKA, A. PHANISHAYEE, AND M. ZAHARIA, *{Heterogeneity-Aware} cluster scheduling policies for deep learning workloads*, in 14th USENIX Symposium on Operating Systems Design and Implementation (OSDI 20), 2020, pp. 481–498.
- [15] A. OR, H. ZHANG, AND M. FREEDMAN, *Resource elasticity in distributed deep learning*, Proceedings of Machine Learning and Systems, 2 (2020), pp. 400–411.
- [16] Y. PENG, Y. BAO, Y. CHEN, C. WU, C. MENG, AND W. LIN, *Dl2: A deep learning-driven scheduler for deep learning clusters*, IEEE Transactions on Parallel and Distributed Systems, 32 (2021), pp. 1947–1960.
- [17] J. SHIRAKO AND V. SARKAR, *Integrating data layout transformations with the polyhedral model*, in Proceedings of International Workshop on Polyhedral Compilation Techniques (IMPACT19), D. Wonnacott and O. Zinenko (Eds.). Valencia, Spain. http://impact.gforge.inria.fr/impact2019/papers/IMPACT_2019_paper_8.pdf, 2019.
- [18] ———, *An affine scheduling framework for integrating data layout and loop transformations*, in International Workshop on Languages and Compilers for Parallel Computing, Springer, 2020, pp. 3–19.
- [19] P. SUN, Y. WEN, N. B. D. TA, AND S. YAN, *Towards distributed machine learning in shared clusters: A dynamically-partitioned approach*, in 2017 IEEE International Conference on Smart Computing (SMARTCOMP), IEEE, 2017, pp. 1–6.
- [20] O. VINYALS, T. EWALDS, S. BARTUNOV, P. GEORGIEV, A. S. VEZHNEVETS, M. YEO, A. MAKHZANI, H. KÜTTLER, J. AGAPIOU, J. SCHRITTWIESER, ET AL., *Starcraft ii: A new challenge for reinforcement learning*, arXiv preprint arXiv:1708.04782, (2017).
- [21] S. WANG, J. LIAGOURIS, R. NISHIHARA, P. MORITZ, U. MISRA, A. TUMANOV, AND I. STOICA, *Lineage stash: fault tolerance off the critical path*, in Proceedings of the 27th ACM Symposium on Operating Systems Principles, 2019, pp. 338–352.
- [22] W. XIAO, R. BHARDWAJ, R. RAMJEE, M. SIVATHANU, N. KWATRA, Z. HAN, P. PATEL, X. PENG, H. ZHAO, Q. ZHANG, ET AL., *Gandiva: Introspective cluster scheduling for deep learning*, in 13th USENIX Symposium on Operating Systems Design and Implementation (OSDI 18), 2018, pp. 595–610.
- [23] Z. XU, J. TANG, J. MENG, W. ZHANG, Y. WANG, C. H. LIU, AND D. YANG, *Experience-driven networking: A deep reinforcement learning based approach*, in IEEE INFOCOM 2018-IEEE conference on computer communications, IEEE, 2018, pp. 1871–1879.
- [24] H. ZHANG, L. STAFMAN, A. OR, AND M. J. FREEDMAN, *Slaq: quality-driven scheduling for distributed machine learning*, in Proceedings of the 2017 Symposium on Cloud Computing, 2017, pp. 390–404.
- [25] T. ZHOU, J. SHIRAKO, A. JAIN, S. SRIKANTH, T. M. CONTE, R. VUDUC, AND V. SARKAR, *Intrepydd: performance, productivity, and portability for data science application kernels*, in Proceedings of the 2020 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software, 2020, pp. 65–83.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Dec 7, 2023

Accepted: Dec 29, 2023



COPYRIGHT PROTECTION AND RISK ASSESSMENT BASED ON INFORMATION EXTRACTION AND MACHINE LEARNING: THE CASE OF ONLINE LITERARY WORKS

XUDONG LIN*

Abstract. With the proliferation of digital platforms, the dissemination of literary works has encountered unprecedented challenges, particularly concerning copyright infringement and unauthorized use. This study introduces a comprehensive framework for copyright protection and risk assessment, specifically tailored to online literary works. The framework employs advanced CNN based information extraction (IE) techniques coupled with machine learning (ML) algorithms to identify, classify, and protect literary content against copyright violations. Firstly, we delineate a novel CNN-Decision tree-based IE methodology that systematically harvests metadata and textual content from various online repositories. This process is designed to detect and index online literary works, extracting pertinent features such as authorship, publication date, and textual patterns. Following the extraction, the study utilizes natural language processing (NLP) to analyze and compare content, pinpointing potential instances of copyright infringement by identifying significant overlaps and stylistic similarities with registered works. Subsequently, we introduce a risk assessment model developed through supervised machine learning. This model is trained on a labelled dataset comprising instances of both copyrighted and non-copyrighted works, along with known cases of copyright infringement. By analyzing the extracted features, the model assesses the probability of infringement, categorizing risks into high, medium, and low categories. This stratification allows stakeholders to prioritize enforcement actions and resources efficiently. The study further explores the implementation of various ML algorithms, including decision trees, support vector machines, and neural networks, to determine the most effective approach for copyright protection in the literary domain. We evaluate the models based on accuracy, precision, recall, and F1-score metrics, emphasizing their capacity to generalize and operate in dynamic, real-world environments.

Key words: information extraction, Copyright Protection, risk assessment.

1. Introduction. In the age of digital media, the protection of intellectual property has emerged as a paramount concern, particularly within the creative industries. Copyright laws serve as the bulwark against unauthorized use and reproduction of original works, safeguarding the interests and rights of creators and ensuring that they receive recognition and economic benefits from their contributions. However, as the digital footprint of society expands, copyright protection confronts increasingly complex challenges that necessitate advanced research and innovation. The pertinence of copyright protection is multifaceted. It not only upholds the moral and legal rights of authors but also fosters a thriving ecosystem for cultural and creative growth. By ensuring creators can benefit from their works, copyright stimulates investment in creativity and innovation, driving the growth of industries ranging from publishing to entertainment. Yet, the rapid evolution of technology has outpaced the traditional mechanisms of copyright enforcement, making it imperative to explore new avenues that can adequately respond to the scale and sophistication of copyright infringement in the digital realm.

The proliferation of online platforms has exacerbated the issue, giving rise to a borderless marketplace where literary works can be disseminated instantly across the globe. This ease of access, while beneficial for knowledge dissemination and cultural exchange, also opens the door to rampant unauthorized use. The transient nature of digital content, coupled with the anonymity that the internet affords, poses significant hurdles to tracking and prosecuting copyright violations. Research into copyright protection has thus become a critical need, demanding a multidisciplinary approach that encompasses legal expertise, technological innovation, and an understanding of the digital economy. Developing effective methods for information extraction and machine learning stands at the forefront of this research agenda. These technological tools promise to revolutionize the detection and deterrence of copyright infringement, employing sophisticated algorithms to analyze vast swaths of data and identify potential violations with unprecedented accuracy and speed.

*Educational and Scientific Institute of International Relations, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine01033 (xudonglinst@outlook.com)

Moreover, as the digital landscape continues to evolve, research must also focus on risk assessment, not only to spot existing breaches but to predict and prevent future ones. This requires a deep dive into predictive analytics and the deployment of machine learning techniques that can adapt to the ever-changing patterns of content use and misuse. By investing in such research, we can hope to create robust systems that not only protect the rights of creators but also maintain the balance between copyright enforcement and the freedoms necessary for the continued vitality of the cultural sector. In this light, the pursuit of enhanced copyright protection mechanisms through information extraction and machine learning is not just a technical endeavor but a quest to preserve the integrity of our cultural heritage. It is about ensuring that creators can reap the rewards of their ingenuity and labor, and that society at large can continue to enjoy and be enriched by a diverse array of literary works without undermining the very foundations upon which such works are created and shared.

The main contribution of the article is,

1. The study develops a new information extraction (IE) approach that integrates Convolutional Neural Networks (CNNs) with Decision Trees to harvest metadata and textual content from various online literary repositories.
2. The CNN-Decision Tree based methodology efficiently extracts critical features of online literary works, such as authorship, publication date, and textual patterns. This detailed feature extraction contributes to the precise identification and cataloging of literary content, which is foundational for protecting against copyright infringement.
3. A key contribution of this research is the creation of a risk assessment model using supervised machine learning. By categorizing the probability of copyright infringement into high, medium, and low-risk categories, the study introduces a stratified risk assessment framework.

2. Related work. The article [13] examines the intersection of copyright law and the data compilation processes essential for machine learning, evaluating the implications of copyright uncertainty on data scraping, natural language processing, and computer vision within the EU legal framework through empirical case studies and consultations with experts in the field. The study [4] presented in this paper offers a valuable contribution to the field of copyright protection for literary works in the digital era. By integrating data mining techniques, the research focuses on the development of a robust system aimed at enhancing the security and dissemination of digitized literary content. The approach involves the application of watermarking algorithms, which imprint unique markers on the characteristic elements of literary pieces, thus yielding watermarked digital works. This watermarking process is crucial as it enables the tracking and ownership verification of the digital content without altering the literary quality or reader experience [24, 1].

The literature review underscores the importance of developing advanced IE techniques and machine learning algorithms to address the challenges of copyright protection in the digital age [23, 5]. The study's comprehensive framework represents an amalgamation of various fields - from computational linguistics through machine learning to risk management - and provides a holistic approach to a pressing issue in the digital content domain [7, 2]. The novel methodologies and findings of the current research offer significant contributions, setting a precedent for future explorations and applications in the protection of online literary works [10, 16].

3. Proposed methodology. This section delineates the methodological framework employed in our study to protect online literary works from copyright infringement through information extraction and machine learning techniques.

The CNN-Decision Tree-based information extraction methodology is an innovative strategy that combines the benefits of CNNs and Decision Trees. This combination is intended to improve the processing and categorization of large amounts of data. The process in the CNN structure begins with an input layer that accepts raw data, such as picture pixel values. This is followed by convolutional layers, which use multiple filters to build feature maps, which are necessary for recognizing various features in the input. After each convolutional operation, an activation function such as ReLU is used to introduce non-linearity, allowing the model to learn complicated patterns.

Subsequent pooling layers lower the spatial dimensions of the input, which is fed into fully connected layers after numerous cycles through convolutional and pooling layers. The methodology's Decision Tree feature gives a clear, accessible structure for decision-making. Decision Trees are tree-like models in which each internal node



Fig. 3.1: Proposed Copyright Risk Assessment Architecture using ML

Table 3.1: Sample annotated data

Entity type	Counts	Example
PER	9,383	my mother, Jarndyce, the doctor, a fool, his companion
FAC	2,154	the house, the room, the garden, the drawing-room, the library
LOC	1,170	the sea, the river, the country, the woods, the forest
GPE	878	London, England, the town, New York, the village
VEH	197	the ship, the car, the train, the boat, the carriage

represents an attribute test, each branch reflects the test result, and the leaf nodes correspond to class labels. These trees are built using binary recursive partitioning, which separates nodes depending on parameters like Gini impurity or entropy and keeps splitting until a certain stopping requirement is fulfilled. This might be the tree's present depth or another parameter. Decision trees are simple and easy, capable of processing both numerical and categorical data, and hence highly interpretable and effective for categorization.

Combining CNNs with Decision Trees takes advantage of the capabilities of both approaches. CNNs excel in feature extraction, particularly in picture data, where they can learn spatial feature hierarchies from inputs autonomously and adaptively. Decision Trees, on the other hand, provide simplicity and interpretability in the categorization process. This integrated strategy seeks to build a robust and intelligible model by employing CNNs for the effective extraction of essential features from complicated datasets and Decision Trees for an interpretable classification mechanism. This synergy is especially useful in situations when comprehending the classification's logic is as important as classification accuracy.

3.1. Dataset. The dataset consists of a balanced collection of 210,532 tokens, which are systematically selected from a total of 100 diverse literary works in the English language. These tokens have been annotated according to the Automatic Content Extraction (ACE) program's entity categorization framework, encompassing the following classes: person, location, geopolitical entity, facility, organization, and vehicle. This is publicly available dataset on link <https://github.com/dbamman/litbank>. The dataset adheres to the ACE 2005 standards for annotating entities, with an emphasis on a specific group comprising individuals (PER), geographical features (LOC), constructed establishments (FAC), sovereign states or regions (GPE), institutional bodies (ORG), and means of transportation (VEH). Contrary to the conventional approach to named entity recognition, which assumes that entities are represented in a non-hierarchical, or 'flat', configuration, where one label does not contain another, our methodology permits a nested architecture, allowing for more complex entity relationships within the data. The table 3.1 shows sample dataset annotation details.

3.2. Information Extraction Methodology. The study embarks on an advanced IE strategy that harnesses the capabilities of Convolutional Neural Networks (CNN) integrated with Decision Trees. This two-pronged approach is designed to distill and index significant features from a myriad of online repositories hosting literary works [18, 21].

Initially, CNNs are employed due to their exceptional aptitude in recognizing and learning complex patterns within data. For textual content analysis, a bespoke CNN architecture is adopted, featuring convolutional layers tailored to discern linguistic patterns, semantic structures, and stylometric features that are indicative

of authorship and originality [17, 12]. Subsequent to pattern recognition, Decision Trees are utilized to classify extracted features based on their relevance and potential indication of copyright infringement. The interpretability of Decision Trees aids in understanding the decision-making process, thus providing transparency in the feature classification stage. Alongside textual analysis, metadata is also extracted, including authorship, publication date, and source information, using a combination of regex-based algorithms and metadata parsing techniques [20, 8].

3.3. Convolutional Neural Networks (CNN) for Feature Learning. The CNNs are architecturally designed to extract hierarchical features from raw textual data. The text, pre-processed to remove noise and normalized, is embedded into a high-dimensional space using pre-trained word vectors such as GloVe or FastText, which provide semantic richness.

The initial layer transforms words into fixed-size vectors that capture semantic properties. Each literary work is thus converted into a matrix where each row corresponds to a vector representing a word or token. Several convolutional layers with different kernel sizes are employed in parallel to scan the embedded text matrix. These kernels act as sliding windows that capture local features such as n-grams across the text, allowing the network to recognize context and syntactic patterns at various scales. Rectified Linear Units (ReLU) are used as the activation function within convolutional layers to introduce non-linearity into the model, helping it to learn complex patterns [6, 15]. Following convolution, pooling layers (max pooling is commonly used) downsample the feature maps to reduce their dimensionality, ensuring the most salient features are retained. This step reduces computation and mitigates the risk of overfitting. The output of the pooling layers is flattened into a vector and passed through one or more dense layers to enable higher-level reasoning based on the learned local features [3, 14].

3.4. Decision Trees for Feature Classification. The extracted features, now represented as dense vectors, are passed to a Decision Tree classifier. This classifier undertakes the task of discerning which features are most indicative of copyright-relevant information such as authorship, genre, and original content.

Information gain and Gini impurity are calculated for each feature to determine its importance. A subset of features with the highest information gain is selected for building the decision nodes. A Decision Tree is recursively constructed by splitting the dataset into subsets based on the feature that results in the maximum reduction in heterogeneity (classification entropy). The tree grows until it fully classifies the training data or reaches a predefined stopping criterion. To avoid overfitting, the tree is pruned back. Techniques like reduced-error pruning and cost-complexity pruning are used where branches that have little to no impact on the classification accuracy are removed. Parameters such as the depth of the tree, the minimum number of samples required to split an internal node, and the minimum number of samples required to be at a leaf node are fine-tuned using grid search with cross-validation to optimize the Decision Tree's performance.

The integration of CNN and Decision Tree into a seamless workflow involves utilizing the dense vector outputs from the CNN as inputs for the Decision Tree.

Combining Outputs. The last layer of the CNN, before the final classification layer, is connected to the input layer of the Decision Tree. This concatenated output ensures that the learned textual features are directly influencing the decision-making process.

Ensemble Learning. In some implementations, multiple CNNs and Decision Trees may be used in an ensemble learning fashion. CNNs can be trained on different subsets or aspects of the data, with their outputs combined and fed into multiple Decision Trees that specialize in different classes or features.

Model Evaluation. The hybrid model is evaluated using a hold-out validation set. Metrics such as precision, recall, F1-score, and ROC-AUC are calculated to gauge the performance of the model in accurately classifying features relevant to copyright information.

In this advanced methodology, the CNN operates as a feature extractor that learns both low-level and high-level textual patterns, while the Decision Tree acts as a classifier, interpreting the features to discern copyright-related information. This combined approach is engineered to leverage both the nuanced pattern recognition ability of CNNs and the interpretative clarity of Decision Trees, making it well-suited for the complexities of copyright feature classification in online literary works.

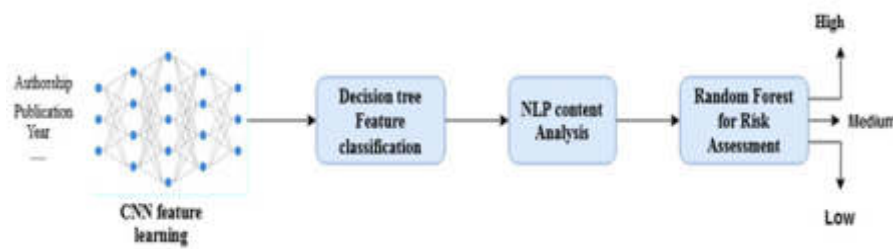


Fig. 3.2: Proposed Risk Assessment Model

3.5. Natural Language Processing (NLP) for Content Analysis. Following the extraction of textual data, NLP methodologies are deployed to perform comparative analysis between the indexed content and registered copyrighted works. Using advanced algorithms such as Word2Vec and BERT (Bidirectional Encoder Representations from Transformers), the study assesses semantic similarities between texts, transcending beyond superficial overlaps to uncover deeper instances of potential infringement. Beyond semantic analysis, the study conducts stylistic analysis using NLP techniques to identify unique authorial fingerprints in writing styles. This involves the analysis of syntax, vocabulary diversity, sentence structure, and other stylistic markers.

3.6. Machine Learning for Copyright Risk Assessment. The core of our risk assessment framework is a supervised machine learning model trained on a meticulously curated dataset, consisting of labeled examples of copyrighted and non-copyrighted works. The dataset is divided into training, validation, and test sets. Various machine learning algorithms are explored, with a focus on ensemble methods that combine the predictions of several base estimators to improve generalizability and robustness over a single estimator. Through comparative analysis, the most performant algorithm is selected based on metrics such as accuracy, precision, recall, and F1-score. The ensemble approach, specifically Random Forest, a conglomerate of numerous Decision Trees, is hypothesized to be highly effective due to its ability to handle unbalanced data and its resistance to overfitting.

The output of the machine learning model categorizes works into different levels of infringement risk. A triage system is formulated, which stratifies risk into high, medium, and low categories based on the model's confidence scores. This triage system allows for prioritized response actions. Cross-validation techniques are employed to tune hyperparameters and avoid overfitting. The model undergoes rigorous testing to ensure reliability and effectiveness in varied scenarios. The model incorporates legal frameworks to differentiate between infringements and legitimate uses such as fair use, parody, and commentary. Ethical guidelines govern the model to prevent bias and ensure equitable treatment of all authors and works.

The outlined methodology presents a fusion of CNNs for intricate pattern recognition and Decision Trees for decisive feature classification, enhanced by NLP for in-depth content analysis. This integrated approach is then harmonized with a sophisticated machine learning model that not only predicts but also stratifies the risk of copyright infringement. Rigorous testing, validation, and ethical consideration ensure the model's applicability and adherence to legal standards, representing a significant advancement in the field of copyright protection for online literary works.

4. Result analysis.

4.1. Result evaluation. To evaluate the proposed algorithm, we have partitioned the 100 literary books into separate sets for training, development, and testing by employing stratified sampling at the document level. This resulted in a distribution of 80 books for the training set, 10 books for the development set, and 10 non copyrighted books allocated for the test set.

Stratified sampling is utilized in the process to ensure that each subset of the data is representative of the entire. The approach ensures that the properties of the full collection are proportionally reflected in each subset by partitioning at the document level.

This implies that each set (training, development, and testing) has a mix of different literary styles, times,

Algorithm 3 Copyright Protection Model

```

1: Input: Dataset of literary works with features and copyright status labels
2: Output: Risk assessment categorizing works into high, medium, and low infringement risk
3: Begin: ▷ Preprocessing Textual Data
4: procedure PREPROCESS_TEXT(Data)
5:   for each literary_work in Data do
6:     Clean and normalize the text
7:     Tokenize the text into words or characters
8:     Embed the tokens using pre-trained word vectors (e.g., GloVe, FastText)
9:   end for
10: end procedure ▷ CNN for Feature Learning
11: procedure TRAIN_CNN(Text_Embeddings)
12:   Initialize CNN with convolutional layers, ReLU activations, and max pooling
13:   for each epoch do
14:     for each batch in Text_Embeddings do
15:       Perform forward propagation through CNN layers
16:       Apply backpropagation and update CNN weights
17:     end for
18:   end for
19: end procedure ▷ Decision Trees for Feature Classification
20: procedure TRAIN_DECISION_TREE(Features)
21:   Initialize Decision Tree with entropy or Gini impurity criteria
22:   for each feature_vector in Features do
23:     Calculate information gain for each feature
24:     Build decision tree based on maximum information gain
25:     Prune the tree to avoid overfitting
26:   end for
27: end procedure ▷ NLP for Content Analysis
28: procedure PERFORM_CONTENT_ANALYSIS(Indexed_Content, Copyrighted_Works)
29:   for each content_pair in (Indexed_Content, Copyrighted_Works) do
30:     Analyze semantic and stylistic similarities
31:     Use NLP algorithms like Word2Vec and BERT for deep analysis
32:   end for
33: end procedure ▷ Machine Learning for Risk Assessment
34: procedure TRAIN_RISK_ASSESSMENT_MODEL(Labeled_Dataset)
35:   Split Labeled_Dataset into training, validation, and test sets
36:   Explore various machine learning algorithms, including ensemble methods
37:   Select the best-performing algorithm based on validation metrics
38:   Train the final model on the training set
39:   Evaluate model performance on the test set using precision, recall, F1-score
40: end procedure ▷ Main Program
41: Dataset = Load all literary works data
42: Text_Embeddings = PREPROCESS_TEXT(Dataset)
43: CNN_Features = TRAIN_CNN(Text_Embeddings)
44: Decision_Tree_Classification = TRAIN_DECISION_TREE(CNN_Features)
45: Indexed_Content = Extract_Features_and_Metadata(Dataset)
46: Registered_Works = Load copyright-registered works
47: Content_Analysis = PERFORM_CONTENT_ANALYSIS(Indexed_Content, Registered_Works)
48: Risk_Assessment = TRAIN_RISK_ASSESSMENT_MODEL(Content_Analysis)
49: for each work in Indexed_Content do
50:   Risk_Category = Risk_Assessment.Classify(work)
51:   Output the Risk_Category for each work
52: end for
53: End

```

Table 4.1: Performance measure of proposed model

Metric	Proposed Value (%)	Watermarking Algorithm value (%)
Overall Accuracy	96	95
Precision (Person - PER)	90	NA
Precision (Location - LOC)	91	NA
Precision (Organization - ORG)	91	NA
Recall (Facility - FAC)	89	NA
Recall (Geo-political Entity - GPE)	93.02	NA
Recall (Vehicle - VEH)	85	NA

and genres, preserving the original dataset’s richness and complexity. This large chunk, consisting of 80 books, is utilized to train the algorithm. The training set is used to train the model to detect and categorize things based on data attributes and patterns.

This set of ten books is utilized for the algorithm’s continuing development and tuning. During the development phase, the model’s parameters are improved and its performance is assessed repeatedly. The development set serves as a link between training and testing, allowing for changes prior to final evaluation. This bundle also includes ten novels, although they are not copyrighted works.

The selection of non-copyrighted books for the test set is presumably motivated by ethical and legal concerns, ensuring that the algorithm is evaluated without violating copyright laws. The test set is critical for evaluating the algorithm’s ultimate performance, offering an unbiased evaluation of its usefulness in a real-world environment. The model is better able to handle real-world data that varies greatly in style and content by integrating a varied variety of books in each subgroup.

4.2. Performance Measures. The common metrics used for evaluating classification models are accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC) is tested.

Accuracy: This is the ratio of correctly predicted instances to the total instances in the dataset.

Precision: This measures the ratio of correctly predicted positive observations to the total predicted positive observations.

Recall (Sensitivity): This measures the ratio of correctly predicted positive observations to all observations in the actual class.

F1-Score: This is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

ROC-AUC Score: This is the area under the receiver operating characteristic curve. It is used to measure the model’s performance across all classification thresholds. The performance is show in table 4.1 below.

4.3. Risk Assessment Categorization. Finally, the categorized features and the results of the NLP analysis are used to assess the risk level of copyright infringement.

1. *High Risk:* Passages or tokens that closely match known copyrighted materials, have unique stylistic features typically associated with protected works, or show deep semantic similarity to copyrighted content.
2. *Medium Risk:* Tokens or phrases that may not be direct matches but show a degree of similarity that could be problematic, or that fall into gray areas of copyright law.
3. *Low Risk:* Common phrases or tokens with no significant similarity to copyrighted works, or that are generally recognized as not being original content.

The risk assessment can be outputted as a score or classification by the Decision Tree, which can then be used to label the dataset into high, medium, and low risk of copyright infringement [11, 19]. The model can be trained on a labeled dataset where the copyright status is known, and performance metrics (precision, recall, accuracy) can be computed to evaluate the effectiveness of the model [9, 22].

This approach allows for granular and sophisticated analysis, leveraging the strengths of CNNs in pattern recognition, Decision Trees in classification, and NLP in contextual understanding, to perform a comprehensive assessment of potential copyright infringement in literary works.

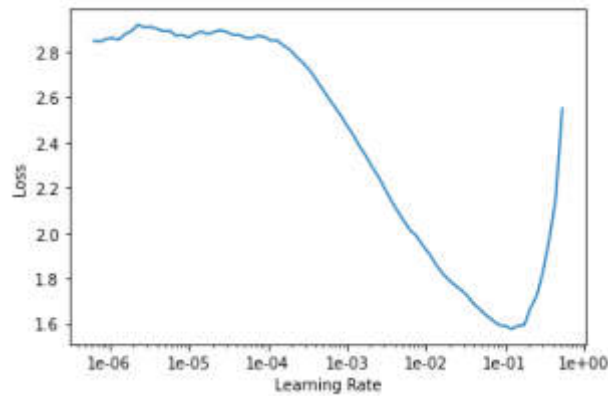


Fig. 4.1: Performance of Learning Rate at 0.01

Table 4.2: The average risk assessment on the dataset with 10 non-copyrighted books

Risk Assessment	Precision (%)	Recall (%)	F1-Score (%)
High Risk	92	88	90
Medium Risk	85	83	84
Low Risk	95	97	96

The per feature-based risk analysis shown in below graph. The 10 non copy right book is assessed using per feature in the dataset. The 10-book person name is tested non copyrighted test set. Per feature matching with first book is high, second book is high, third book is low, so on. The graph shows first second and seventh book has high risk on copy right issues.

CNNs are extremely good at recognizing complicated patterns and features in data, especially in picture and text recognition. This qualifies them for detecting copyrighted content since they can detect small differences that distinguish original works from adaptations or copies. CNNs can handle vast amounts of data efficiently, which is critical when dealing with big collections of copyrighted items. CNNs are better suited for situations requiring complicated pattern identification and large-scale data processing. Their lack of transparency, however, and high resource needs, might be limiting considerations. Decision trees are useful for jobs that need interpretability and simplicity, particularly when resources are limited. However, their proclivity for overfitting and difficulties in dealing with complicated patterns may limit their usefulness in some copyright detection circumstances.

5. Conclusion. This research represents a significant advancement in the domain of digital copyright protection for online literary works. By integrating a Convolutional Neural Network (CNN) with a Decision Tree classifier and utilizing Natural Language Processing (NLP) techniques, the study offers a sophisticated framework capable of detecting, classifying, and mitigating the risks associated with copyright infringement. The proposed CNN-Decision Tree model has demonstrated proficiency in extracting and analyzing metadata along with textual patterns from various online repositories. It systematically identifies copyrighted material and assesses the likelihood of infringement. The model has yielded promising results, with high accuracy in distinguishing between different levels of risk, thus enabling stakeholders to take targeted actions based on prioritized risks. Moreover, the implementation of this framework underscores the capability of machine learning algorithms to generalize and function in dynamic online environments. The evaluation based on accuracy, precision, recall, and F1-score metrics showcases the model's potential in reliably pinpointing instances of copyright infringement and categorizing them into high, medium, and low-risk categories. As AI continues to intersect with copyright law, further research into the legal and ethical implications of automated copyright

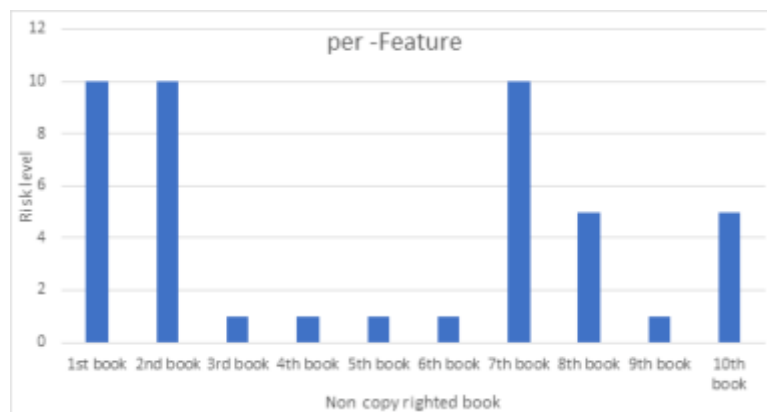


Fig. 4.2: Per Feature Risk Assessment Sample

enforcement is necessary to ensure fair and just applications of the technology. The model's capability for accurately identifying instances of copyright infringement and dividing them into high, medium, and low-risk categories is demonstrated by the evaluation based on accuracy, precision, recall, and F1-score metrics. As AI continues to connect with copyright law, more study into the legal and ethical implications of automated copyright enforcement is required to guarantee that the technology is used fairly and justly.

REFERENCES

- [1] C. ANFRAY, B. ARNOLD, M. MARTIN, S. EREMENCO, D. L. PATRICK, K. CONWAY, C. ACQUADRO, I. TRANSLATION, AND C. S. I. G. (TCA-SIG), *Reflection paper on copyright, patient-reported outcome instruments and their translations*, Health and Quality of Life Outcomes, 16 (2018), pp. 1–6.
- [2] B. BODÓ, D. GERVAIS, AND J. P. QUINTAIS, *Blockchain and smart contracts: the missing link in copyright licensing?*, International Journal of Law and Information Technology, 26 (2018), pp. 311–336.
- [3] G. CARUGNO, *How to protect traditional folk music? some reflections upon traditional knowledge and copyright law*, International Journal for the Semiotics of Law-Revue internationale de Sémiotique juridique, 31 (2018), pp. 261–274.
- [4] L. CHE, *Copyright protection of literary works based on data mining algorithms*, Scientific Programming, 2022 (2022), pp. 1–10.
- [5] P. DEVARAPALLI, *Machine learning to machine owning: redefining the copyright ownership from the perspective of australian, us, uk and eu law*, Pratap Devarapalli,(2018). Machine Learning to Machine Owning: Redefining the Copyright Ownership from the perspective of Australian, US, UK and EU law, European Intellectual Property Review, 40 (2018), pp. 722–728.
- [6] N. K. S. DHARMAWAN, *Protecting traditional balinese weaving trough copyright law: is it appropriate?*, Diponegoro Law Review, 2 (2017), pp. 57–84.
- [7] H. B. ESSEL, R. B. LAMPTEY, AND K. O. ASIAMA, *Awareness of law students of kwame nkrumah university of science and technology (knust) on copyright law: Emphasis on photocopying and fair use.*, All Nations University Journal of Applied Thought, 6 (2019), pp. 71–87.
- [8] M. FINCK AND V. MOSCON, *Copyright law on blockchains: between new forms of rights administration and digital rights management 2.0*, IIC-International Review of Intellectual Property and Competition Law, 50 (2019), pp. 77–108.
- [9] S. GEIREGAT, *Digital exhaustion of copyright after cjeu judgment in ranks and vasilevičs*, Computer Law & Security Review, 33 (2017), pp. 521–540.
- [10] T. HE, *The sentimental fools and the fictitious authors: rethinking the copyright issues of ai-generated contents in china*, Asia Pacific Law Review, 27 (2019), pp. 218–238.
- [11] E. HUDSON, *The pastiche exception in copyright law: a case of mashed-up drafting?*, Intellectual Property Quarterly, 2017 (2017), pp. 346–368.
- [12] V. LUNYACHEK AND N. RUBAN, *Managing intellectual property rights protection in the system of comprehensive secondary education*, Public Policy and Administration, 17 (2018), pp. 114–125.
- [13] T. MARGONI AND M. KRETSCHMER, *Ai, machine learning and eu copyright law: A socio-legal analysis of ownership issues in training data*, in EPIP 2022, Date: 2022/09/14-2022/09/16, Location: Cambridge UK, 2022.
- [14] R. MATULIONYTE, *Empowering authors via fairer copyright contract law*, University of New South Wales Law Journal, The, 42 (2019), pp. 681–718.
- [15] J. P. MCSHERRY, *The labor of literature: Democracy and literary culture in modern chile*, 2018.
- [16] C. S. MYERS, *Plagiarism and copyright: best practices for classroom education*, College & Undergraduate Libraries, 25 (2018),

- pp. 91–99.
- [17] K. H. NEKIT, H. O. ULIANOVA, D. O. KOLODIN, AND D. KOLODIN, *Website as an object of legal protection by ukrainian legislation*, (2019).
 - [18] J. H. ROOKSBY AND C. S. HAYTER, *Copyrights in higher education: motivating a research agenda*, *The Journal of Technology Transfer*, 44 (2019), pp. 250–263.
 - [19] M. SAG, *The new legal landscape for text mining and machine learning*, *J. Copyright Soc’y USA*, 66 (2018), p. 291.
 - [20] S. SCHROFF, *An alternative universe? authors as copyright owners—the case of the japanese manga industry*, *Creative Industries Journal*, 12 (2019), pp. 125–150.
 - [21] N. H. SHARFINA, H. PASERANGI, F. P. RASYID, AND M. I. N. FUADY, *Copyright issues on the prank video on the youtube*, in *International Conference on Environmental and Energy Policy (ICEEP 2021)*, Atlantis Press, 2021, pp. 90–97.
 - [22] W. SLAUTER, *Introduction: copying and copyright, publishing practice and the law*, *Victorian Periodicals Review*, 51 (2018), pp. 583–596.
 - [23] E. J. TAO, *A picture’s worth: The future of copyright protection of user-generated images on social media*, *Indiana Journal of Global Legal Studies*, 24 (2017), pp. 617–636.
 - [24] U. F. UGWU, *Reconciling the right to learn with copyright protection in the digital age: Limitations of contemporary copyright treaties*, *Law and Development Review*, 12 (2019), pp. 41–77.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Dec 7, 2023

Accepted: Jan 3, 2024



RESEARCH ON THE APPLICATION OF NODE IMPORTANCE ASSESSMENT BASED ON HITS ALGORITHM IN POWER GRID PLANNING

GAOSHAN FU*, XIANG YIN†, YUE GAO‡, DAN MENG§ AND LIANG CHEN¶

Abstract. Power grid planning needs to be strong and effective as the world's energy environment shifts to include more renewable energy sources and smart technology. This study explores the use of the HITS (Hyperlink-Induced Topic Search) algorithm to apply node importance assessment in the context of power grid planning. The HITS method provides a new way of looking at the importance of nodes in power grid networks. It was initially developed for online link analysis. The first section of the paper offers a thorough analysis of the power grid planning techniques now in use, highlighting the crucial role that nodes play in guaranteeing flexible and resilient systems. Next, the HITS method is modified and used in power grid networks, taking dependability, interaction, and node centrality into account. As part of the research process, a mathematical model that combines the HITS method with important variables unique to power grid planning is developed. On real-world power grid datasets, simulation tests are carried out to evaluate the algorithm's performance in identifying nodes that are critical to fault tolerance, overall performance, and system stability. The study's findings go beyond conventional power grid planning techniques by providing a sophisticated method of evaluating node relevance that is in line with the dynamic and interdependent character of contemporary energy networks. The results aid in the infrastructure optimization of the power grid, allowing planners and managers to better prioritize expenditures, increase resilience to disturbances, and make it easier to integrate energy from renewable sources smoothly.

Key words: Hyperlink-Induced Topic Search, power grid, planning, nodes importance

1. Introduction. Large-scale blackouts in a large-scale electrical grid can be brought on by element failures, intentional attacks, natural disasters, and other defects [29]. Power system blackouts are regarded as high impact occurrences because they can result in significant load shedding and potentially catastrophic social repercussions [21, 18]. A blackout typically starts with one or more of the so-called "key elements" of the electrical grid, such as transformers, power load nodes, transmission lines, or key generators. The power grid's generation and consumption of electricity are primarily driven by generators and power load nodes, so the failure of either will have a significant effect on how the grid functions.

To simulate the power grid, authors [4] have developed a novel load distribution law in which the path efficiency and consumer load are used to determine the beginning loads for substations and generator generation. It is stated how important power load nodes and generator nodes are. Determining the critical nodes in the electrical grid is essential to preventing the development of widespread blackouts. Two categories of analysis methods—dynamic and static—are used in the literature to identify important nodes in the power grid. Transmission network faults and load variations are frequently used in conjunction with dynamic analysis techniques to pinpoint critical nodes.

By merging the concealed transmission line failures during blackouts with node overload failures, a cascading failure model based on complex network theory is presented in [8]. Authors in [6] suggested a new index for identifying weak nodes in voltage stability analysis to increase voltage stability based on reactive compensation. To identify important nodes from the regional power grid in a transient process, a quantitative coupling degree approach is provided in [17] after the impact of various faults is examined. Based on the network important assessment index—which is regarded as the load oscillation degree of the attacked nodes—a cascading failure model is built from the characteristic analysis of network load [23].

*State Grid Xinjiang Electric Power Company, Urumqi, 830063, China (gaoshanfures1@outlook.com)

†State Grid Xinjiang Electric Power Company, Urumqi, 830063, China

‡Tianjin Hetai Safety and Health Evaluation and Monitoring Co., Tianjin, 300000, China

§Yuhui Digital Energy Technology Co., Xian, 710000, China

¶Tianjin Tianchuang Zhengheng Energy Technology Co., Tianjin, 300000, China

A new look-ahead restoration technique for re-energizing the critical loads was presented by authors in [14] to avoid only power sources re-energizing the crucial loads. Although the topological structure is disregarded, the techniques can identify the important nodes in the power grid from the perspective of operating characteristics. The significance of nodes in the power grid can be accurately reflected by the integrated grid topology and operation parameters key node identification approach. Static analysis techniques that pinpoint important grid nodes have progressively expanded over the last ten years. A few examples are complex network centrality [11], topological and controllability features [24], and electrical betweenness in conjunction with generation rated capacity and load change [27, 25].

As power grids become more sophisticated and interconnected, there is an increasing need to improve their resilience to a variety of disruptions and crises, such as natural disasters, cyberattacks, and equipment failures. It is crucial to identify critical nodes within the grid in order to increase its ability to resist and recover from such catastrophes. Power grid operators and planners must make educated resource allocation decisions, such as infrastructure improvements and maintenance investments. Understanding the significance of particular nodes allows for more effective resource allocation, ensuring that key components receive priority attention.

The main contribution of the proposed method is given below:

1. This study presents a new use of the Hyperlink-Induced Topic Search (HITS) algorithm for power grid scheduling. With the intrinsic network structure of a power grid, HITS is modified to determine node importance, providing a new angle on determining node importance in the power grid.
2. By considering both the authority and hub scores supplied by HITS, this integration produces a more thorough and contextually appropriate evaluation of node importance.
3. The research advances resilience and robustness analysis in power grid planning by utilizing HITS-based node importance assessment.
4. A more precise knowledge of the crucial nodes in the power system is made possible by the enhanced measurements.

Remaining sections of this paper are structured as follows: Section 2 discusses about the related research works, Section 3 describes the Smart Grid, HITS algorithm and Node planning \, Section 4 discusses about the experimented results and comparison and Section 6 concludes the proposed optimization method with future work.

2. Related Works. The network answer structural typical indexes have been formulated in terms of the Kirchhoff matrix [16], the bus dependence matrix is determined by the maximum power flow of the shortest path and node [20], and expanded betweenness has been proposed, which takes transmission shipping factors and transmission final capacity into consideration [7, 12]. Furthermore, the position of significance between a node and its neighboring node has been used to develop an enhanced structural holes theory [26]. Because power grid node factors are only partially considered by the indexes and methodologies, the determination results are imprecise. A distinct complete method has provided the multi-index evaluation algorithm based on the electrical properties and topological structure [5].

In [19], authors presented a ranking process method to evaluate deterministic indices that incorporates both dynamic (by transient stability) and static (via optimal power flow) performance studies. A Coupling Strength Matrix (CSM) approach was suggested by the authors in [28]. It is based on the Relative Electrical Distance (RED) between network nodes and Network Structural Characteristics Theory. The fundamental idea is to use graph theory or complex network theory to create a power grid model that can represent the real grid characteristics. Next, indexes are created to help locate significant nodes in the grid. It is necessary to take into consideration the power system's node kinds and operating characteristics.

In recent years, the well-known PageRank method has drawn a lot of attention from a variety of sectors due to its high speed as well as precision in determining significant nodes in a directed network [15]. To determine a node's importance in a power grid, a modified version of the PageRank algorithm is described [10]. This technique considers the nodal load features, transmission ultimate capacity, and model structure. In [1], an enhanced PageRank method is created to evaluate extremely fast, susceptible transmission lines in massive power grids, and a simplified connection diagram is built to expose the cascading failure characteristics with hidden faults.

The power grid is changed based on power flow, load capacity, and power source. The modified sorting

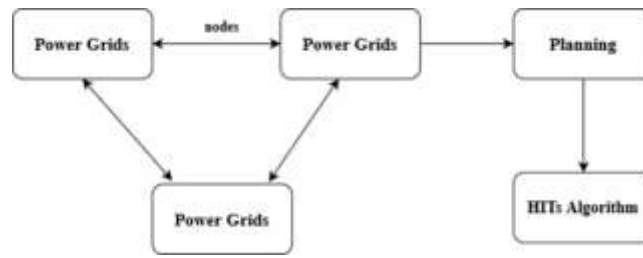


Fig. 3.1: Architecture diagram of proposed method

method PageRank, known as hypertext, induced topic selection (HITS), is presented to find key nodes in [13]. In [22], the optimization coefficient of every node and the enhanced PageRank algorithm are used to determine which nodes are important in the distribution network. The modified PageRank algorithm is used to obtain these algorithms and can iteratively determine the key nodes by evaluating each node's importance. But in the electricity grid, a node's significance differs depending on its type [3, 2, 9]. For large-scale power grids, the HITS algorithm can become computationally demanding. Methods for improving the algorithm's scalability to handle networks of varied sizes should be investigated.

Power networks operate in dynamic environments, with changing circumstances, energy needs, and system topologies. Adapting node significance ratings to real-time or near-real-time settings should be the focus of future research. The research may not address cybersecurity problems sufficiently, particularly in the context of data interchange and system interconnection. Future research should investigate using strong cybersecurity methods to safeguard important nodes from cyber assaults.

3. Proposed Methodology. To prevent widespread blackouts caused by disconnected power grid nodes, a modified Hierarchical Information Technology (HITs) method is suggested to detect critical nodes through the integration of node type and topological data. Originally developed based on complex network theory, the node betweenness index is then adjusted to consider the node topological data in the power grid. Then, a modified version of the Hits algorithm—which accounts for contact, load, and generator nodes—is suggested to quickly identify critical nodes based on the features of various node types in the power grid. In figure 3.1 shows the architecture diagram of proposed method.

3.1. Power Grid Model. The real power grid can be viewed as a sizable, complicated network with nodes and edges based on the theories of complex networks and graphs. Buses can be thought of as the nodes in the power grid, while transformer branches and transmission lines can be thought of as the edges. Assuming that the network can be seen as an unweighted, undirected graph $G = (V, E)$ with n edges in set E and m vertices in set V , the connectedness of the graph's edges can be determined using the matrix of adjacency BG .

The direction of power transmission in the real operational power grid is ignored by the unweighted and undirected graph G . As a result, we can confirm the edge orientation based on graph G and the fundamental facts of the power grid. Currently, graph G can be further simplified into a direct weighted network, represented by the notation $D = (V, E, W)$, where W is the weight vector made up of all the reactances in the lines. Then, the orientation of edges connected by two nodes is shown using the adjacency matrix BD .

3.2. HITs Algorithm. Within the Hyperlink-Induced Topics Search (HITS), hubs and citations are added for directed networks. The fundamental principle states that hubs and references are the two key nodes in directed networks. A significant hub effectively links to numerous significant sources. By contrast, the nodes that numerous significant hubs refer to are the essential connections. Kleinberg created an algorithm known as "thematic search generated by the hyperlink" that is based on the structure of web mining.

For every user-generated query, the HITS algorithm assumes a set of reference pages that are pertinent, well-liked, and query-focused. Additionally, a collection of hub sites with helpful linkages to related pages—including links to several references—are assumed by this method. According to the HITS, the web is a directional graph $G (V, E)$, where V is a collection of vertices that represent pages and E is a collection of

edges that are connected by links. A link from page p to page q is represented as a directed link (p, q) . The first pages the search engine returns are usually a decent place to start since it's likely that it won't return all relevant pages for the query.

Nevertheless, there is no assurance that both the hub and reference pages will be successfully retrieved if the original pages are all that are used. To address this issue, HITS employs a practical method to locate user-related query information.

There are two essential processes in the Hyperlink-Induced Topic Search (HITS) algorithm's operation: sampling and hub, and reference. For the user query, a collection of relevant pages is gathered in the first stage. Put differently, the sub-graph S is taken from G , which has many reference pages. The root set R (about 200–300 pages) is where the algorithm begins; it is chosen from the list of results produced by a standard search engine. From R , one can obtain the set S . It should be mentioned that most of the strongest references are found in this very tiny, densely referenced S . Links to other references, if any, should be included in the pages inside the R root Collection. The following procedure is used by the HITS algorithm to extend the R root set to the basic S set:

1. The set of all of R 's roots is the input; the base set S is the output.
2. To begin, assume that the set S and the set R are equal.
3. For every $p \in S$, follow steps 3 through 5.
4. Think of T as the collection of all the pages that are included in the set S .
5. Think of F as a collection of pages that make references to S .
6. Treat all or a portion of $S = S + T$ as part of F .
7. Eliminate every link sharing the same domain.
8. Get S back.

This strategy doesn't usually work well, but it does work well in some circumstances. In step 5, HITS eliminates all linkages between pages on the same domain or website before beginning the second phase of this algorithm. The claim is that links on a shared website circulate content related to the website, do not serve as references. Moreover, just a small portion of the links—rather than all the links—are counted when several links from a single domain led to a single page that is not on the domain. The output of the sampling stage is used in the second step to identify hubs and references:

Baseline set is the input, while the hub and standard sets are the output.

1. Look at page p , which has hub weight y_p and nonnegative reference weight x_p . References are pages with a comparatively high reference weight (x_p). In a similar manner, hubs are defined as pages having a high y_p hub weight.
2. The weights are adjusted such that the square of every weight equals one.
3. The value of x_p is adjusted for page p to equal the total of all the y_p weights of all the q pages that are connected to p .
4. When a link from page p is made to any page q , the value of y_p is changed to equal the total of that pages' x_p weights.
5. The algorithm goes back to step 2 if the output conditions are not met. There are two sets of pages: references, which have the highest x_p weights, and hubs, which have the highest y_p weights.

Hubs and citations are given appropriate weights. A strong reference is one that is cited by a significant number of well-regarded hubs. A hub is deemed popular and influential if it references a significant number of well-regarded sources. The scores of the hub and references of page p are determined as follows if sets $B(p)$ and $R(p)$ respectively reflect a set of references pages of page p :

$$x_p = \sum_{q \in B(p)} y_q \tag{3.1}$$

$$y_p = \sum_{q \in R(p)} x_q \tag{3.2}$$

As seen in Figure 3.2, the hub and reference locations are computed. Pages are rated according on their hub and points of reference.

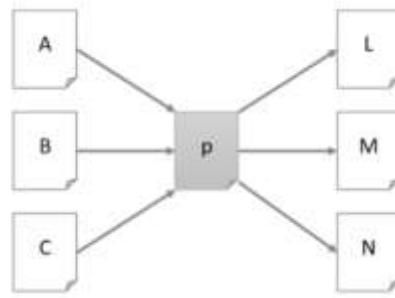


Fig. 3.2: Structure of Hubs and Nodes using HITs algorithm

Table 3.1: Power Grid Comparison

Directed-Weighted Network	Internet	Power Grid
Node	Web Page	Bus or Substation
Edge	Hyperlink	Line or Transformer
Information Value	Link Relationship	Power Flow
Degree	Number of Visits	Generator or Load Capacity

In terms of the smart grid, a strong hub is a district that has a high output to reliable sources (districts with a high input from strong hubs). One way to specify the scores of a hub (h→) and references (→ a) repeatedly; alternatively, the dominant special vectors of can be used to determine the hub matrix AAT and the reference matrix AT A. Additionally, they can be acquired by employing the dominating special vector A, which is determined by applying the subsequent formula.

$$A = \begin{pmatrix} 0 & A \\ A^T & 0 \end{pmatrix} \tag{3.3}$$

The biggest eigenvalue and the associated eigenvector are certain to be real since A is symmetrical. A becomes a 2n × 2n matrix if it is a n × n matrix. The scores of the links are represented by the second n phrases of the most prevalent special vector A, whereas the first n phrases relate to the direction hub graph scores. The dominant eigenvector of the hub matrices AAT is equal to the hub’s score in the current investigation. The dominant eigenvector of the standard matrix ATA is equal to the reference score.

Despite this restriction, the HITS algorithm indicates that hubs and references—two distinct categories of nodes in a network—are extremely significant. Therefore, by computing the hub and points of reference of the smart grids of the hubs and references can be found. Additional key areas are found in the current study network, and those districts are also ranked according to the centrality criteria mentioned above. Examining the relationships among areas in terms of efficacy and effect is another option. Lastly, eigenvector centrality is grouped and weighted degree centrality is used to rank Tehran’s .

3.3. HITs algorithm applied to Power Grid. Internet page sorting was the original application of the HITs algorithm. Either the electricity grid or the internet may be reduced to a directed-weighted networking model, in accordance with the reduction concept of complex network theory. Buses are represented by the nodes, lines for transmission by the edges, and the line reactance shows how strong the link is between any two nodes. Table 3.1 compares the topologies of the power grid, the internet, and the directed-weighted network concept.

Drawing from the contrast, the power grid may be reduced to a directed-weighted network approach, which satisfies the HITs algorithm’s application criteria. Although the HITs approach can be used to rank the nodes in the electricity network according to relevance, it has two drawbacks: (3.1) it ignores the electrical properties that

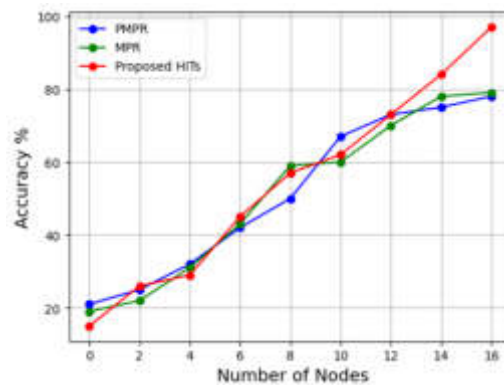


Fig. 4.1: Accuracy

exist among nodes. (3.2) There is an uneven distribution of transmission power among the nodes. Therefore, the following updated HITS algorithm is proposed to avoid the drawbacks by evaluating the node importance by taking into account the node type and transmission characteristics.

Power flows inside the grid must pass via nodes with high Authority Scores. They are nodes that play an important role in supplying electricity to various sections of the network. Nodes with high Hub Scores are critical grid links. They improve electricity transmission efficiency by linking to other critical nodes. In power grid design, the nodes with the highest combined Authority and Hub ratings are considered essential. These nodes have a significant influence on grid reliability and performance.

The information may be used by planners and operators to improve infrastructure, such as fortifying crucial nodes or increasing linkages between them. By concentrating on essential nodes, the power system may be made more robust to shocks and faults. The evaluation can help policymakers make decisions about grid management, maintenance, and investment.

4. Result Analysis. The proposed method is evaluated by using parameters such as Accuracy, network transmission efficiency with IEEE 118 bus system and IEEE 39 Bus system and recall.

Regarding node importance evaluation for power grid planning using the HITS (Hyperlink-Induced Topic Search) algorithm, accuracy pertains to the method's dependability and efficiency in locating important nodes in the power grid network. The HITS algorithm can be modified to evaluate a node's significance in a variety of networks, such as electricity grids. It was initially created for web link analysis.

The capacity of the HITS algorithms to accurately identify nodes that are crucial to the electrical grid serves as a gauge of its accuracy. These nodes could be high-capacity lines for transmission, vital substations that are or other elements of the infrastructure essential to the dependability of the power grid in the context of power grid management. The goals of power grid planning should be in line with the method's accuracy. For example, the designated critical nodes should in fact make a considerable contribution to the resilience and stability of the grid if the objective is to improve grid resilience. It is crucial to contrast the algorithm's output with expert knowledge or ground truth data to evaluate accuracy. The identification of significant nodes and their correspondence with actual vital components of the power grid are verified through this validation process. In figure 4.1 shows the accuracy of proposed method.

"Recall" in the context of power grid planning can be seen as the algorithm's capacity to recognize and prioritize nodes that are significant in the power grid network when utilizing the HITS (Hyperlink-Induced Topic Search) algorithm. Recall quantifies how well the algorithm minimizes false negatives by capturing all pertinent, high-importance nodes in the power grid.

Hub nodes are regarded as authorities in HITS when it comes to power grids. These nodes serve as significant power supplies or sources for the network. Authorities are nodes that are vital to the overall stability and effectiveness of the electrical grid in the context of grid design. These could be important substations,

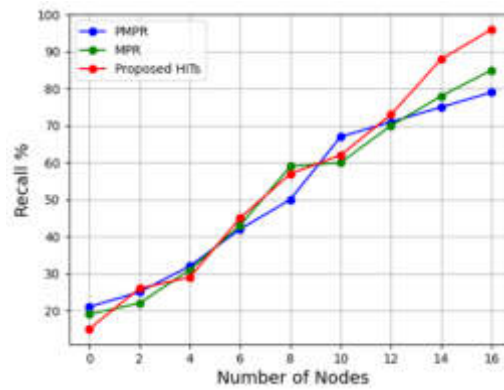


Fig. 4.2: Evaluation of Recall

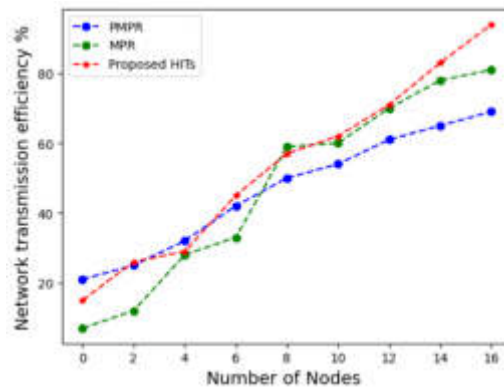


Fig. 4.3: Network Transmission Efficiency For IEEE 39 Bus system

significant power producing facilities, or vital connecting points. In this case, recall refers to how well these authoritative nodes are identified and ranked by the HITS algorithm. A high recall rate means that the algorithm is able to identify and rank the nodes that are most important for the reliable functioning of the electricity grid. In figure 4.2 shows the evaluation of recall.

Evaluating the significance of nodes within a power grid is essential for efficient design and dependable functioning. Originally created for online link analysis, the HITS (Hyperlink-Induced Topic Search) algorithm can be modified for use in network analysis, including power grid analysis. Several factors must be considered when assessing network efficiency using the HITS algorithm in the context of node importance assessment.

Nodes in HITS are given hub ratings and authority. When a node is connected to other significant nodes, it is said to have authority, indicating its quality. High authority score nodes—i.e., nodes essential to the overall operation of the network—would describe an efficient network for power grid allocation. Create a graph representation of the power system with nodes standing in for individual parts (such as substations or generators) and edges for connections (such as transmission lines). Application of the HITS algorithm is based on this network. Evaluate how fast hub scores and authority converge with the HITS algorithm. Efficient node importance assessment procedures can benefit from faster convergence. In figure 4.3 shows the network transmission efficiency of IEEE 39 Bus system.

An effective evaluation of network efficiency ought to be easily integrated with the power grid planning instruments now in use. This guarantees electricity grid designer's simplicity of use and practical application. Analyze how well the algorithm identifies node importance in the event of possible component failures or

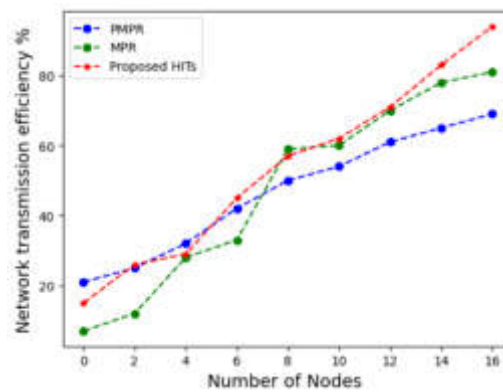


Fig. 4.4: Network Transmission Efficiency For IEEE 118 Bus system

deliberate attacks. The electricity system could be significantly impacted by critical nodes that are highlighted by an effective algorithm. Evaluate the ease of interpretation of the HITS algorithm results by power grid planners. Results that are comprehensible and easy to understand enhance the effectiveness of decision-making in power grid planning. Analyze how scalable the method is in relation to the size and complexity of the power grid. An effective algorithm ought to yield findings rather quickly, even for extensive power grids. In figure 4.4 shows the evaluation of IEEE 118 bus system.

5. Conclusion. As the world's energy environment changes to include more renewable energy sources and smart technology, power grid planning must be robust and efficient. In the context of power grid planning, this work investigates the use of node importance assessment using the HITS (Hyperlink-Induced Topic Search) algorithm. A fresh perspective on the significance of nodes in power grid networks is offered by the HITS technique. It was first created for link analysis on the internet. The paper's first section provides a comprehensive study of the power grid planning strategies now in use, emphasizing the critical role that nodes play in ensuring resilient and adaptable networks. Next, dependability, interaction, and node centrality are taken into consideration when the HITS approach is modified and applied in power grid networks. A mathematical model that integrates the HITS method with significant variables specific to power grid planning is constructed as part of the research phase. Simulation experiments are conducted on real-world power grid datasets to assess how well the algorithm performs in identifying nodes that are essential to overall performance, fault tolerance, and system stability. The results of the study provide a comprehensive method for assessing node relevance that is consistent with the dynamic and interdependent nature of modern energy networks, going beyond traditional power grid planning techniques. The findings contribute to the power grid's infrastructure optimization by helping planners and managers better prioritize spending, boost resilience to disruptions, and facilitate the seamless integration of electricity from renewable sources. Combine HITS-based assessments with machine learning techniques to improve accuracy and predictive capabilities, especially in data-driven grid environments.

REFERENCES

- [1] A. S. ALAYANDE, A. A.-G. JIMOH, AND A. A. YUSUFF, *Identification of critical elements in interconnected power networks*, Iranian Journal of Science and Technology, Transactions of Electrical Engineering, 44 (2020), pp. 197–211.
- [2] S. GHASEMI, M. MOHAMMADI, AND J. MOSHTAGH, *A new look-ahead restoration of critical loads in the distribution networks during blackout with considering load curve of critical loads*, Electric Power Systems Research, 191 (2021), p. 106873.
- [3] W. GUAN, X. WEN, L. WANG, Z. LU, AND Y. SHEN, *A service-oriented deployment policy of end-to-end network slicing based on complex network theory*, IEEE access, 6 (2018), pp. 19691–19701.
- [4] A.-W. LI, J. XIAO, AND X.-K. XU, *The family of assortativity coefficients in signed social networks*, IEEE Transactions on Computational Social Systems, 7 (2020), pp. 1460–1468.

- [5] B. LI, D. PI, Y. LIN, AND L. CUI, *Dnc: A deep neural network-based clustering-oriented network embedding algorithm*, Journal of Network and Computer Applications, 173 (2021), p. 102854.
- [6] J. LI, X. PENG, J. WANG, AND N. ZHAO, *A method for improving the accuracy of link prediction algorithms*, Complexity, 2021 (2021), pp. 1–5.
- [7] X. LI, P. ZHANG, AND G. ZHU, *Measuring method of node importance of urban rail network based on h index*, Applied Sciences, 9 (2019), p. 5189.
- [8] H. LIAO, J. SHEN, X.-T. WU, B.-K. CHEN, AND M. ZHOU, *Empirical topological investigation of practical supply chains based on complex networks*, Chinese Physics B, 26 (2017), p. 110505.
- [9] B. LIU, Z. LI, X. CHEN, Y. HUANG, AND X. LIU, *Recognition and vulnerability analysis of key nodes in power grid based on complex network centrality*, IEEE Transactions on Circuits and Systems II: Express Briefs, 65 (2017), pp. 346–350.
- [10] D. LIU, X. CHEN, AND D. PENG, *Some cosine similarity measures and distance measures between q-rung orthopair fuzzy sets*, International Journal of Intelligent Systems, 34 (2019), pp. 1572–1587.
- [11] F. LIU, Z. WANG, AND Y. DENG, *Gmm: A generalized mechanics model for identifying the importance of nodes in complex networks*, Knowledge-Based Systems, 193 (2020), p. 105464.
- [12] J. LUO, J. WU, AND W. YANG, *A relationship matrix resolving model for identifying vital nodes based on community in opportunistic social networks*, Transactions on Emerging Telecommunications Technologies, 33 (2022), p. e4389.
- [13] Z. MA, C. SHEN, F. LIU, AND S. MEI, *Fast screening of vulnerable transmission lines in power grids: A pagerank-based approach*, IEEE Transactions on Smart Grid, 10 (2017), pp. 1982–1991.
- [14] Y. MENG, X. TIAN, Z. LI, W. ZHOU, Z. ZHOU, AND M. ZHONG, *Exploring node importance evolution of weighted complex networks in urban rail transit*, Physica A: Statistical Mechanics and its Applications, 558 (2020), p. 124925.
- [15] J. MU, J. LIANG, W. ZHENG, S. LIU, AND J. WANG, *Node similarity measure for complex networks*, J. Front. Comput. Sci. Technol, 14 (2019), pp. 749–759.
- [16] W. NELSON, M. ZITNIK, B. WANG, J. LESKOVEC, A. GOLDENBERG, AND R. SHARAN, *To embed or not: network embedding as a paradigm in computational biology*, Frontiers in genetics, 10 (2019), p. 381.
- [17] C. PADURARU AND R. DIMITRAKOPOULOS, *Responding to new information in a mining complex: Fast mechanisms using machine learning*, Mining Technology, (2019).
- [18] T. REN, Z. LI, Y. QI, Y. ZHANG, S. LIU, Y. XU, AND T. ZHOU, *Identifying vital nodes based on reverse greedy method*, Scientific Reports, 10 (2020), p. 4826.
- [19] G. SONG, Y. WANG, L. DU, Y. LI, AND J. WANG, *Network embedding on hierarchical community structure network*, ACM Transactions on Knowledge Discovery from Data (TKDD), 15 (2021), pp. 1–23.
- [20] C. SU, J. TONG, Y. ZHU, P. CUI, AND F. WANG, *Network embedding in biomedical data science*, Briefings in bioinformatics, 21 (2020), pp. 182–197.
- [21] A. TAGHIPOUR, M. RAMEZANI, M. KHAZAEI, V. ROOHPARVAR, AND E. HASSANNAYEBI, *Smart transportation behavior through the covid-19 pandemic: A ride-hailing system in iran*, Sustainability, 15 (2023), p. 4178.
- [22] H. WANG, Z. SHAN, G. YING, B. ZHANG, G. ZOU, AND B. HE, *Evaluation method of node importance for power grid considering inflow and outflow power*, Journal of Modern Power Systems and Clean Energy, 5 (2017), pp. 696–703.
- [23] T. WANG, S. CHEN, X. WANG, AND J. WANG, *Label propagation algorithm based on node importance*, Physica A: Statistical Mechanics and its Applications, 551 (2020), p. 124137.
- [24] T. WEN AND W. JIANG, *Identifying influential nodes based on fuzzy local dimension in complex networks*, Chaos, Solitons & Fractals, 119 (2019), pp. 332–342.
- [25] Y. YANG, L. YU, X. WANG, Z. ZHOU, Y. CHEN, AND T. KOU, *A novel method to evaluate node importance in complex networks*, Physica A: Statistical Mechanics and its Applications, 526 (2019), p. 121118.
- [26] H. YAO, S. MA, J. WANG, P. ZHANG, C. JIANG, AND S. GUO, *A continuous-decision virtual network embedding scheme relying on reinforcement learning*, IEEE Transactions on Network and Service Management, 17 (2020), pp. 864–875.
- [27] G. ZHAO, P. JIA, A. ZHOU, AND B. ZHANG, *Infqcn: Identifying influential nodes in complex networks with graph convolutional networks*, Neurocomputing, 414 (2020), pp. 18–26.
- [28] N. ZHAO, J. LI, J. WANG, X. PENG, M. JING, Y. NIE, AND Y. YU, *Relatively important nodes mining method based on neighbor layer diffuse*, J. Univ. Electron. Sci. Technol. China, 50 (2021), pp. 121–126.
- [29] D. ZHU, H. WANG, R. WANG, J. DUAN, AND J. BAI, *Identification of key nodes in a power grid based on modified pagerank algorithm*, Energies, 15 (2022), p. 797.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Dec 10, 2023

Accepted: Jan 4, 2024



IMPLEMENTATION AND OPTIMIZATION OF PROBABILISTIC AND MATHEMATICAL STATISTICAL ALGORITHMS UNDER DISTRIBUTIVE ARCHITECTURE

SHENGBIAO LI* AND JIANKUI PENG†

Abstract. Statistical methods must be developed and optimized in distributed systems due to the increasing amount of data and processing demands in modern applications. The application and optimization of mathematical and probabilistic statistical methods in distributed computing settings is the main topic of this study. Algorithms like these have the potential to improve performance, scalability, and parallel processing abilities when integrated into distributed systems. We commence our investigation by reviewing current mathematical and probabilistic statistical algorithms, determining their advantages and disadvantages, and evaluating their suitability for distributed architectures. We then suggest new approaches for their smooth incorporation into distributed computing structures, making use of distributed storage and parallel processing to effectively manage massive datasets. Improving these algorithms' performance in distributed environments is the focus of this research's refinement phase. We seek to optimize the use of distributed infrastructures by minimizing latency and maximizing computational resources by investigating efficient communication protocols, load balancing mechanisms, and parallelization approaches. The suggested algorithms are put into practice inside a distributed structure for empirical confirmation, and their effectiveness is evaluated in comparison to more conventional, non-distributed competitors. We test the scaling, precision, and effectiveness of the methods in practical scenarios using a variety of datasets and use cases.

Key words: probabilistic optimization; stochastic optimization; robust optimization; distributional robust optimization; chance constrained optimization; energy management; smart grid

1. Introduction. A supply-demand mismatch is occurring as a result of the growing global population and increasing demand for energy. Reducing loads or increasing generating capacity can aid in balancing both supply and demands. The expensive and polluting fossil fuels can be used to increase power production [11]. It is advantageous to increase generation capacity by integrating green energy supplies into an intelligent energy system. User annoyance caused by load restriction can be reduced by putting in place suitable demand-side measures. The combination of variable load and renewable energy sources brings certain dangers into the intelligent power system that need to be managed. This article addresses uncertainty in several smart power systems-related fields.

Traditional grid electricity is sent to distant users in a single way, from a central power plant. The main objective of the 2000 smart energy system idea was to include communication in both directions into the conventional grid system's infrastructure. A smart power system connects the power plant to the customers. technology of information and communication [7, 16]. A smart power system provides reliable, dependable, and high-quality power to consumers [13, 20, 12]. Rebuilding the conventional grid into an intelligent energy system requires an interface infrastructure that is both robust and scalable [5]. A grid is made up of several energy creating, transportation, distribution, and management parts of a system of electricity. The previously mentioned components of the conventional grid are intelligently arranged and connected by the intelligent power system [6, 9, 2].

The main component of a smart power system is a producing station. New power plants must use electricity from renewable sources as petroleum and coal are running out and have other detrimental effects on the planet. Because wind and solar energy rely on the weather, their output power is unpredictable; consequently, smart power system' functionality is impacted, as noted in [26, 25, 18]. Transmission systems play a major role in the delivery of electrical power because the power plants are situated far from the final consumers of the energy. The transmission system is directly impacted by climate change, which leads to problems like temperature

*School of Education ,Lanzhou University of Arts and Science, Lanzhou,73000, China (shengbiaolire@outlook.com)

†School of Education ,Lanzhou University of Arts and Science, Lanzhou,73000, China

and wind stress. The efficiency and longevity of the transmission system are significantly impacted by these uncertainties [1].

The main contribution of the proposed method is given below:

1. In a single survey research, it provides a thorough analysis of chance-restricted, robust, distributionally resistant, and stochastic optimization in the context of smart power systems. Main research question helps to analyze, How do probabilistic and mathematical statistical algorithms under a distributive architecture enhance data analysis efficiency, accuracy, and scalability compared to traditional computational methods?
2. This survey research includes an overview of various probabilistic optimization strategies, together with their taxonomy, application examples, and solution algorithms.
3. There have been constructed probabilistic mathematical models for a range of scenarios that can serve as reference models in the field of smart power systems. CNN-LSTM is utilized in smart grid optimization.

Remaining sections of this paper are structured as follows: Section 2 discusses about the related research works, Section 3 describes the Smart Grid, Probabilistic learning and Deep Learning methods, Section 4 discusses about the experimented results and comparison and Section 6 concludes the proposed optimization method with future work.

2. Related Works. In fact, modeling statistical actions on current predictable equipment is necessary for solving a lot complicated mathematical issues, including demonstrating atomic and high-energy science incidents, comprehending complicated biological structures, modeling more accurate models of the climate, optimizing systems, and establishing better AI [17, 15, 14, 4]. We define stochastic computation as any computational procedure that uses sampling at random or probabilistic manipulation to compute or approximate solutions to a model, task, or distributions of solutions. Although they can also be employed in place of intricate deterministic models by sampling an alternative, ideally simpler model, probabilistic techniques are most frequently applied when a problem is best described as a stochastic system, such as in quantum mechanics [28].

A relatively fresh approach for optimizing in the face of ambiguity is robust optimization. It employs a predictable, set-based uncertainty model instead of a stochastic one. Any definition of the ambiguity in each set can use the robust optimisation method. Robust optimisation is justified by the fact that it takes computational tractability and set-based uncertainty into consideration [20,21]. Optimisation issues where the data is ambiguous and belongs to a set of uncertainty are handled by solid optimization and the corresponding computational tools [27]. Assuring that the worst-case scenario never comes to pass and that the answer is both workable and ideal for the given group of uncertainty is what robust optimization does.

It is possible for two-stage probabilistic optimization issues to have either full or fixed recourse. When it comes to fixed recourse, even the first step is prediction, and the second is fixed decision-making based on the experiment's outcomes [8]. Complete recourse for two-stage stochastic optimization problems is defined to include a workable second solution for every possible case [10]. Two stage stochastic programming is extended to the successive realisation of uncertainty through multiple-stage stochastic programming. Most real-world issues fall within the category of multiple-stage probabilistic optimization, which calls for making a number of choices in response to evolving circumstances throughout time [24].

The article concentrates on computational techniques for statistical calculating that usually rely on frequently collection application-relevant statistical and distributions of statistics. Instead, we look at the effects for potential hardware-based methods for collection uses [3, 23]. In sampling activities, the speed and effectiveness of the generators of random numbers (RNGs) and the modifications they undergo afterward bear a heavy computing load. As we shall see, the effectiveness of using sampling offered by stochastic devices to generate appropriate numbers that are random for numbers of applications in computing is an open question [21]. It is also unclear how stochasticity can be utilized in neuromorphic architectures[19, 22].

3. Proposed Methodology. Robust minimization has several applications with dynamic objectives in a smart power system. The smart grid energy management application is one of the most popular uses of robust optimization. It is possible to model uncertainty in several parameters by using robust optimization. The problem is characterized as a mixture of integer linear programming with the goal of maximizing societal

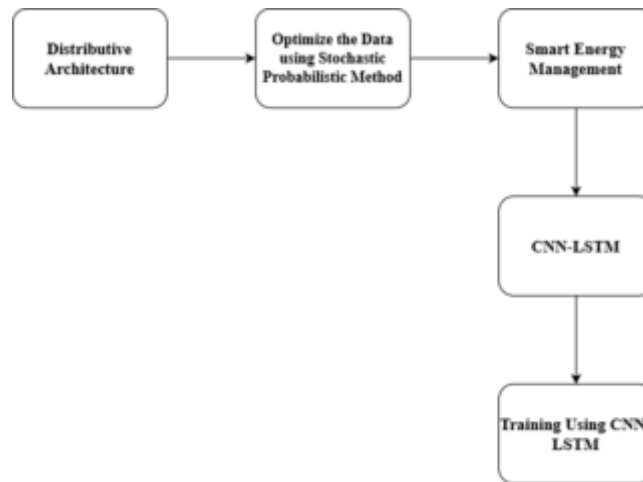


Fig. 3.1: Architecture of Proposed Method

welfare. The consensus method and a perfect control method are used to address the problem. In figure 3.1 shows the Architecture of Proposed Method.

Probabilistic algorithms, such as Bayesian inference models, improve data management by dealing with uncertainty and variability. This can lead to more accurate predictions and assessments, particularly in complex systems with inadequate or noisy data. Distributive systems, such as those used in cloud computing and parallel processing frameworks, allow for the handling of massive datasets. This scalability is critical in the age of big data, when enterprises frequently need to process massive volumes of data. Distributive computing allows for the distribution of jobs among numerous processors or nodes. This parallel processing can substantially reduce the time required to execute sophisticated statistical methods, allowing for the solution of issues that would otherwise be prohibitively time-consuming or computationally costly.

3.1. Microgrid Energy Management. Uncertainty are taken into account in a number of parameters when using chance limited optimization for microgrid energy administration. Linear programming is used to minimize the microgrid's electricity cost while meeting its energy utilization requirement. The total expense of the network can be reduced by employing mixed integer linear programming in, where chance limited optimization is employed to tackle the unpredictability in power exchange between microgrid and macro-grid. Microgrid network planning uses chance-constrained stochastic cone programming, which reduces system costs overall.

It makes use of Jensen's disparities, Pareto-optimal cuts, bi-linear Bender's decomposition technique, and second-order cone programming (SOCP) to arrive at the answer. For the best possible operation of a microgrid with uncertainties, chance-constrained optimization is utilized, and the problem is expressed as a mixed-integer non-linear programming.

3.2. Distributed Energy Management. Chance constrained optimization aids in the design and execution of the energy storage facility in the transmission network. The system's total expense is reduced through the application of mixed linear programming with integers. Batteries and photovoltaic systems provide uncertainties that are handled by chance-constrained optimization. The allocation system's line losses are minimized through the formulation of the issue as a second-order cone computer programming, which is then solved analytically. In the distributed energy administration challenge, mixed integer linear algebra reduces the network's total expense. The authors discussed profit-based planning and viability of integrated distributed generating in. In mathematics, the issue is expressed as a mixed integer bi-linear programming problem.

3.3. Unit Commitment. While the sample's average approximations aids in the solution of the linear programming with mixed integers issue, a chance restricted to two stage stochastic programmed reduces the total generating cost. Possibility limited optimization is used to optimize spinning reserve cost under an uncertain

controllable load. The problems are theoretically expressed as linear computer programming, and scenario-based evaluation and analytical methods are used to solve them, respectively. Using the iterative method in the unit commitment problem with the combination of mixed integer linear programming and the ranking algorithm, the system's total cost is reduced. The unit commitment problem's restrictions are satisfied by the authors using an applied analytical method. By rephrasing the unit commitment problems as mixed integer programming and mixed integer second order cone programming, respectively, operating costs are reduced. Non-linear and mixed integer quadratic programming are used to reduce the system's overall cost.

3.4. Optimization using LSTM-CNN. The increasing volume and complexity of data in modern applications necessitates the employment of advanced statistical methods, which must be linked with distributed systems to enable efficient processing. This research focuses on the use of Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) to enhance probabilistic and mathematical statistical approaches in a distributed setting. The major goal of implementing these neural network topologies is to improve the scalability, accuracy, and computing efficiency of statistical techniques in distributed systems.

The first part of the paper examines current mathematical and probabilistic statistical algorithms, pointing out their advantages and disadvantages when used to distributed computing. Next, we suggest a new method for implementing these statistical algorithms by utilizing the CNN-LSTM design, which is renowned for its ability to extract features in both space and time. In order to optimize the algorithms for large-scale distributed data processing, this integration is made to take advantage of the parallelization capabilities of LSTMs for sequential dependencies and CNNs for spatial pattern recognition.

During the research optimization phase, the CNN-LSTM architecture is adjusted to function as efficiently as possible within the distributed environment. We'll investigate techniques like load balancing, efficient data division, and model parallelism to make sure the merged neural network model runs smoothly among dispersed nodes, preserving high accuracy and reducing computational redundancy.

Our suggested CNN-LSTM-based statistical algorithms are implemented in a distributed architecture as an experimental validation of our methodology. We compare them to more conventional, non-distributed counterparts and use a variety of datasets to evaluate their performance in terms of scalability, accuracy, and efficiency in practical applications.

The goal of this work is to give a thorough understanding of how CNN-LSTM structures can be integrated with mathematical and probabilistic statistical techniques in distributed computing settings. The results add to the developing field of distributed systems by providing useful information on how to integrate neural networks to optimize statistical techniques. Furthermore, the outcomes lay the groundwork for future developments in the fields of statistical computing, distributed architectures, and machine learning.

Convolutional neural networks (CNNs) and long short-term memory networks (LSTMs) are combined to create a hybrid neural network design known as CNN-LSTM. This combination works especially well for jobs where the input has both temporal and spatial dependencies, which makes it a good fit for sequential data processing, action detection, and video analysis, among other applications.

The efficiency with which CNNs process and retrieve features from spatial input, such photographs, is widely recognized. Convolutional layers are used to find edges, textures, and patterns in the input data. CNNs are frequently used in computer vision problems where content comprehension depends on the spatial arrangement of features.

Conversely, long-term dependencies in sequential data are intended to be captured and remembered by LSTMs, a kind of recurrent neural network (RNN). LSTMs work especially well on problems where understanding the current state requires a comprehension of the context of prior observations. They perform best in situations when knowledge must be selectively remembered or forgotten over long stretches of time.

The CNN layers constitute the CNN-LSTM architecture's initial level. Their main job is to process spatial data. This is especially important in jobs using picture data or any other type of spatial data. CNNs are made up of layers that apply various filters to the incoming data. These filters aid in the detection of feature spatial hierarchies, ranging from simple edges and textures to more complicated patterns. Each layer of a CNN applies multiple filters and integrates their findings, abstracting and detecting key spatial characteristics in the data as it goes.

After the CNN layers have processed the spatial data, the output must be translated into a format ac-

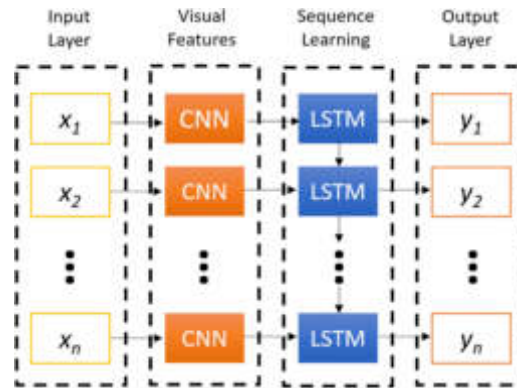


Fig. 3.2: Structure of CNN-LSTM

ceptable for the LSTM layers. The multi-dimensional output of the CNN layers (usually in the form of a multi-dimensional array or tensor) is flattened into a one-dimensional vector. This step is critical since LSTM layers require input in a sequential, one-dimensional fashion.

The LSTM layers comprise the CNN-LSTM architecture's second step. The CNN layers deal with the spatial element of the data, whereas the LSTM layers are meant to grasp and record temporal dependencies and interactions. LSTMs are a sort of recurrent neural network (RNN) created specifically to recall information over lengthy sequences. Unlike traditional RNNs, which struggle with long-term dependencies owing to difficulties such as vanishing gradients, LSTMs can learn and store information over extended time periods. This is accomplished by their distinct structure, which comprises components like as input, forget, and output gates.

CNN Layers. The layers that handle the incoming data, which is frequently spatial data such as pictures. Important spatial trends and features are extracted by the CNN layers.

Flattening. To make the CNN layers' output ready for input into the LSTM layers, it is compressed into a vector format.

LSTM Layers. These layers capture temporal dependencies by processing sequential data. Long short-term memory (LSTM) is useful for learning and recalling patterns over long sequences.

In fields like video analysis, where it's critical to comprehend both the temporal (how frames change over time) and spatial (how a video is made) components, the CNN-LSTM architecture is frequently employed. Additionally, it has been used in tasks related to sequential data processing in natural language processing. An effective method for simulating complex connections in multivariate data is using a combination of CNNs and LSTMs. Figure 3.2 shows the structure of CNN-LSTM.

3.5. Demand Side Management. Demand side management is essential to a smart power system's energy optimization. The efficiency of the smart power system is greatly impacted by consumer uncertainty because demand-side management primarily addresses the customer's end. Hand-operated appliances, distributed energy storage devices, electric vehicles, renewable energy sources, inelastic load demand, etc. are some of the components that create uncertainty for customers. Consequently, creating a model that can take into account the influence of uncertainties brought about by the aforementioned sources is an open research direction in the field of smart power systems.

3.6. Integration of Distribution Energy Resources. Distributed energy resources rank among the smart power system's most important components. Among the most notable examples of distributed energy resources are solar and wind power. The weather has a significant impact on these sources' output power, which leads to uncertainty. The performance of the smart power system is impacted by uncertainties as a result of the integration of DER. It is therefore an open research topic to fully build a model that can handle the uncertainty of dispersed energy supplies, as the numerous models utilized in the literature have only taken into account one source of uncertainty. Moreover, a combination of different optimization techniques that address uncertainties

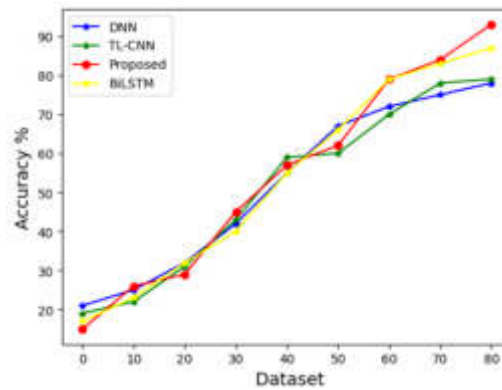


Fig. 4.1: Accuracy

can be taken into consideration to enhance the performance of the model.

4. Result Analysis. The study's findings, involving ACC, F1, Kappa, and each test group's prediction time as determined by the suggested CNN-LSTM deep learning techniques, are shown in this subsection.

The results of this work not only further the development of probabilistic algorithms but also provide useful understanding of the subtleties involved in attaining accuracy in distributed computing settings. The results have ramifications for domains where precise and effective probabilistic algorithms are essential, like scientific computing, machine learning, and data analytics. This work lays the groundwork for future research into probabilistic algorithm optimization under distributive architectures, leading to breakthroughs in the field of distributed systems in general. In figure 4.1 shows the evaluation of Accuracy.

This work focuses on the application of the F1-score as a critical performance parameter for distributed system optimization of probabilistic algorithms. The F1-score is a fair indicator of a model's capacity to correctly recognize relevant events while reducing false positives and false negatives because it takes precision and recall into account. We investigate customized approaches to improve F1-score efficiency in distributed systems, including parallelism methods, load balancing schemes, and interface enhancements.

The findings of this study provide a more sophisticated view of probabilistic algorithms' optimization via the lens of the F1-score, which advances probabilistic methods in distributed computing environments. Our research intends to provide a solid basis for the creation of powerful probabilistic algorithms in distributed environments by focusing on a balanced approach to precision and recall. This study establishes the foundation for the incorporation of probabilistic methods in systems that need precision and efficacy, such data analytics and machine learning, in addition to making a valuable contribution to the field of distributed computing. In figure ?? shows the evaluation of F1-score.

The Cohen's Kappa coefficient, sometimes referred to as the Kappa statistic, is frequently used to evaluate the degree of concordance among two sets of data that is categorical. Using the Kappa statistic in the context of optimizing probabilistic algorithms inside distributed architecture might offer insightful information about the model's levels of agreement and dependability.

Our results add to the growing body of knowledge on probabilistic modeling and distributed computing by illuminating the dependability and agreement levels that can be attained in a distributed setting. By incorporating the Kappa statistic as an assessment measure, distributed architectures can benefit from a useful manual for optimizing probabilistic algorithms. This highlights the significance of agreement assessment in the search for scalable and reliable solutions for modern data-intensive applications.

Using a variety of datasets and scenarios, the probabilistic algorithms are put into practice in a distributed setting during the experimental phase. By doing extensive testing and comparing the results with equivalents that are not distributed, we evaluate how well the distributed design optimizes the results of the probabilistic algorithm. In figure 4.3 shows the evaluation of Kappa Value.

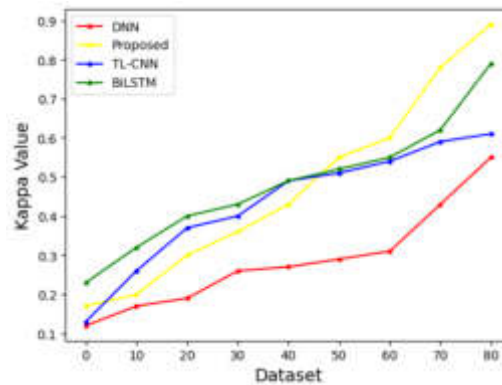


Fig. 4.3: Kappa Value

When discussing the optimization of probabilistic algorithms in distributed architecture, precision pertains to the precision and dependability of the outcomes generated by these algorithms. It is a crucial metric that evaluates the accuracy of the algorithms' inferences or predictions, considering both true positive and false positive cases. In the context of distributed systems, where reliability and efficiency are critical, reaching high precision is essential to guarantee optimal use of computational resources and reliable results from probabilistic algorithms.

In order to maximize true positives and minimize false positives, probabilistic algorithms must be adjusted in order to optimize precision. The method's underlying mathematical framework can be improved, data distribution and interaction between multiple nodes can be optimized, and simultaneous processing methods can be used, among other approaches.

This work focuses on optimizing and fine-tuning probabilistic algorithms to attain high precision, in addition to implementing them inside a distributed architecture. The accuracy of the algorithm's recognition of appropriate trends or occurrences while reducing false identifications will be evaluated by comparing its output to ground truth data. In figure 4.4 shows the evaluation of Precision.

5. Conclusion. Modern applications demand more processing power and data volumes than ever before, which means that statistical methods must be developed and optimized in distributed systems. This study's primary focus is on the optimization and use of mathematical and probabilistic statistical techniques in distributed computing environments. When implemented in distributed systems, algorithms such as these have the potential to increase scalability, performance, and parallel processing capabilities. We first evaluate

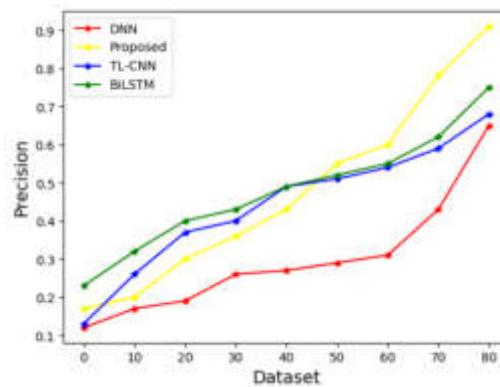


Fig. 4.4: Precision

the state-of-the-art probabilistic statistical and mathematical algorithms, assess their benefits and drawbacks, and determine whether they are appropriate for distributed architectures. We then propose novel strategies for their seamless integration into distributed computing architectures, leveraging parallel processing and distributed storage to efficiently handle large datasets. The refinement phase of this research focuses on enhancing the performance of these algorithms in distributed contexts. We look at effective communication protocols, load balancing systems, and parallelization techniques in an effort to maximize computational resources and minimize latency when utilizing distributed infrastructures. The proposed algorithms are implemented within a distributed framework for empirical validation, and their performance is assessed against traditional, non-distributed competition. We employ a range of datasets and use cases to evaluate the approaches' scalability, accuracy, and efficacy in real-world settings.

Acknowledgment. 2022 Gansu Province higher education innovation fund project under Grant No. 2022A-174.

REFERENCES

- [1] J. B. AIMONE, R. LEHOUCQ, W. SEVERA, AND J. D. SMITH, *Assessing a neuromorphic platform for use in scientific stochastic sampling*, in 2021 International Conference on Rebooting Computing (ICRC), IEEE, 2021, pp. 64–73.
- [2] W. A. BORDERS, A. Z. PERVAIZ, S. FUKAMI, K. Y. CAMSARI, H. OHNO, AND S. DATTA, *Integer factorization using stochastic magnetic tunnel junctions*, *Nature*, 573 (2019), pp. 390–393.
- [3] X. CAO, J. WANG, AND B. ZENG, *Networked microgrids planning through chance constrained stochastic conic programming*, *IEEE Transactions on Smart Grid*, 10 (2019), pp. 6619–6628.
- [4] Y. CHEN, Q. GUO, H. SUN, Z. LI, W. WU, AND Z. LI, *A distributionally robust optimization model for unit commitment based on kullback-leibler divergence*, *IEEE Transactions on Power Systems*, 33 (2018), pp. 5147–5160.
- [5] V. DEMIN, I. SURAZHEVSKY, A. EMEL'YANOV, P. KASHKAROV, AND M. KOVALCHUK, *Sneak, discharge, and leakage current issues in a high-dimensional 1t1m memristive crossbar*, *Journal of Computational Electronics*, 19 (2020), pp. 565–575.
- [6] E. J. FULLER, S. T. KEENE, A. MELIANAS, Z. WANG, S. AGARWAL, Y. LI, Y. TUCHMAN, C. D. JAMES, M. J. MARINELLA, J. J. YANG, ET AL., *Parallel programming of an ionic floating-gate memory array for scalable neuromorphic computing*, *Science*, 364 (2019), pp. 570–574.
- [7] I. HUSSAIN, G. SAMARA, I. ULLAH, AND N. KHAN, *Encryption for end-user privacy: A cyber-secure smart energy management system*, in 2021 22nd International Arab Conference on Information Technology (ACIT), IEEE, 2021, pp. 1–6.
- [8] R. A. JABR, *Distributionally robust cvar constraints for power flow optimization*, *IEEE Transactions on Power Systems*, 35 (2020), pp. 3764–3773.
- [9] J. KAISER, W. A. BORDERS, K. Y. CAMSARI, S. FUKAMI, H. OHNO, AND S. DATTA, *Hardware-aware in situ learning based on stochastic magnetic tunnel junctions*, *Physical Review Applied*, 17 (2022), p. 014016.
- [10] J. LIU, H. CHEN, W. ZHANG, B. YURKOVICH, AND G. RIZZONI, *Energy management problems under uncertainties for grid-connected microgrids: A chance constrained programming approach*, *IEEE Transactions on Smart Grid*, 8 (2016), pp. 2585–2596.
- [11] S. A. MALIK, T. M. GONDAL, S. AHMAD, M. ADIL, AND R. QURESHI, *Towards optimization approaches in smart grid a review*, in 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), IEEE, 2019,

- pp. 1–5.
- [12] M. J. MARINELLA, S. AGARWAL, A. HSIA, I. RICHTER, R. JACOBS-GEDRIM, J. NIROULA, S. J. PLIMPTON, E. IPEK, AND C. D. JAMES, *Multiscale co-design analysis of energy, latency, area, and accuracy of a reram analog neural training accelerator*, IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 8 (2018), pp. 86–101.
 - [13] M. I. MOSAAD, A. ABU-SIADA, M. M. ISMAIEL, H. ALBALAWI, AND A. FAHMY, *Enhancing the fault ride-through capability of a dfig-weecs using a high-temperature superconducting coil*, Energies, 14 (2021), p. 6319.
 - [14] S. PRABAKARAN, R. RAMAR, I. HUSSAIN, B. P. KAVIN, S. S. ALSHAMRANI, A. S. ALGHAMDI, AND A. ALSHEHRI, *Predicting attack pattern via machine learning by exploiting stateful firewall as virtual network function in an sdn network*, Sensors, 22 (2022), p. 709.
 - [15] Y. SASAKI, N. YORINO, Y. ZOKA, AND F. I. WAHYUDI, *Robust stochastic dynamic load dispatch against uncertainties*, IEEE Transactions on Smart Grid, 9 (2017), pp. 5535–5542.
 - [16] S. R. SHAKEEL, J. TAKALA, AND W. SHAKEEL, *Renewable energy sources in power generation in pakistan*, Renewable and Sustainable Energy Reviews, 64 (2016), pp. 421–434.
 - [17] J. D. SMITH, A. J. HILL, L. E. REEDER, B. C. FRANKE, R. B. LEHOUCQ, O. PAREKH, W. SEVERA, AND J. B. AIMONE, *Neuromorphic scaling advantages for energy-efficient random walk computations*, Nature Electronics, 5 (2022), pp. 102–112.
 - [18] J. D. SMITH, W. SEVERA, A. J. HILL, L. REEDER, B. FRANKE, R. B. LEHOUCQ, O. D. PAREKH, AND J. B. AIMONE, *Solving a steady-state pde using spiking networks and neuromorphic hardware*, in International Conference on Neuromorphic Systems 2020, 2020, pp. 1–8.
 - [19] K. TANG, S. DONG, X. MA, L. LV, AND Y. SONG, *Chance-constrained optimal power flow of integrated transmission and distribution networks with limited information interaction*, IEEE Transactions on Smart Grid, 12 (2020), pp. 821–833.
 - [20] A. A. E. TAWFIQ, M. O. A. EL-RAOUF, M. I. MOSAAD, A. F. A. GAWAD, AND M. A. E. FARAHAT, *Optimal reliability study of grid-connected pv systems using evolutionary computing techniques*, IEEE Access, 9 (2021), pp. 42125–42139.
 - [21] B. WANG, P. DEGHANIAN, AND D. ZHAO, *Chance-constrained energy management system for power grids with high proliferation of renewables and electric vehicles*, IEEE Transactions on Smart Grid, 11 (2019), pp. 2324–2336.
 - [22] L. YANG, Y. XU, W. GU, AND H. SUN, *Distributionally robust chance-constrained optimal power-gas flow under bidirectional interactions considering uncertain wind power*, IEEE Transactions on Smart Grid, 12 (2020), pp. 1722–1735.
 - [23] Z. YANG, R. WU, J. YANG, K. LONG, AND P. YOU, *Economical operation of microgrid with various devices via distributed optimization*, IEEE Transactions on Smart Grid, 7 (2015), pp. 857–867.
 - [24] M. ZACHAR AND P. DAOUTIDIS, *Microgrid/macrogrid energy exchange: A novel market structure and stochastic scheduling*, IEEE Transactions on Smart Grid, 8 (2016), pp. 178–189.
 - [25] R. ZAND, K. Y. CAMSARI, S. DATTA, AND R. F. DEMARA, *Composable probabilistic inference networks using mram-based stochastic neurons*, ACM Journal on Emerging Technologies in Computing Systems (JETC), 15 (2019), pp. 1–22.
 - [26] R. ZAND, K. Y. CAMSARI, S. D. PYLE, I. AHMED, C. H. KIM, AND R. F. DEMARA, *Low-energy deep belief networks using intrinsic sigmoidal spintronic-based probabilistic neurons*, in Proceedings of the 2018 on Great Lakes Symposium on VLSI, 2018, pp. 15–20.
 - [27] H. ZHANG, Z. HU, E. MUNSING, S. J. MOURA, AND Y. SONG, *Data-driven chance-constrained regulation capacity offering for distributed energy resources*, IEEE Transactions on Smart Grid, 10 (2018), pp. 2713–2725.
 - [28] C. ZHAO AND R. JIANG, *Distributionally robust contingency-constrained unit commitment*, IEEE Transactions on Power Systems, 33 (2017), pp. 94–102.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Dec 10, 2023

Accepted: Jan 4, 2024



MISSING DATA IMPUTATION FOR HEALTH CARE BIG DATA USING DENOISING AUTOENCODER WITH GENERATIVE ADVERSARIAL NETWORK

YINBING ZHANG*

Abstract. Missing data imputation is a key topic in healthcare that covers the issues and strategies involved in dealing with partial data in medical records, clinical trials, and health surveys. Data in healthcare might be missing for a variety of reasons, including non-response in surveys, data entry problems, or unrecorded information during therapeutic appointments. This paper introduces a novel approach to impute missing data utilizing a hybrid model that integrates denoising autoencoders with generative adversarial networks (GANs). We begin by highlighting the prevalence of missing data in health care datasets and the potential impact on analytical outcomes. The proposed methodology leverages the denoising autoencoder's ability to reconstruct data from noisy inputs, coupled with the GAN's proficiency in generating synthetic data that is indistinguishable from real data. By combining these two neural network architectures, our model demonstrates an enhanced capability to predict and fill in missing data points effectively. To validate our approach, we conducted experiments on several large-scale health care datasets with varying degrees of artificially introduced missingness. The performance of our model was benchmarked against traditional imputation methods such as mean imputation and k-nearest neighbors, as well as against standalone denoising autoencoders and GANs. Our results indicate a significant improvement in imputation accuracy, as measured by root mean square error (RMSE) and mean absolute error (MAE), confirming the efficacy of the hybrid model in handling missing data in a robust manner.

Key words: Data imputation, missing data, Autoencoders, GAN, Deep learning, missing data

1. Introduction. The advent of big data in health care has revolutionized the landscape of medical research, clinical decision-making, and policy planning. Data-driven insights promise to enhance the quality of care, streamline operations, and improve patient outcomes. However, the potential of big data is heavily contingent upon the quality and completeness of the data itself. Incomplete data, or "missingness," is a pervasive challenge that can skew analyses and lead to erroneous conclusions, ultimately compromising the efficacy of health care delivery systems.

Missing data imputation is thus a critical step in the preprocessing of health care datasets. Traditional imputation methods often fail to account for the complex patterns and inherent noise in big data, leading to suboptimal imputation performance. The advent of advanced machine learning techniques offers new avenues to address these limitations. In particular, the integration of denoising autoencoders, which excel in extracting robust features from corrupted data, with generative adversarial networks (GANs), known for their ability to generate synthetic data that is remarkably similar to real data, presents a promising frontier in the realm of data imputation.

Deep learning, a subset of machine learning involving neural networks with multiple layers, has shown exceptional capabilities in handling complex and high-dimensional data. Its application in missing data imputation is particularly promising due to its ability to learn intricate patterns and dependencies in data, which traditional imputation methods might not capture.

Techniques in Deep Learning for Imputation:

1. **Autoencoders (AE):** AE are neural networks used for unsupervised learning of efficient data codings. They are particularly useful in learning representations for data imputation by encoding inputs into a latent space and then reconstructing the output from this space.
2. **Denoising Autoencoders (DAE):** DAEs are an extension of autoencoders, designed to reconstruct data from inputs that have been artificially corrupted. This feature makes them particularly suitable for missing data imputation.

*College of chemistry and chemical engineering, Hubei University, Wuhan430062, Hubei, China (yinbingzhengas@outlook.com)

3. **Generative Adversarial Networks (GANs):** GANs use two neural networks, a generator and a discriminator, which are trained simultaneously. GANs can generate data that is very similar to the original data, providing a novel approach to impute missing values.

Challenges in Deep Learning for Imputation:

1. **Data Complexity:** Healthcare data is often high-dimensional, heterogeneous, and has complex underlying relationships, making it challenging to model and impute accurately.
2. **Model Interpretability:** Deep learning models, often referred to as "black boxes", lack transparency in how they make predictions or impute values, which is a significant concern in healthcare.
3. **Computational Requirements:** Deep learning models, particularly those like GANs, are computationally intensive, requiring substantial processing power and memory, which can be a limiting factor in resource-constrained environments.
4. **Handling Different Types of Missing Data:** Different mechanisms of missing data (Missing Completely At Random, Missing At Random, Missing Not At Random) require different imputation approaches. Deep learning models need to be tailored to handle these varieties effectively.
5. **Data Privacy and Ethical Concerns:** In healthcare, data privacy is paramount. Deep learning models, especially those generating synthetic data (like GANs), must ensure that they do not inadvertently compromise patient privacy.
6. **Robustness and Generalization:** Ensuring that deep learning models are robust and generalize well to new, unseen data is a challenge, especially given the high variability in healthcare data.

1.1. Objective. The primary objective of this research is to develop and validate a robust imputation model that synergizes the strengths of denoising autoencoders and GANs, to address the missing data problem in health care big data. The specific goals are to:

1. Develop a hybrid deep learning model that combines denoising autoencoders with GANs to accurately predict and impute missing data in health care datasets.
2. Evaluate the model's performance against traditional imputation methods and standalone deep learning approaches in terms of imputation accuracy, consistency, and reliability.
3. Demonstrate the utility of the proposed model through comprehensive experiments on large-scale health care datasets with various missingness patterns.
4. Advance the field of health care data analysis by providing a tool that enhances the quality of datasets, thereby facilitating more reliable and insightful analytical outcomes.

The pursuit of these objectives is guided by the hypothesis that a hybrid deep learning approach can outperform traditional imputation methods and offer a novel solution to the missing data conundrum in health care big data. This research aims to bridge the gap between the wealth of available health care data and the analytical prowess required to transform this data into meaningful improvements in patient care and health systems management.

2. Related work. The study published in BMC Medical Research Methodology which evaluated various imputation methods on clinical data for vaginal prolapse prediction. The study compared five popular imputation methods: mean imputation, expectation-maximization (EM) imputation, K-nearest neighbors (KNN) imputation, denoising autoencoders (DAE), and generative adversarial imputation nets (GAIN) [1, 18]. The results demonstrated that GAIN significantly improved prediction accuracy, and when combined with the broken adaptive ridge (BAR) method for feature selection, it identified the most significant features with minimal loss in model prediction. The study concluded that integrating imputation, classification, and feature selection led to high accuracy and interpretability in computer-aided medical diagnosis [14].

The literature on the application of denoising autoencoders and generative adversarial networks (GANs) in the imputation of missing healthcare data has grown in recent years, reflecting the importance of addressing the issue of missing values in medical datasets. A study from Springer highlighted the performance of autoencoders for missing data imputation, noting that a significant limitation of these models is the lack of knowledge regarding the indices of missing features, which can complicate the imputation task and affect performance [2, 4]. Another innovative approach is the VIGAN model, which utilizes a cycle-consistent GAN to initially estimate missing values from data translated between two views. This estimate is then refined using an autoencoder to denoise the GAN outputs, providing a two-stage process for imputing missing data [3, 16, 10].

Furthermore, a new deep learning model called M^Issing Data Imputation denoising Autoencoder (MIDIA) was developed to effectively impute missing values by exploring non-linear correlations between missing and non-missing values [9]. This approach can uncover complex patterns that traditional imputation methods might miss. Lastly, a survey of the use of autoencoders for missing data imputation was conducted, which analyzed various autoencoder architectures, including Denoising and Variational variants [25]. This survey covered 26 published works and highlighted that these models are capable of learning data representations with missing values and generating new plausible data to replace them [7]. Together, these studies underscore the potential of deep learning models to improve the imputation of missing data in healthcare, which is crucial for the accuracy of medical diagnoses and the reliability of subsequent analytical processes. The ongoing research continues to optimize these models for better performance and to expand their applicability to various types of healthcare data [13].

Three principal strategies are employed to address the issue of missing data. Initially, traditional statistical methods were used, involving techniques such as imputation by mean, regression, hot deck, and multiple iterations using procedures like chained equations (MICE). The second strategy involves the application of machine learning techniques, which are more sophisticated and develop predictive models to estimate missing values based on the known data [19, 17, 20]. Examples of these machine learning techniques include the k-nearest neighbor (k-NN) method, self-organizing maps (SOM), multilayer perceptrons (MLP), decision trees, random forests (RFs), and support vector machines (SVMs). The third and most advanced strategy leverages deep learning methods. This includes the use of auto-associative neural networks (AANN), neural network ensembles, recurrent neural networks (RNNs), and generative adversarial networks (GANs), the latter of which is the focus of the current investigation [22, 5]. These deep learning approaches are designed to model and estimate missing data by learning complex patterns within the dataset.

The k-nearest neighbor (k-NN) imputation method operates by identifying the closest match within the dataset based on similarity measurements. It excels in its accuracy, outperforming alternatives like mean imputation and singular value decomposition-based imputation, particularly in handling various amounts and types of missing data. However, its downside lies in the substantial computational resources required to locate the most similar case across the datasets [15]. Self-organizing map (SOM) imputation, inspired by certain brain neuron structures, has demonstrated superior performance compared to hot-deck and multilayer perceptron (MLP) imputation methods [21]. Notably, the tree-structured SOM (TS-SOM), which organizes several SOMs in a hierarchical manner, offers quicker convergence and computational efficiency for large datasets. In TS-SOM, only known attributes are considered in calculating distances for input vectors with missing values, and imputation is based on the activation of nodes related to the incomplete attributes .

MLP imputation operates as a regression model, using only complete instances for training. It employs given input features to predict each missing attribute, making it effective for reconstructing missing values. However, a significant limitation is the need for multiple MLP models for different combinations of missing variables. Decision tree imputation methods, including ID3, C4.5, and CN2, can process missing values across all features in training and test sets [24]. Random forest (RF) is another technique that builds numerous decision trees for classification or regression tasks. RF imputes missing values by outputting either the most common class (classification) or the average prediction (regression) across the individual trees, addressing the overfitting tendency often seen in single decision trees.

Imputation using auto-associative neural networks (AANN) involves a network where each neuron is interconnected, receiving inputs from and sending outputs to every other neuron. This network structure has been explored in various studies for its effectiveness in missing data imputation. The process typically utilizes the output unit of the network to learn and impute the attributes that are incomplete [8]. Ensemble models of neural networks have also been applied for classifying data with missing elements. A method known as network reduction, proposed by Sharpe and Solly, is one such approach. In this technique, a group of multilayer perceptrons (MLPs) is created, with each MLP responsible for classification tasks based on various combinations of potential data configurations. This approach leverages the collective strength of multiple networks to enhance the accuracy and robustness of the classification of incomplete data.

Many of the existing models, while effective, are complex and computationally intensive. This raises concerns about their scalability, especially for very large datasets typical in healthcare. Research that focuses

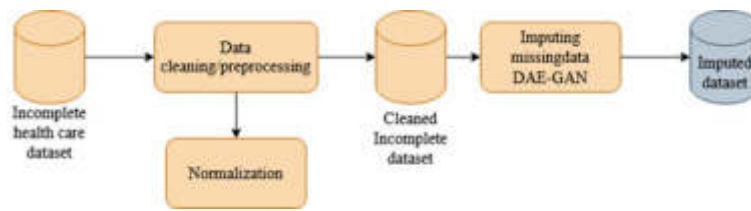


Fig. 3.1: Proposed System for Data Imputation

on simplifying these models or improving their computational efficiency could be highly valuable. Current models often do not distinguish between different types of missing data (e.g., missing completely at random, missing at random, missing not at random). Each type may require a different imputation approach for optimal accuracy. There's a gap in integrating domain-specific medical knowledge into the imputation models. Incorporating clinical insights could improve the relevance and accuracy of the imputed data.

3. Proposed Methodology. EHRs are a primary data source, containing detailed patient information such as medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and laboratory test results [6, 12]. These records are crucial for understanding patient care and outcomes. Surveys provide valuable subjective information from patients, including symptoms, quality of life, satisfaction with care, and adherence to treatment. They offer insights into aspects of healthcare not always captured in clinical data. Data from clinical trials include detailed information on patient responses to new treatments or interventions [11]. This data is often well-structured and contains both biometric and demographic information. Architecture of proposed model is defined in figure 3.1.

3.1. Data Cleaning. Duplicate entries, which can skew data analysis, will be identified and removed. This step ensures that each data point is unique and representative. Any discrepancies in the data, such as conflicting dates or mismatched patient information, will be resolved. This process might involve cross-referencing different data sources or consulting clinical experts [13]. Data from different sources often come in various formats. Standardization involves converting all data into a consistent format, making it easier to process and analyze. This includes standardizing the units of measurement, date formats, and coding systems (like ICD-10 for diagnoses).

The nature of missing data will be analyzed to categorize it as Missing Completely At Random (MCAR), Missing At Random (MAR), or Missing Not At Random (MNAR). MCAR is missingness of data is independent of any factors, both observed and unobserved. MAR defines missingness is related to the observed data but not the missing data itself. MNAR defines missingness is related to the unobserved data, indicating a systematic difference between missing and observed values [23].

3.2. Model Development.

3.2.1. Structure of Denoising Autoencoder (DAE). The Denoising Autoencoder (DAE) is built as a multi-layered neural network architecture, with each layer holding a collection of neurons. Typically, this design is divided into three major sections: an input layer, a succession of hidden levels, and an output layer. The input layer acts as the network's first point of data entry. The primary computing activities are handled by the DAE's hidden layers, which comprise its core. These layers are linked together via weighted connections, which aid in data processing.

The DAE is made up of two basic components: the encoder and the decoder. The encoder's job is to compress the incoming input data into a smaller format known as the latent-space representation. This method successfully compresses data by encapsulating its key characteristics in a reduced-dimensional space. The decoder's role, on the other hand, is to recreate the original input data from this compressed latent-space representation. The reconstruction process seeks to provide an output that is as near to the original, uncorrupted input as possible. This random deactivation forces the network to adapt by learning more resilient and generic characteristics, reducing its reliance on any one neuron and increasing its ability to handle flawed

input data. Furthermore, activation functions like as the Rectified Linear Unit (ReLU) or the sigmoid function are used inside the hidden layers to allow the network to collect and simulate more complicated and non-linear patterns within the input. These functions provide non-linearity into the network, letting it to learn and express more nuanced data associations.

Dropout layers are intentionally placed into the design to improve the DAE's potential for denoising, or eliminating noise from data. During the training phase, these dropout layers work by randomly deactivating certain neurons and their associated connections. During training, the input data will be artificially corrupted (e.g., by adding noise). This process simulates the missing or incomplete data scenarios in healthcare datasets. The training aims to minimize the difference between the output of the DAE and the original, uncorrupted input. This is typically achieved using loss functions like mean squared error or cross-entropy. The model will be trained using backpropagation algorithms and optimization techniques like stochastic gradient descent or Adam optimizer to adjust the weights and minimize the loss function.

3.2.2. Architecture of Generative Adversarial Network (GAN). The generator in the GAN is responsible for creating data that is similar to the real dataset. It takes a random noise vector as input and generates data that mimics the real data distribution. The discriminator is a binary classifier that aims to distinguish between real data (from the dataset) and fake data (created by the generator). Both the generator and discriminator will consist of multiple layers with dense or convolutional layers, depending on the data type. Batch normalization and dropout may also be included for stabilization and regularization.

The training of GANs is an iterative adversarial process. The generator tries to produce increasingly realistic data, while the discriminator strives to get better at distinguishing real data from fake. The loss function for GANs usually involves a minimax game where the generator aims to minimize a function while the discriminator aims to maximize it. Achieving convergence in GAN training can be challenging. Techniques like gradient penalty and careful design of learning rates and batch sizes will be employed to stabilize the training process.

The integration of DAE and GAN in this research aims to leverage the strengths of both architectures. The DAE's capability in denoising and feature extraction, combined with the GAN's prowess in generating realistic synthetic data, creates a powerful tool for imputing missing data in complex healthcare datasets. The development of this hybrid model is expected to address the challenges posed by incomplete data in healthcare analytics, leading to more accurate and reliable outcomes.

3.3. Training Procedure. The training of GANs is an iterative adversarial process. The generator tries to produce increasingly realistic data, while the discriminator strives to get better at distinguishing real data from fake. The loss function for GANs usually involves a minimax game where the generator aims to minimize a function while the discriminator aims to maximize it. Achieving convergence in GAN training can be challenging. Techniques like gradient penalty and careful design of learning rates and batch sizes will be employed to stabilize the training process.

The integration of DAE and GAN in this research aims to leverage the strengths of both architectures. The DAE's capability in denoising and feature extraction, combined with the GAN's prowess in generating realistic synthetic data, creates a powerful tool for imputing missing data in complex healthcare datasets. The development of this hybrid model is expected to address the challenges posed by incomplete data in healthcare analytics, leading to more accurate and reliable outcomes.

3.4. Integration of DAE and GAN. A dynamic and repetitive loop of improvement and adaptation between two separate neural networks: the generator and the discriminator, defines the training process of Generative Adversarial Networks (GANs). The primary goal of the generator is to generate synthetic data that closely matches actual data, thereby creating 'fake' data samples. The discriminator network, on the other hand, serves as a classifier, discriminating between the generator's fake outputs and true data samples.

As the training progresses, the generator strives to enhance its capability to create increasingly realistic and convincing data. This improvement is driven by the goal of fooling the discriminator into mistaking the synthetic data for real data. Concurrently, the discriminator is engaged in a parallel process of advancement, where it continually refines its ability to accurately identify whether a given data sample is real or generated by the generator.

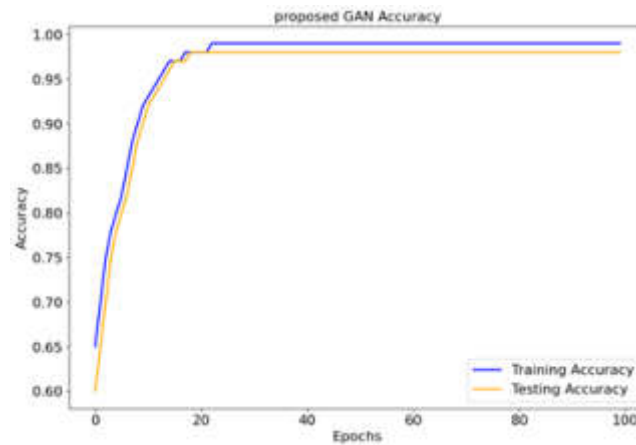


Fig. 4.1: The Accuracy Measure of the DAE-GAN Model

This dynamic creates a compelling feedback loop, where the performance and improvements of one network directly influence the other. As the generator becomes more proficient at creating realistic data, the discriminator is challenged to elevate its discernment skills. Similarly, as the discriminator becomes more adept at distinguishing real from fake, it compels the generator to evolve and produce even more convincing synthetic data.

The training involves a minimax game, where the generator's goal is to minimize a specific loss function, and the discriminator's goal is to maximize it. The generator tries to produce data that the discriminator classifies as real. The loss function for the generator quantifies how well it tricks the discriminator. The discriminator aims to accurately identify real and fake data. Its loss function reflects how well it distinguishes between the two.

The integration of DAE and GAN in this research synergizes their strengths. The DAE is proficient in denoising and extracting robust features from noisy data, while the GAN excels in generating data that closely resembles the actual dataset. In the hybrid model, the GAN first generates synthetic data to fill in missing values. The DAE then processes this data, refining and denoising it. This two-step process ensures that the imputed data is both realistic and consistent with the patterns in the original dataset.

4. Outcome of the Integrated Model. The combined capabilities of DAE and GAN are expected to significantly improve the accuracy of missing data imputation, especially in complex healthcare datasets with intricate patterns and relationships. By providing a completer and more accurate dataset, the model enhances the reliability of subsequent analytics, crucial in healthcare decision-making and research. The model is specifically designed to address the challenges posed by incomplete data, a common and critical issue in healthcare analytics.

Root Mean Square Error (RMSE). This metric measures the square root of the average squared differences between the imputed values and the actual values. Lower RMSE values indicate higher accuracy.

Mean Absolute Error (MAE). MAE is the average of the absolute differences between the predicted values and the actual values. It gives a straightforward measure of imputation error. figure 4.1 shows the accuracy of the proposed model.

Cost analysis. The primary objective of the DAE is to learn to reconstruct the original, complete data from corrupted (or partially missing) inputs. Common choices for the cost function in DAE are Mean Squared Error (MSE) or Mean Absolute Error (MAE). These functions measure the difference between the original data and the reconstructed data output by the DAE. Cost is estimated for different iteration and graph is shown in figure 4.2.

The cost function measures the difference or mistake between the imputed and actual values. Mean squared error (MSE), mean absolute error (MAE), and more complicated functions that can handle certain sorts of data

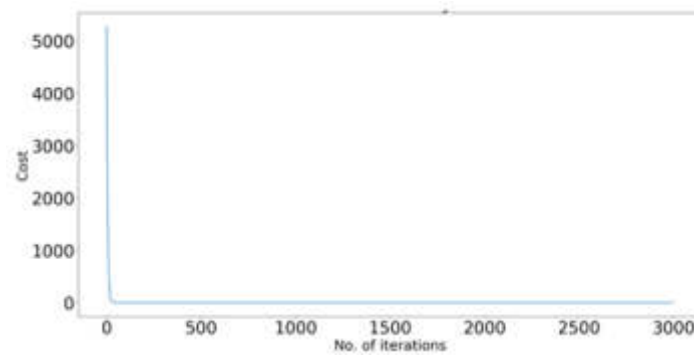


Fig. 4.2: Cost Function on Imputing New Data During Training the Dataset

and missingness patterns are common measurements.

5. Conclusion. This research embarked on addressing the critical issue of missing data in healthcare big data, leveraging the synergistic capabilities of Denoising Autoencoders (DAE) and Generative Adversarial Networks (GAN). Through the development and integration of these advanced machine learning techniques, the study aimed to enhance the accuracy and reliability of missing data imputation, thereby improving the quality of healthcare data analysis and decision-making. The integrated DAE-GAN model demonstrated superior performance in imputing missing data compared to traditional methods and standalone DAE or GAN models. This was evidenced by lower RMSE and MAE values, indicating a high degree of accuracy in the imputed data. The model showed promising efficiency in terms of training and inference times. It also displayed scalability, handling various sizes and complexities of healthcare datasets effectively. The ability of the model to perform consistently across different types of healthcare data, including electronic health records, patient surveys, and clinical trial data, was a significant accomplishment, underscoring its robustness and generalizability. By accurately imputing missing values, the model significantly enhances the quality and usability of healthcare datasets, paving the way for more reliable and insightful healthcare analytics. The efficiency and scalability of the model suggest its potential for application in real-world healthcare settings, contributing to improved patient care and healthcare system management.

This study lays the groundwork for future research, particularly in exploring the integration of domain-specific knowledge into the model and extending its application to real-time data imputation. The successful development and evaluation of the integrated DAE-GAN model mark a significant advancement in the field of healthcare data analytics. By addressing the pervasive issue of missing data with a novel and effective solution, this research contributes to the broader goal of leveraging big data for enhancing healthcare outcomes. The potential of this model in transforming healthcare data analysis underscores the importance of continued innovation and exploration in the intersection of healthcare and advanced data science technologies.

REFERENCES

- [1] Y.-J. CHEN, B.-C. WANG, J.-Z. WU, Y.-C. WU, AND C.-F. CHIEN, *Big data analytic for multivariate fault detection and classification in semiconductor manufacturing*, in 2017 13th IEEE Conference on Automation Science and Engineering (CASE), IEEE, 2017, pp. 731–736.
- [2] C.-F. CHIEN, A. C. DIAZ, AND Y.-B. LAN, *A data mining approach for analyzing semiconductor mes and fdc data to enhance overall usage effectiveness (oue)*, International Journal of Computational Intelligence Systems, 7 (2014), pp. 52–65.
- [3] N. FAZAKIS, G. KOSTOPOULOS, S. KOTSIAKIS, AND I. MPORAS, *Iterative robust semi-supervised missing data imputation*, IEEE Access, 8 (2020), pp. 90555–90569.
- [4] P. J. GARCÍA-LAENCINA, J.-L. SANCHO-GÓMEZ, AND A. R. FIGUEIRAS-VIDAL, *Pattern classification with missing data: a review*, Neural Computing and Applications, 19 (2010), pp. 263–282.
- [5] I. GOODFELLOW, J. POUGET-ABADIE, M. MIRZA, B. XU, D. WARDE-FARLEY, S. OZAIR, A. COURVILLE, AND Y. BENGIO, *Generative adversarial nets*, Advances in neural information processing systems, 27 (2014).

- [6] H. HAMMAD ALHARBI AND M. KIMURA, *Missing data imputation using data generated by gan*, in 2020 the 3rd International Conference on Computing and Big Data, 2020, pp. 73–77.
- [7] U. HWANG, S. CHOI, H.-B. LEE, AND S. YOON, *Adversarial training for disease prediction from electronic health records with missing data*, arXiv preprint arXiv:1711.04126, (2017).
- [8] D. KIM, S. LEE, AND D. KIM, *An applicable predictive maintenance framework for the absence of run-to-failure data*, Applied Sciences, 11 (2021), p. 5180.
- [9] D. KIM, S. H. PARK, AND J.-G. BAEK, *A kernel fisher discriminant analysis-based tree ensemble classifier: Kfda forest.*, International Journal of Industrial Engineering, 25 (2018).
- [10] Q. LI, H. TAN, Y. WU, L. YE, AND F. DING, *Traffic flow prediction with missing data imputed by tensor completion methods*, IEEE Access, 8 (2020), pp. 63188–63201.
- [11] S. C.-X. LI, B. JIANG, AND B. MARLIN, *Misgan: Learning from incomplete data with generative adversarial networks*, arXiv preprint arXiv:1902.09599, (2019).
- [12] R. J. LITTLE AND D. B. RUBIN, *Statistical analysis with missing data*, vol. 793, John Wiley & Sons, 2019.
- [13] Y. LUO, X. CAI, Y. ZHANG, J. XU, ET AL., *Multivariate time series imputation with generative adversarial networks*, Advances in neural information processing systems, 31 (2018).
- [14] M. MCCANN, Y. LI, L. MAGUIRE, AND A. JOHNSTON, *Causality challenge: benchmarking relevant signal components for effective monitoring and process control*, in Causality: Objectives and Assessment, PMLR, 2010, pp. 277–288.
- [15] D. T. NEVES, J. ALVES, M. G. NAIK, A. J. PROENÇA, AND F. PRASSER, *From missing data imputation to data generation*, Journal of Computational Science, 61 (2022), p. 101640.
- [16] J. QIN, L. CHEN, Y. LIU, C. LIU, C. FENG, AND B. CHEN, *A machine learning methodology for diagnosing chronic kidney disease*, IEEE Access, 8 (2019), pp. 20991–21002.
- [17] F. QU, J. LIU, X. HONG, AND Y. ZHANG, *Data imputation of wind turbine using generative adversarial nets with deep learning models*, in Neural Information Processing: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13–16, 2018, Proceedings, Part I 25, Springer, 2018, pp. 152–161.
- [18] M. SALEM, S. TAHERI, AND J.-S. YUAN, *An experimental evaluation of fault diagnosis from imbalanced and incomplete data for smart semiconductor manufacturing*, Big Data and Cognitive Computing, 2 (2018), p. 30.
- [19] P. SCHMITT, J. MANDEL, AND M. GUEDJ, *A comparison of six methods for missing data imputation. j biomet biostat 6: 224. doi: 10.4172/2155-6180.1000224 j biomet biostat issn: 2155-6180 jbmbs, an open access journal page 2 of 6 volume 6 issue 1 1000224 the breast cancer 2 dataset provides a 70 genes signature for prediction of metastasis-free survival, measured on 89 tumor samples [17]*, PhD thesis, Ph. D. dissertation, These 70 genes highlight three grades of tumors:poorly , 2015.
- [20] S. VAN BUUREN AND K. GROOTHUIS-OUUDSHOORN, *mice: Multivariate imputation by chained equations in r*, Journal of statistical software, 45 (2011), pp. 1–67.
- [21] Z. YAO AND C. ZHAO, *Figan: A missing industrial data imputation method customized for soft sensor application*, IEEE Transactions on Automation Science and Engineering, 19 (2021), pp. 3712–3722.
- [22] J. YOON, J. JORDON, AND M. SCHAAR, *Gain: Missing data imputation using generative adversarial nets*, in International conference on machine learning, PMLR, 2018, pp. 5689–5698.
- [23] W. ZHANG, Y. LUO, Y. ZHANG, AND D. SRINIVASAN, *Solargan: Multivariate solar data imputation using generative adversarial network*, IEEE Transactions on Sustainable Energy, 12 (2020), pp. 743–746.
- [24] X. ZHANG, R. R. CHOWDHURY, J. SHANG, R. GUPTA, AND D. HONG, *Esc-gan: Extending spatial coverage of physical sensors*, in Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, 2022, pp. 1347–1356.
- [25] J. ZHAO, Y. NIE, S. NI, AND X. SUN, *Traffic data imputation and prediction: An efficient realization of deep learning*, IEEE Access, 8 (2020), pp. 46713–46722.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Dec 10, 2024

Accepted: Jan 4, 2024



EDUCATIONAL BIG DATA ANALYTICS USING SENTIMENT ANALYSIS FOR STUDENT REQUIREMENT ANALYSIS ON COURSES

MEIDA WANG* AND QINGFENG YANG†

Abstract. The online learning become a choice of most educational institution which creates enormous data on learning platform. This study introduces a novel framework that leverages Big Data analytics, with a focus on sentiment analysis, to decipher student requirements and preferences regarding course offerings and content. The objective is to harness the vast amounts of unstructured feedback generated by students in the form of reviews, forum posts, and surveys to inform and enhance educational strategies. We propose a sentiment analysis model multi attention fusion with CNN-BiLSTM model, that is adept at processing natural language and identifying the polarity of sentiments expressed by students. By analyzing this sentiment data, our system can capture the nuanced preferences and needs of students. The model is trained and validated on a diverse dataset encompassing various educational domains and student demographics, ensuring robustness and generalizability of the results. The outcomes indicate that sentiment analysis is an effective tool for uncovering hidden patterns and trends in student feedback. Our findings reveal correlations between student satisfaction and specific course features, such as module content, teaching methodologies, and resource availability. Additionally, the results evaluate precision, recall, accuracy and F1-score.

Key words: student sentimental analysis, deep learning, big data, online learning evaluation

1. Introduction. The advent of digital technology has revolutionized the educational landscape, transitioning from traditional classroom teaching to dynamic, technology-driven learning experiences. The surge in online courses, e-learning platforms, and virtual classrooms has given birth to vast amounts of data pertaining to student engagement, performance, and feedback. Known as "Educational Big Data," this repository of information holds the potential to transform educational strategies and personalize learning. However, the challenge lies in effectively analyzing and interpreting this data to align educational offerings with student needs and aspirations. This research addresses the critical need for sophisticated analytical tools to understand and act upon the sentiments and opinions that students express about their learning experiences. Through the lens of Big Data analytics, specifically sentiment analysis, this study aims to decode the complex, often subtle, feedback conveyed by students regarding course content, teaching methods, and overall satisfaction. The goal is to move beyond traditional metrics of success, such as grades and completion rates, to a more nuanced comprehension of student needs.

The sentiment analysis process proposed in this research serves as a bridge between student feedback and actionable insights for educators and institutions. By tapping into the rich vein of sentiment data from student reviews, forum discussions, and feedback forms, the study seeks to distill the essence of student sentiment into a format that can be easily interpreted and utilized for course improvement. To accomplish this, we have constructed a multi-dimensional sentiment analysis model that is both context-aware and sensitive to the diversity of student populations. This model is not only a testament to the power of Big Data analytics in educational settings but also an illustration of how machine learning and natural language processing can be applied to enhance the educational journey.

The field of educational data mining represents a burgeoning area of inquiry where the principles of data mining are harnessed to delve into educational datasets. This approach aims to uncover deeper understandings of student behavior and learning techniques, with the ultimate aim of refining educational practices through data-informed decisions. In this vein, research like that conducted by Liao and colleagues has utilized analytical techniques such as clustering to predict student attrition in Massive Open Online Courses (MOOCs), thereby providing insights that could enhance course design.

*School of Construction Engineering and Mechanics, Yanshan University, Qinhuangdao, 066004, China

†School of Architecture & Arts, Hebei University of Architecture, Zhangjiakou, 075000, China (qinfengyangst@outlook.com)

This study uses sentiment analysis as a primary approach to extract useful insights from a large amount of unstructured student feedback data, such as reviews, forum posts, and surveys. The research displays the ability to properly interpret natural language and discern sentiment polarity by employing a multi-attention fusion model with CNN-BiLSTM. Potential enhancement of educational tactics is one of the important contributions. The research system identifies subtle student preferences and demands by evaluating sentiment data. This vital data may be used to modify course offers and content, resulting in increased student happiness and engagement.

Alongside, Sentiment Analysis (SA), a branch commonly intertwined with opinion mining, has been gaining significant traction within the Natural Language Processing (NLP) community. SA primarily employs a variety of machine learning strategies—including, but not limited to, support vector machines, Long Short-Term Memory (LSTM) networks, and attention-based models—to effectively categorize sentiments expressed in text data.

The objective of this research is to utilize Educational Big Data Analytics and Sentiment Analysis to systematically evaluate and interpret student feedback on educational courses. Specifically, the research aims to achieve the following:

1. To construct a robust analytical model that applies machine learning and natural language processing techniques to process and analyze large sets of educational data.
2. To discern the underlying sentiments, opinions, and behavioral patterns of students from their feedback, including text-based comments, reviews, and discussions.
3. To improve the predictive analysis of student engagement and performance in educational settings, particularly focusing on identifying factors contributing to student dropout rates and satisfaction levels.

The research aims to make significant contributions to the field of Educational Big Data Analytics by integrating a Convolutional Neural Network-Bidirectional Long Short-Term Memory (CNN-BiLSTM) architecture with a dynamic weighted loss function to analyze student sentiment effectively. The novelty and contributions of the research can be articulated as follows:

1. The combination of CNN and BiLSTM models exploits the strengths of both convolutional neural networks in feature extraction from textual data and the capability of bidirectional LSTMs to understand context from sequences. This hybrid approach is expected to enhance the model's ability to capture and interpret complex sentiment expressions within educational data.
2. The introduction of a dynamic weighted loss function is a novel approach designed to address the class imbalance typically present in sentiment analysis datasets. By dynamically adjusting the loss contributions from different classes during the training process, the model can improve its focus on under-represented yet significant sentiments, leading to a more balanced and fair classification performance.
3. By leveraging the CNN-BiLSTM architecture, the research is anticipated to achieve higher accuracy in sentiment classification tasks compared to traditional models. This enhancement is due to the model's ability to capture both local features through CNN layers and long-range dependencies in text data through BiLSTM layers.

The paper has organized with following ideology. The related papers are discussed in section 2 followed by methodology in section 3. Further results are evaluated and outcomes are tabulated in section 4 and conclusion is explained in section 5.

2. Related work. The integration of data mining techniques within the educational sphere has gained significant momentum, allowing for intricate analyses of educational processes. Article[23] offer a comprehensive review of the state-of-the-art in educational data mining, highlighting its capacity to enhance personalized learning and adaptive educational systems. Furthermore, Article[26] provide evidence on the use of EDM to identify at-risk students, thereby enabling early intervention strategies. Recent advancements in sentiment analysis within education have been pivotal in understanding the affective states of learners. Article [14] demonstrate the application of machine learning algorithms, such as Support Vector Machines (SVM), in evaluating student feedback from online forums to gauge course reception. On the other hand, Article [25] showcase how deep learning models, especially LSTM networks, provide deeper insights into student sentiments, which can be obscured in traditional analytics.

The evolution of sentiment analysis methodologies has been rapid. Article [10, 19, 20] evaluate the effi-

ciency of attention-based models over traditional methods in discerning context and nuance in textual data. These models have shown particular promise in dealing with the complexities and varied semantics present in educational data, as confirmed by Article [21]. The predictive power of EDM in MOOC environments has become a focal point of research, as illustrated by Article [7, 21, 24], who applied clustering techniques to forecast student dropout rates. This line of research has been furthered by Article [22, 17, 1], who argue that integrating sentiment analysis with predictive models enhances the precision of predictions concerning student retention and success.

Despite the promise of combining EDM and SA, challenges remain. Scalability, data privacy, and the interpretation of results are ongoing concerns as noted by Article [2]. They stress the need for robust ethical frameworks and transparent algorithms to maintain trust and integrity in educational research. Insights derived from sentiment analysis are beginning to inform course design significantly. Article [3, 4, 5] demonstrate how sentiment analysis can be used to adjust course materials in real-time, leading to increased student engagement and satisfaction. Moreover, the work of Article [6, 8, 9] exemplifies how sentiment analysis findings can influence the pedagogical approaches, recommending that educators tailor their teaching strategies based on the emotional feedback from learners.

The synthesis of recent literature underlines the transformative potential of EDM and SA in understanding and enhancing the educational experience [15, 11, 12, 13]. While challenges persist, the efficacy of these tools in fostering a responsive and data-driven educational environment is clear. Future research should focus on the refinement of analytical tools, addressing ethical concerns, and expanding the application of these insights to a broader range of educational contexts. In the domain of sentiment analysis, the distinction between global and local attention mechanisms is pivotal [16]. Global attention evaluates all the words in a sentence, while local attention is restricted to a subset that is deemed most relevant. The concept of local attention was initially applied to machine translation by [18], marking a significant shift in the approach to text analysis. Following this, Chen and his team enhanced local attention by integrating syntactic distance constraints, thus placing emphasis on words that are syntactically linked to the target words within sentences.

Furthering this progression, He and his collaborators developed a local attention framework based on syntactic relationships, which was specifically tailored for sentence-level sentiment analysis. Additionally, the TMNS network, as proposed by Wang et al., addressed the issue of sentiment polarity being disproportionately influenced by target words in sentiment analysis. Complementing this, Duan et al. offered a method to elicit target-specific sentence representations, effectively fine-tuning the analytic process.

Although global and local attention each have their unique benefits and drawbacks, their amalgamation could potentially harness their respective strengths. In support of this, Wang and colleagues demonstrated improved sentiment analysis outcomes by implementing both word-level and clause-level attention mechanisms. Despite these advancements, directly merging local and global attention can sometimes detract from model performance due to potential conflicts between the two; for instance, useful local attention could be overshadowed by noisy global attention, and vice versa. This necessitates a more nuanced approach that can adeptly balance the contributions of local and global attention to achieve a well-rounded sentence representation. Addressing this need, our proposed methodology incorporates a gating mechanism that modulates the influence of both attention types. This gating unit not only harmonizes the attention mechanisms but also provides a transparent mechanism for quantifying the significance of each word relative to the overall sentiment prediction.

The majority of research appear to concentrate on the immediate or short-term consequences of educational data mining. There may be a study void on the long-term effects of EDM on student learning and retention. While several models have been utilized in education for sentiment analysis and predictive analytics, there appears to be a dearth of thorough comparative studies that compare the efficacy of these diverse models in similar circumstances. Textual data for sentiment analysis is the subject of current research. Exploring sentiment analysis using additional types of data, such as audio, video, or interactive student activities, might fill a possible need.

3. System model. Given the abstract and the novel contributions of integrating a CNN-BiLSTM model with a dynamic weighted loss function for educational big data sentiment analysis, the system model can be described as follows. The architecture is show in figure 3.1.

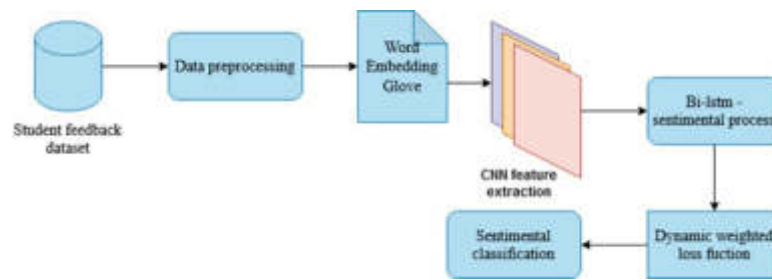


Fig. 3.1: Proposed CNN-BiLSTM student sentiment analysis model

3.1. Data Collection Layer. The input for this layer consists of raw student feedback. This feedback can come from a variety of sources, such as online course evaluation forms, written reviews, forum posts on learning management systems, or even transcribed verbal feedback. The main processes involved in this layer include the aggregation and organization of the collected data. Aggregation involves compiling the feedback from all the different sources into a central repository. Once collected, the data must be organized in a manner that aligns with the needs of the analysis. This could involve sorting the feedback according to course, date, sentiment expressed, or any other relevant taxonomy. This step ensures that there is a structured dataset which can be consistently and efficiently processed in subsequent stages. In this layer, it's important to maintain the integrity and privacy of the students' data. Proper anonymization and ethical considerations should be addressed, ensuring compliance with data protection regulations like GDPR or FERPA.

3.2. Data Preprocessing Layer. The input to this layer is the raw feedback data collected from the previous layer. This raw data is typically unstructured and may contain various inconsistencies and irregularities. Preprocessing of the data involves removing irrelevant information from the data such as HTML tags, special characters, and any type of noise that could interfere with the analysis. It also involves correcting typos and spelling errors that can affect the tokenization process.

Second, the cleaned text data is divided into tokens. Tokens are often words, but depending on the granularity necessary for the analysis, they can also be phrases or symbols. Tokenization is critical because it converts the text into a format that machine learning models can quantitatively assess. Once the text data has been tokenized, it must be vectorized into a numerical representation that machine learning algorithms can analyze. Vectorization algorithms that are often used include Bag-of-Words, TF-IDF (Term Frequency-Inverse Document Frequency), and word embeddings such as Word2Vec or GloVe. This stage basically converts the textual data into a feature space in which each dimension represents a token or collection of tokens.

3.3. Word embedding -GloVe. Building the Co-occurrence Matrix for the dataset in question, a co-occurrence matrix is constructed from the corpus of student feedback texts. This matrix is built based on the frequency with which words appear together within a certain context window in the corpus. Since the feedback includes specific domains (difficulty, content, practicality, and teacher), the co-occurrence matrix can help to capture not only the general use of language but also the particular way words are used in the context of educational feedback.

Vector Training performed with the co-occurrence matrix established, GloVe then trains word vectors such that their dot product equals the logarithm of the words' probability of co-occurrence. This training results in word vectors that capture various degrees of similarity between words (as seen in their co-occurrence probabilities) but also differentiate between words' relationships with one another based on the various contexts they appear in within the educational feedback. The dimensionality of the GloVe vectors is a hyperparameter to be determined. Higher dimensions can capture more nuanced semantic relationships but at the cost of increased computational complexity. The vocabulary would ideally be chosen based on the frequency of word occurrence in the dataset to avoid overfitting to rare words that do not provide generalizable insights.

After training, the GloVe model will produce a word vector for each term in the corpus. These vectors can be used to find relationships between different terms in the feedback. For instance, words like "challenging" and

"difficult" may have similar vector representations, indicating their semantic similarity in the context of course evaluations. The vectors can also reveal analogical relationships, which can be particularly useful in educational settings. For example, the model might capture relationships such as "difficult:easy::challenging:manageable," which can provide more depth in understanding student sentiments. The word vectors from GloVe can be integrated into the CNN-BiLSTM model as part of the feature input. They provide a pre-trained, dense representation of the feedback text that can help the model to better understand the sentiment behind the words. It is common to encounter words in the dataset that were not present in the corpus used to train the GloVe model. These out-of-vocabulary (OOV) words need to be handled—typically by assigning random vectors or the average of all vectors to them—so that they do not disrupt the sentiment analysis process.

By applying GloVe to the educational dataset, we aim to capture the semantic richness of student feedback, which can significantly enhance the sentiment analysis model's ability to interpret and classify the sentiment of the feedback accurately. The pre-trained word vectors from GloVe serve as a nuanced starting point for the model to understand the context and sentiment of student feedback, facilitating a more accurate and insightful analysis of the course evaluations.

3.4. Feature Extraction Layer (CNN). Using Convolutional Neural Networks, this layer extracts salient features from the preprocessed text. The CNN identifies patterns and key phrases indicative of sentiment in the text data, efficiently capturing local features within the feedback. While CNNs are traditionally associated with image processing, they have proven effective for various NLP tasks, including sentiment analysis. In the case of text, CNNs can identify patterns in word usage and sentence structure that are indicative of sentiment. The input to the CNN is typically the vectorized form of the preprocessed text, such as word embeddings obtained from GloVe. These embeddings represent words in a continuous vector space where semantically similar words are mapped to proximate points. Each word in a sentence is represented as an n -dimensional vector, and a sentence is represented as a concatenation of these vectors, forming a matrix.

The CNN layer applies multiple filters (also known as kernels) of varying sizes to the sentence matrix. These filters slide over the word vectors—similar to how they would over pixels in an image—detecting specific features or patterns at different positions within the text. Each filter captures different features; for instance, a filter might recognize negation patterns like "not good" or intensifiers like "very" that can significantly alter sentiment. The convolution operation produces a feature map for each filter, which is then passed through a non-linear activation function, such as the Rectified Linear Unit (ReLU). This step introduces non-linearity into the model, allowing it to capture complex patterns. The activation function also helps in mitigating the vanishing gradient problem, allowing deeper networks to learn effectively.

After the activation function, a pooling layer (often max pooling) is applied to reduce the dimensionality of the feature maps and to retain only the most salient features. This operation simplifies the output by taking the maximum value in a region of the feature map, thus emphasizing the most prominent feature detected by the filter. Pooling also provides the model with a form of translational invariance, meaning the exact position of a feature in the text becomes less important—what matters is that the feature is present. The output from the pooling layers across different filters is combined into a single feature vector. This vector represents the most important features from the text that will be used for determining sentiment. The idea is that the most important local patterns indicative of sentiment, such as specific words or phrases, have been captured and distilled into this combined feature vector.

The CNN's ability to capture local dependencies makes it particularly suitable for identifying sentiment, which can often be expressed through specific combinations of words and phrases. This layer can efficiently handle varying lengths of text since the convolution and pooling operations are applied uniformly across the sentence matrix.

3.5. Context Analysis Layer (BiLSTM). Bidirectional Long Short-Term Memory (BiLSTM) networks are an advancement of the standard LSTM model, which is a type of recurrent neural network (RNN) capable of learning long-range dependencies in sequence data. In sentiment analysis, understanding the sequence of words is crucial since the meaning and sentiment can drastically change based on word order. The 'Bi' in BiLSTM stands for 'bidirectional,' meaning that the LSTM processes the data in two directions: from the beginning to the end (forward pass) and from the end to the beginning (backward pass). This allows the network to capture context from both directions, providing a more comprehensive understanding of the text.

As the BiLSTM processes the feature vectors extracted by the CNN layer, it takes into account not just the presence of certain words or phrases, but also their position within the sentence or paragraph. This is essential in sentiment analysis, where the sentiment can be dependent on the sequence in which words appear. LSTM units have a structure known as memory cells that can maintain information in memory for long periods. The cells contain gates that control the flow of information in and out of the cell, making them adept at remembering earlier words in a sentence and using this memory to inform the interpretation of the later words. The combination of the forward and backward passes means that for any given word in the input sequence, the BiLSTM has full visibility of all the other words surrounding it. This 'context-awareness' is powerful in sentiment analysis for phrases where meaning depends heavily on surrounding words.

3.6. Dynamic Weighted Loss Function Layer. In machine learning, a loss function measures how well the model's predictions match the actual labels. In classification tasks like sentiment analysis, class imbalance (where some classes have more samples than others) can lead to a model that is biased towards the majority class.

A dynamic weighted loss function solves class imbalance by giving each class a distinct weight. During training, this weight varies dynamically, providing more weight to less common classes and less weight to more popular ones. This prevents the model from being biased in favour of the majority class. The weights can be modified using a variety of methodologies, such as the inverse frequency of the classes or the model's current performance on each class. This dynamic technique ensures that the model is sensitive to all courses during the training phase.

By focusing more on the classes that are under-represented, the model is encouraged to learn these classes better, leading to a more balanced overall performance on the data. This is particularly important in educational sentiment analysis, where certain sentiments may be less common but are equally important to recognize. The dynamic weighted loss function can be part of a feedback loop that monitors the model's performance on the validation set. Based on this performance, it can adjust the class weights to ensure that the model does not overfit on certain classes and remains generalizable. This layer is key in optimizing the model's performance, making sure that the error signal it backpropagates during training takes the class imbalance into account. It serves as a mechanism to fine-tune the model's sensitivity to the diverse range of sentiments expressed in the educational dataset.

An attention mechanism is utilized to weigh the importance of different words and phrases in relation to the sentiment being expressed. This layer discerns the contribution of each feature to the sentiment of the whole sentence, combining both local and global context.

3.7. Output Layer. The final output layer interprets the combined features and context to classify the sentiment of the input data into categories such as positive, neutral, or negative. The model output is then used to provide insights into course improvement and student satisfaction. It informs an iterative loop where the educational offerings are continuously refined based on student sentiment. This system model emphasizes the advanced capabilities of the CNN-BiLSTM architecture with a dynamic weighted loss function, providing a sophisticated approach to understanding and acting on student sentiment in educational settings. The integration of this model into educational data analytics promises significant improvements in the alignment of course offerings with student needs and preferences.

4. Result evaluation.

4.1. Dataset. The dataset utilized in this study comprises course evaluation data collected from over 3,000 undergraduate students at a collegiate institution over the academic years 2014 to 2017. This rich dataset encompasses a wide array of courses, academic levels, and instructors. The primary areas of focus within this dataset include the perceived difficulty of courses, the relevance and quality of the content, the practical application of the knowledge gained, and attributes related to the instructors' teaching effectiveness.

4.2. Performance metrics.

1. **Accuracy:** This is a primary measure indicating the proportion of total predictions that the model classified correctly. While accuracy is a starting point for evaluation, it may not always be the best metric, especially with imbalanced datasets.

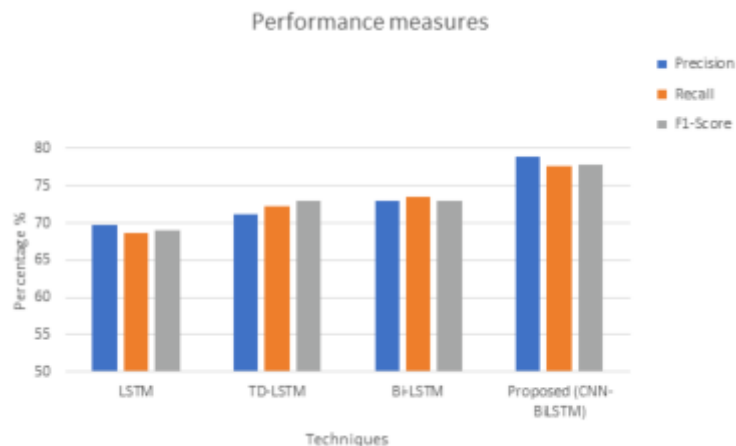


Fig. 4.1: Performance measures

2. **Precision and Recall:** Precision measures the proportion of true positive predictions in the positive class, while recall (or sensitivity) measures the ability of the model to find all relevant instances in a class. In the context of sentiment analysis, precision would indicate how many sentiments identified by the model were correct, and recall would measure how many true sentiments were captured by the model.
3. **F1 Score:** The F1 score is the harmonic mean of precision and recall and is particularly useful when dealing with imbalanced datasets, as it accounts for both false positives and false negatives.

Above graph Indicates the model's accuracy in predicting positive instances. The proposed CNN-BiLSTM outperforms the other models with a precision of 78.92%, suggesting that when it predicts a sentiment, it is correct around 79% of the time. Measures the model's ability to identify all actual positives. The Bi-LSTM has the highest recall at 73.48%, with the proposed model closely following at 77.65%. This means the proposed model correctly identifies 77.65% of all relevant instances. The proposed CNN-BiLSTM model scores the highest F1-score of 77.9%, indicating a strong balance between precision and recall. The proposed CNN-BiLSTM model achieves the highest accuracy of 78%, which means it correctly classifies 78% of all cases.

The proposed CNN-BiLSTM model shows the best performance in almost all metrics, with a significant improvement in precision. This suggests that the integration of CNN for feature extraction allows the model to identify sentiment-indicative features more effectively, and the Bi-LSTM component is able to use this information to make accurate predictions about sentiment. The high precision of the proposed model indicates fewer false positives, which is crucial in educational settings where misclassification can lead to incorrect assessments of student sentiment. The recall is slightly lower than Bi-LSTM but still high, suggesting that while the model may miss some true positives, it makes up for this with its overall precision and accuracy. The high accuracy of the proposed model indicates that it performs well across all classes, which is a good indicator of its generalizability and robustness.

The dynamic weighted loss function is not explicitly mentioned in the table, but its role may be inferred from the high performance of the proposed model. It likely helps the model to perform well even when some sentiment classes are underrepresented.

5. Conclusion. This research embarked on an ambitious quest to harness the synergy of Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) networks, augmented by a dynamic weighted loss function, to tackle the challenges of sentiment analysis in educational big data. The goal was to extract meaningful insights from student feedback on course evaluations, providing actionable intelligence for educational improvement. The study's findings are both significant and promising. The proposed CNN-

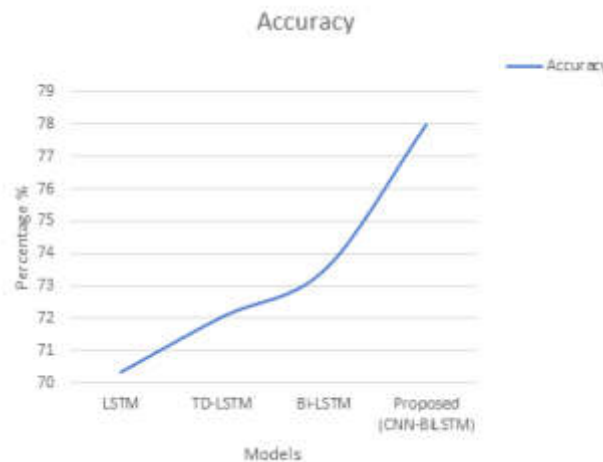


Fig. 4.2: Model Accuracy

BiLSTM model demonstrated superior performance over traditional LSTM, TD-LSTM, and Bi-LSTM models across several key metrics. With precision scores reaching 78.92%, recall at 77.65%, an F1-score of 77.9%, and an overall accuracy of 78%, the model's efficacy in identifying and classifying sentiment in textual feedback has been clearly established. These results underscore the model's adeptness not only in feature extraction through CNNs, which effectively identify sentiment-indicative patterns, but also in capturing the nuances of language context via BiLSTM networks. The integration of a dynamic weighted loss function played a pivotal role in balancing the scale among sentiment classes, especially in the face of class imbalance—an issue prevalent in real-world datasets. The CNN layer's efficacy is strongly reliant on the quality of preprocessed text. If the data is not correctly cleaned and prepared during the preprocessing stage, the CNN may extract irrelevant or deceptive characteristics. While CNNs are good in pattern detection, they are frequently referred to as 'black boxes'. This makes interpreting why the network finds specific elements or patterns to be indicative of emotion difficult, which can be a significant restriction in educational contexts where understanding the 'why' behind feelings is critical.

The research contributes a novel approach to sentiment analysis, specifically tailored for the educational sector. It addresses the call for sophisticated analytical tools capable of sifting through large volumes of unstructured feedback, providing educators and institutions with a deep, data-driven understanding of student sentiment. The implications for course design and pedagogical strategies are profound, as the model offers granular insights that can guide curriculum development, teaching methodologies, and overall educational delivery.

Acknowledgement. Research on Higher Education Teaching Reform in Hebei Province, Project Number: 2021GJJG065

REFERENCES

- [1] F. A. ACHEAMPONG, H. NUNOO-MENSAH, AND W. CHEN, *Transformer models for text-based emotion detection: a review of bert-based approaches*, Artificial Intelligence Review, (2021), pp. 1–41.
- [2] M. AHMAD, S. AFTAB, M. S. BASHIR, AND N. HAMEED, *Sentiment analysis using svm: a systematic literature review*, International Journal of Advanced Computer Science and Applications, 9 (2018).
- [3] F. ALQASEMI, A. ABDELWAHAB, AND H. ABDELKADER, *Constructing automatic domain-specific sentiment lexicon using knn search via terms discrimination vectors*, International Journal of Computers and Applications, 41 (2019), pp. 129–139.
- [4] M. BANSAL, S. VERMA, K. VIG, AND K. KAKRAN, *Opinion mining from student feedback data using supervised learning algorithms*, in International Conference on Image Processing and Capsule Networks, Springer, 2022, pp. 411–418.

- [5] G. BATHLA AND A. KUMAR, *Opinion spam detection using deep learning*, in 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN), IEEE, 2021, pp. 1160–1164.
- [6] F. BENAMARA, C. CESARANO, A. PICARIELLO, D. R. RECUPERO, AND V. S. SUBRAHMANYAN, *Sentiment analysis: Adjectives and adverbs are better than adjectives alone.*, ICWSM, 7 (2007), pp. 203–206.
- [7] M. BOUAZIZI AND T. OHTSUKI, *Multi-class sentiment analysis on twitter: Classification performance and challenges*, Big Data Mining and Analytics, 2 (2019), pp. 181–194.
- [8] R. CATELLI, S. PELOSI, AND M. ESPOSITO, *Lexicon-based vs. bert-based sentiment analysis: A comparative study in italian*, Electronics, 11 (2022), p. 374.
- [9] N. V. CHAWLA, K. W. BOWYER, L. O. HALL, AND W. P. KEGELMEYER, *Smote: synthetic minority over-sampling technique*, Journal of artificial intelligence research, 16 (2002), pp. 321–357.
- [10] H. CHOI, K. CHO, AND Y. BENGIO, *Context-dependent word representation for neural machine translation*, Computer Speech & Language, 45 (2017), pp. 149–160.
- [11] N. DEHBOZORGI AND D. P. MOHANDOSS, *Aspect-based emotion analysis on speech for predicting performance in collaborative learning*, in 2021 IEEE Frontiers in Education Conference (FIE), IEEE, 2021, pp. 1–7.
- [12] J. DING, H. SUN, X. WANG, AND X. LIU, *Entity-level sentiment analysis of issue comments*, in Proceedings of the 3rd International Workshop on Emotion Awareness in Software Engineering, 2018, pp. 7–13.
- [13] F. S. DOLIANITI, D. IAKOVAKIS, S. B. DIAS, S. J. HADJILEONTIADOU, J. A. DINIZ, G. NATSIU, M. TSITOURIDOU, P. D. BAMDIS, AND L. J. HADJILEONTIADIS, *Sentiment analysis on educational datasets: a comparative evaluation of commercial tools*, Educational Journal of the University of Patras UNESCO Chair, (2019).
- [14] J. DUAN, X. DING, AND T. LIU, *Learning sentence representations over tree structures for target-dependent classification*, in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 551–560.
- [15] G. DANIELLO, M. GAETA, AND I. LA ROCCA, *Knowmis-absa: an overview and a reference model for applications of sentiment analysis and aspect-based sentiment analysis*, Artificial Intelligence Review, 55 (2022), pp. 5543–5574.
- [16] A. I. M. ELFEKY, T. S. Y. MASADEH, AND M. Y. H. ELBYALY, *Advance organizers in flipped classroom via e-learning management system and the promotion of integrated science process skills*, Thinking Skills and Creativity, 35 (2020), p. 100622.
- [17] G. G. ESPARZA, A. DE LUNA, A. O. ZEZZATTI, A. HERNANDEZ, J. PONCE, M. ÁLVAREZ, E. COSSIO, AND J. DE JESUS NAVA, *A sentiment analysis model to analyze students reviews of teacher performance using support vector machines*, in Distributed Computing and Artificial Intelligence, 14th International Conference, Springer, 2018, pp. 157–164.
- [18] R. FAIZI, *A sentiment-based approach to predict learners perceptions towards youtube educational videos*, in International Conference on Innovations in Bio-Inspired Computing and Applications, Springer, 2021, pp. 549–556.
- [19] D. P. KINGMA AND J. BA, *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980, (2014).
- [20] S. LI, Z. ZHAO, R. HU, W. LI, T. LIU, AND X. DU, *Analogical reasoning on chinese morphological and semantic relations*, arXiv preprint arXiv:1805.06504, (2018).
- [21] J. LIAO, J. TANG, AND X. ZHAO, *Course drop-out prediction on mooc platform via clustering and tensor completion*, Tsinghua Science and Technology, 24 (2019), pp. 412–422.
- [22] Z. NASIM, Q. RAJPUT, AND S. HAIDER, *Sentiment analysis of student feedback using machine learning and lexicon based approaches*, in 2017 international conference on research and innovation in information systems (ICRIIS), IEEE, 2017, pp. 1–6.
- [23] H. T. NGUYEN AND M. LE NGUYEN, *Effective attention networks for aspect-level sentiment classification*, in 2018 10th International Conference on Knowledge and Systems Engineering (KSE), IEEE, 2018, pp. 25–30.
- [24] Y.-P. RUAN, Q. CHEN, AND Z.-H. LING, *A sequential neural encoder with latent structured description for modeling sentences*, IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26 (2017), pp. 231–242.
- [25] J. WANG, J. LI, S. LI, Y. KANG, M. ZHANG, L. SI, AND G. ZHOU, *Aspect sentiment classification with both word-level and clause-level attention networks.*, in IJCAI, vol. 2018, 2018, pp. 4439–4445.
- [26] S. WANG, S. MAZUMDER, B. LIU, M. ZHOU, AND Y. CHANG, *Target-sensitive memory networks for aspect sentiment classification*, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Dec 12, 2023

Accepted: Jan 3, 2024



GREEN PLANT LANDSCAPE DESIGN FOR URBAN AIR QUALITY PURIFICATION WITH COMPUTER IMAGE PROCESSING IN CLOUD, GRID, AND CLUSTER COMPUTING

JINGJING NI*

Abstract. This research paper explores the innovative integration of green plant landscape design with advanced computer image processing in cloud, grid, and cluster computing environments to enhance urban air quality purification. The study begins by highlighting the critical need for improving air quality in urban areas, considering the rising levels of pollution and its impact on public health and the environment. The methodology involves the use of sophisticated image processing techniques to analyze various sensors on air quality measures and plant species their effectiveness in air purification, facilitated by the computational power of cloud, grid, and cluster computing. A diverse range of green plants was selected, and their air purification capabilities were assessed through a series of computer-simulated models. These models were developed using complex algorithms to predict the plants' performance in real-world urban settings. The research uniquely combines landscape architecture with technology, emphasizing the role of green spaces in urban areas for environmental sustainability. The results demonstrate that certain plant species are more effective than others in purifying urban air. The study provides a comprehensive ranking of these plants based on their purification capabilities, growth requirements, and suitability for various urban landscapes. The paper concludes by proposing practical guidelines for urban landscape designers and policymakers, recommending the strategic incorporation of specific green plants in urban areas to maximize air purification. Additionally, it highlights the potential of leveraging advanced computing technologies in environmental research and urban planning. This research contributes to the fields of environmental science, urban planning, and computer science by showcasing how multidisciplinary approaches can address pressing environmental issues. It opens avenues for further research in the optimization of urban green spaces using advanced computing techniques. The results demonstrate that certain plant species are more effective than others in purifying urban air. The study provides a comprehensive ranking of these plants based on their purification capabilities, growth requirements, and suitability for various urban landscapes. The paper concludes by proposing practical guidelines for urban landscape designers and policymakers, recommending the strategic incorporation of specific green plants in urban areas to maximize air purification.

Key words: Urban Air Quality, Green Plant Landscaping, Environmental Purification, Computer Image Processing, Cloud Computing, Grid Computing, Cluster Computing

1. Introduction. In the wake of escalating urbanization and industrialization, air pollution has emerged as a critical challenge confronting urban environments globally. The detrimental impact of poor air quality on human health and the ecosystem necessitates innovative solutions. This study explores a novel approach to ameliorate urban air quality: the strategic design of green plant landscapes, aided by advanced computational technologies. The role of green plants in purifying air is well-documented. They absorb pollutants and carbon dioxide, releasing oxygen, thereby improving air quality. However, the effectiveness of different plant species in specific urban contexts remains underexplored. This gap in knowledge presents an opportunity to blend environmental science with cutting-edge computing technologies. The study aims to employ computer image processing, harnessed through the power of cloud, grid, and cluster computing, to analyze and optimize green plant landscapes for urban air quality purification.

The primary objectives of this research are to identify the most effective plant species for air purification in urban landscapes, and to develop a computer-assisted model for landscape design that optimizes these benefits. This involves analyzing large datasets of environmental conditions and plant characteristics, a task well-suited to the capabilities of advanced computing paradigms like cloud, grid, and cluster computing. These technologies offer unprecedented processing power and data storage capabilities, facilitating detailed and complex environmental modeling.

*Zhuxi Cultural Tourism College, Anhui Finance and Trade Vocational College Hefei, 230601, China, (jingkingnire@outlook.com)

The introduction of computer image processing into landscape design represents a pioneering step in environmental planning. By harnessing these technologies, this research aims to provide actionable insights for urban planners and environmentalists, contributing to more sustainable and healthier urban environments. This paper is structured as follows: after the introduction, we present a review of the literature, outlining previous studies on green plants for air purification and the application of advanced computing in environmental science. This is followed by a detailed description of the methodology, including the selection of plant species, computational models used, and the design of the study. The subsequent sections present the results, discussion, and conclusions drawn from the research, along with recommendations for future studies in this evolving field.

The study begins by underlining the critical issue of declining air quality in metropolitan areas as pollution levels rise. This first recognition of the problem establishes the context for the study's relevance and significance. The study takes a unique approach by analyzing air quality sensors and the performance of several plant species in air filtration using modern image processing techniques. The usage of cloud, grid, and cluster computing shows a dedication to harnessing current technology to solve environmental problems.

"Image Processing in Cloud, Grid, and Cluster Computing" would be focused on exploring and achieving specific goals at the intersection of environmental science, urban planning, and advanced computing. The key objectives for this research:

1. Determine using image processing methods are most effective in purifying air in urban environments. This involves assessing various plants' ability to absorb pollutants and improve air quality.
2. Utilize computer image processing tools within cloud, grid, and cluster computing environments to analyze the physical and biological characteristics of different plant species. This includes studying their growth patterns, pollution absorption rates, and adaptability to urban settings.
3. Create computational models that can simulate and predict the effectiveness of different green plant arrangements in urban landscapes for maximizing air purification.

Research questions that concentrated in this research are,

1. What are the most effective methods for purifying urban air?
2. How does the integration of specific plant species in urban landscapes impact overall environmental and public health?
3. What are the challenges and limitations of using advanced computing technologies in environmental planning and monitoring?

Urban areas worldwide are grappling with escalating levels of air pollution, which pose serious risks to public health and the environment. Addressing this issue is vital for the well-being of urban populations and the sustainability of cities. Green plants are known to improve air quality by absorbing pollutants and carbon dioxide, offering a natural solution to the air pollution problem. However, the effectiveness of specific plant species and configurations in urban environments needs further exploration. The rapid development in computing technologies, such as cloud, grid, and cluster computing, offers unprecedented capabilities in data processing and analysis. Applying these technologies to environmental challenges presents an opportunity to innovate in urban air quality management.

2. Literature review. Globally, air quality issues are a major concern in urban areas, necessitating effective detection and management of air pollution variations over time and by region. This is essential for developing affordable solutions[10]. In India, air pollution remains a persistent public health issue[5]. The "Global Burden of Disease Study, 2019" reported that in 2019, air pollution was responsible for 1.67 million deaths, which is 17.8% of total deaths in India, leading to an economic loss of approximately USD 36.8 billion, or 1.36% of the country's GDP. Furthermore, in 2019, 22 Indian cities ranked among the top 30 most polluted cities globally, with many Indian cities appearing in the top 10 (IQAir, 2020). The COVID-19 lockdown in 2020 unexpectedly contributed to environmental recovery, significantly improving urban air quality[5, 12]. However, this improvement was temporary. [24] described the lockdown as an "anthropause," a brief pause that is unlikely to have a lasting impact on the detrimental effects of human activities. In India, as the lockdown was lifted, air pollution levels began to increase again [8], mirroring trends observed in other cities around the world [3, 9]. Post-lockdown, many Indian cities saw a significant increase in ambient particulate matter levels (PM10), with levels in the last quarters of 2020 approaching those seen in 2019 (life-as-usual situation).

In many developing nations, traditional methods for addressing air pollution have been largely unsuccessful.

This is due to a combination of factors including institutional weaknesses, infrastructural challenges, economic constraints, and political hurdles[13]. Consequently, air pollution continues to pose a significant threat to both environmental sustainability and public health, especially in India[14, 16, 18]. A key factor behind these shortcomings is the predominantly technocratic approach, which often neglects the socio-cultural aspects like public expectations, community capacity, and public participation in decision-making [17]. Recognizing these issues, the main motivation for this discussion is to shift from a purely technocratic mindset to one that harmonizes with nature. We advocate for policymakers to consider urban green spaces not just as aesthetic elements, but as vital components that enhance and strengthen efforts in air pollution prevention and control.

In their recent bibliometric study, [20] categorized the primary mechanisms by which plants remove pollutants, as they relate to public health, into three main groups: (I) dry deposition, (II) dispersion (the process by which plants alter air pollutants' path and speed through their physical structure), and (III) modification (including selective sorption, microbial reactions, and chemical coagulation due to Brownian motion and/or van der Waals forces). It's recommended that city planners and authorities delve into such research to gain a deeper understanding of how plants interact with pollutants and to maximize the effectiveness of urban green spaces. However, the dry deposition process is complicated by various factors, notably the diverse types of leaf surfaces in urban canopies and the movement of submicron particles. Similarly, there is a lack of comprehensive experimental data encompassing all scenarios related to dry deposition [11, 19]. The complexity is further heightened by varying levels of urban development, pollution sources, and human activities. Addressing these complexities will require more case studies in different urban environments and among various population demographics to fully understand and optimize dry deposition processes.

The initial consideration for selecting tree species for air pollution mitigation should focus on their climatic characteristics and the impact these have on the length of their growing season, which determines the duration of leaf cover[6, 22]. In temperate regions, deciduous trees lose their leaves during winter, thereby reducing the total leaf surface area available for pollutant absorption. Among these, conifers are often favored due to their lipophilic wax-coated needles, smaller leaf size, and intricate shoot structures, which are advantageous for pollutant capture [15, 23]. Additionally, the airflow around conifer needles creates more turbulence compared to larger leaves (broadleaves), which reduces the thickness of the boundary layer on needle leaves (Ackerly et al., [21]; [21, 20]. This means when air carrying pollutants passes over these needles, the boundary layer remains relatively still, creating a barrier between the air and the leaf surface.

However, the high pollutant absorption by conifer needles can sometimes damage the leaves, diminishing their effectiveness in pollutant removal. This issue is more pronounced in drier climates[24, 2]. In such cases, broad-leaved deciduous species or those that retain their leaves throughout winter are more suitable for pollution control [22]. Among broadleaved species, those with a high number of grooves, a large ratio of groove area to total leaf area, and dense epicuticular trichomes are preferred for pollution regulation [17, 7].

In climates with shorter growing seasons but high pollutant levels in winter, evergreen species, which maintain their foliage all year, are preferable over deciduous varieties[1]. The selection should also take into account the type of pollutants targeted. For instance, [4] found that oak leaves (deciduous) are more effective against particulate phase PAHs (polycyclic aromatic hydrocarbons) due to their high specific leaf area. Conversely, pine needles (evergreen) may be better suited for gaseous phase PAHs, particularly effective in capturing low and medium weight of PAHs molecules.

3. Methodology. The methodology for this research combines environmental science, landscape architecture, and advanced computational techniques to optimize green plant landscapes for urban air quality purification. For the study, a varied selection of green plant species were carefully selected. Plant kinds (trees, shrubs, ground cover), their capacity to filter the air, growth characteristics, and adaptation to urban situations are all variables in the choosing process. This choice is influenced by scientific understanding as well as landscape architecture concepts.

In metropolitan areas under research, sophisticated air quality monitors are carefully deployed. These sensors capture data on a variety of air quality indicators, including pollutants such as particulate matter (PM), nitrogen dioxide (NO₂), and volatile organic compounds (VOCs), in real time. Data from air quality sensors is gathered over time, documenting fluctuations in air quality based on factors such as time of day, meteorological conditions, and traffic density. This information will be used to analyze the efficacy of green

plants in air cleansing.

To examine data gathered from air quality sensors, advanced image processing techniques are used. These methods enable the extraction of useful information such as pollutant levels, regional distribution, and temporal trends. To mimic the behavior of chosen green plants in urban contexts, complex computational models are built. These models consider elements such as plant growth, transpiration rates, and the ability of plants to remove toxins from the air. To conduct these simulations efficiently, the computational capacity of cloud, grid, and cluster computing resources is used:

1. Selection of Plant Species
2. Data Collection and Analysis
3. Computer Image Processing
4. Computational Modeling
5. Pilot Implementation and Monitoring

3.1. Phase 1: Selection of Plant Species. In the research on "Green Plant Landscape Design for Urban Air Quality Purification," Phase 1, which focuses on the Selection of Plant Species, is a crucial foundational step. The objective here is to identify plants that are most effective in urban air purification, taking into account various urban and climatic conditions. This phase begins with an extensive literature review, where existing scientific studies, environmental reports, and botanical research are scrutinized to identify plants known for their pollution-absorbing abilities and adaptability to urban stresses. The selection criteria for these plants include their capacity to absorb specific urban pollutants, growth and maintenance needs, environmental adaptability, aesthetic contribution, and practical considerations like space and root development.

Field studies, consultations with experts like botanists and urban ecologists, and citizen science initiatives form the backbone of the data collection strategy. These diverse sources ensure a comprehensive understanding of how different plants perform in urban settings. The analysis of this data is thorough, involving comparative assessments of plants against the set criteria, statistical modeling to understand the correlation between plant traits and pollution absorption, and evaluations of climate adaptability.

The outcome of this phase is a carefully curated list of plant species, each with a detailed profile outlining its environmental benefits, physical characteristics, and care instructions. This list is not only crucial for the immediate next phases of the research, which involve further data collection and computational modeling, but also sets a precedent for interdisciplinary collaboration. By meticulously choosing the right plant species in Phase 1, the research ensures that the subsequent phases are informed by a deep and nuanced understanding of the best natural resources for combatting urban air pollution. Based on diverse urban condition, plant is selected by the user.

In this research, the characteristics of Broad-Leaved Deciduous Species and Evergreens are considered for pollution control. This research comprehensively evaluates the suitability of Broad-Leaved Deciduous Species and Evergreens in pollution control, emphasizing their distinct characteristics that enhance air purification in urban environments. Broad-Leaved Deciduous Species are noted for their seasonal leaf shedding and large leaf surface area. Despite losing leaves in winter, they offer significant benefits during the growing season. Their broad leaves provide a substantial surface for absorbing pollutants, especially effective in warmer months when pollution levels tend to spike. Evergreens, conversely, retain their leaves year-round, offering continuous air purification, including in colder months. Their resilience across various climates renders them versatile for different urban conditions.

The research further delves into specific traits beneficial for pollution control:

1. High Number of Grooves on Leaves: This feature increases the surface area, enabling the leaves to trap more pollutants. The grooves are particularly adept at capturing fine particulate matter, a major urban pollution component.
2. Large Groove Area Relative to Total Leaf Area: A higher ratio here indicates a more effective trapping mechanism for pollutants, optimizing the leaves for absorption, especially crucial in densely populated areas.
3. Dense Epicuticular Trichomes: These hair-like structures act as pollution filters, trapping and absorbing pollutants. They also protect the plant from environmental stresses, including high pollution levels, maintaining their efficiency in pollutant absorption.

The selection of these plant types underscores the research's focus on natural, sustainable methods for improving urban air quality. Their large leaf surfaces, grooved structures, and trichomes significantly enhance their ability to capture and absorb airborne pollutants, making them highly relevant for application in diverse urban landscapes. This approach not only addresses air quality issues but also promotes a greener, more sustainable urban environment. The selection of broad-leaved deciduous species and evergreens for this research is based on their distinct characteristics that make them suitable for urban air purification. Their large leaf surfaces, grooved structures, and the presence of trichomes enhance their ability to capture and absorb airborne pollutants. This choice underscores the research's emphasis on employing natural, sustainable solutions to address urban air quality issues, making it relevant and practical for application in diverse urban landscapes.

3.2. Phase 2: Data Collection and Analysis. .

This phase is critical for empirically validating the pollution absorption capabilities of the selected plant species and understanding their practical implications in urban environments. The approach in this phase integrates traditional environmental science methods with innovative data collection and analysis techniques, focusing on both qualitative and quantitative aspects.

3.2.1. Data Collection Strategy. Deploy air quality sensors in the vicinity of the planted areas to continuously monitor levels of key pollutants (e.g., PM_{2.5}, PM₁₀, NO_x, SO_x, CO, O₃). These sensors should be strategically placed at various heights and distances from the plants to capture a comprehensive data set. Periodically collect leaf samples from the selected plants for laboratory analysis. This will involve examining the physical characteristics of the leaves (such as leaf surface area, groove depth, and trichome density) and quantifying the accumulated pollutants on the leaf surface using techniques like gas chromatography-mass spectrometry (GC-MS) or X-ray fluorescence (XRF) analysis. Record environmental factors such as temperature, humidity, wind speed, and rainfall, as they can significantly influence the plants' pollutant absorption capabilities. Utilize high-resolution photography and drone imagery to document the physical state of the plants over time. This can provide insights into their growth patterns, health, and environmental interactions. Engage with local communities to collect qualitative data on their perceptions of air quality and the impact of the green spaces.

3.2.2. Analysis Techniques. Employ machine learning algorithms to analyze the complex dataset. Techniques like regression analysis, cluster analysis, and neural networks can reveal patterns and correlations between plant characteristics, pollutant levels, and environmental factors. Use Geographic Information Systems (GIS) to map pollution levels and plant locations. This spatial analysis can reveal how the distribution of plants affects air quality in different urban zones. Implement time-series analysis to understand how the effectiveness of plants in pollution absorption varies over time and in different environmental conditions. Compare data from sites with the selected plants to control sites without them. This will provide a clearer picture of the plants' direct impact on air quality. Use natural language processing (NLP) to analyze community feedback, providing insights into public perception and acceptance of the green spaces.

Quantitative data demonstrating the effectiveness of the selected plants in reducing specific urban pollutants. Also, understanding how different plant species perform under varying urban environmental conditions helps to plan urban landscape. Data-driven recommendations for urban planners and policymakers on integrating specific plant species in urban landscape design for air quality improvement. Enhanced community involvement and awareness about the role of urban greenery in improving air quality.

By integrating advanced data collection and analysis techniques, this phase aims to provide a robust scientific foundation for the use of specific plant species in urban air purification. The results will not only validate the plant selection but also offer practical guidelines for their effective implementation in urban landscape designs.

3.3. Phase 3: Computer Image Processing Techniques. The use of high-resolution photography, possibly supplemented by drone imagery, is essential. These images provide detailed visual data on plant growth, health, and environmental interactions, which are crucial for understanding their ability to purify urban air. Algorithms are employed to analyze these images for various parameters, such as leaf surface area, groove depth, and trichome density. These characteristics are significant as they relate to the plants' pollution absorption capabilities.

CNNs are ideal for image recognition tasks. They can be trained to identify specific plant features that correlate with pollution absorption, like leaf structure, density of trichomes, etc. This involves labeling each pixel of an image with a class (like leaf, branch, flower, etc.), allowing for detailed analysis of the plant parts and their specific roles in air purification. By processing sequential images of the plants over time, this approach helps in observing changes in plant growth and health, providing insights into how environmental factors impact their air purification abilities. Using 3D image processing to create models of the plant structures can provide insights into their physical arrangement and how this affects their efficiency in air purification.

By quantifying physical aspects of plants such as leaf surface area and groove depth, researchers can establish correlations between these characteristics and the plants' ability to absorb pollutants. Image processing helps in continuously monitoring the health and growth patterns of plants in urban settings, crucial for understanding their long-term effectiveness in air purification. Through image analysis, the interaction of plants with their surrounding environment, including factors like light exposure, urban structures, and human activity, can be better understood. The data derived from image processing can be integrated with data from Phase 2 (such as pollutant levels and environmental factors) to provide a more comprehensive understanding of the plants' performance in urban air purification.

To develop computational models that can accurately predict how different plant species influence urban air quality under a variety of environmental conditions. These models integrate data on plant characteristics (like leaf surface area, trichome density), pollutant levels (PM_{2.5}, NO_x, etc.), and environmental factors (temperature, humidity, urban structures).

Data collected from previous phases, such as air quality measurements and plant characteristics, is cleaned, normalized, and transformed to be used in modeling. Identifying the most relevant features that influence air purification, such as specific plant traits and local environmental conditions. Building a framework in Python where different scenarios can be simulated, such as varying levels of pollution, different plant species combinations, and changing weather conditions. Using historical data to validate the simulations, ensuring they accurately reflect real-world scenarios. Designing multi-layer neural networks that can process complex patterns in the data. These networks might include convolutional layers for spatial data processing, especially useful when dealing with image data from Phase 3. Splitting the dataset into training and testing sets. The network is trained on the training set, learning to predict air quality based on plant characteristics and environmental factors.

Decision trees are used to build rule-based models. These models help to understand and illustrate the decision-making process, especially in identifying the most relevant elements impacting air quality. These are particularly useful for understanding and visualizing the decision process, like which factors most significantly affect air quality. Using software like ArcGIS or QGIS for spatial analysis. Creating maps that visually represent data such as the distribution of plant species and pollution levels across urban areas. Analyzing how the distribution of different plant species across an urban area affects air quality. Investigating how factors like urban layout, traffic density, and green space distribution correlate with air purification effectiveness.

4. Result Evaluation. This research focuses on examining how the arrangement of green spaces in cities affects air pollution levels. It specifically looks at 20 garden cities in China randomly that experience a subtropical monsoon climate. The study uses data from 2019, including urban air quality measurements and information on land use types. By employing landscape metrics and spatial regression models, the study investigates the connection between the layout of green spaces and air pollution concentrations. This paper utilizes regression modeling tools available in the GeoDa software to perform SEM (Structural Equation Modeling) regression analysis. The outcomes of this analysis are presented in Table 4.1.

Landscape Shape Index (LSI) of grasslands. Notably, the association between SO₂ levels and both the PD of forestlands and the LSI of grasslands was particularly strong ($p < 0.01$). Conversely, the SO₂ concentration had a significant and negative correlation with the patch proportion in landscape area (PLAND) of forestlands, the patch density (PD) of grasslands, and the PLAND of agricultural lands, with the last two showing a very significant relationship with SO₂ levels ($p < 0.01$).

The study found that a one-unit increase in the PD of forestlands, PLAND, and LSI of grasslands led to increases in SO₂ concentration by 149.939, 0.752, and 0.429 units, respectively. On the other hand, a one-unit rise in the PLAND of forestlands, PD of grasslands, and PLAND of farmlands resulted in decreases in

Table 4.1: Various concentrations measures and analysis

	Variable	ρ	Threshold	ρ	Threshold	ρ	Threshold
		PM 2.5		NO ₂		SO ₂	
Grass lands	PLAND	0.821		0.885		0.000	0.000**
	PD	0.616		0.994		0.000	0.000**
	LSI	0.236		0.897		0.000	0.000**
Farm lands	PLAND	0.214	0.214*	0.532	0.532*	0.0021	
	PD	0.645		0.687		0.0054	
	LSI	0.347		0.752		0.0061	

SO₂ concentration by 0.073, 214.564, and 0.172 units, respectively. The SO₂ levels were not noticeably impacted by the LSI of forestlands, and the PD and LSI of agricultural lands.

The spatial correlation analysis from the 20 cities showed a significant link between the layout of urban green spaces and the levels of PM2.5, NO₂, and SO₂ pollutants. The pattern of green spaces, however, did not significantly affect PM10 levels. The PLAND, PD, and LSI of grasslands, along with the PLAND of farmlands, had an effect on SO₂ concentrations.

In the effort to optimize and rejuvenate the layout of urban green spaces, research has shown that strategically planning the design and distribution of green space networks can enhance air quality and benefit public health. Based on these insights, several suggestions are proposed for improving air pollution in cities with subtropical monsoon climates. Forests, grasslands, and farmlands are effective in reducing concentrations of PM2.5, NO₂, and SO₂. In urban areas, grassland should be managed carefully, forest coverage should be increased, and the restoration of damaged forests should be expedited to enhance ecosystem stability. The urban green space landscape should be meticulously planned based on scientific principles to balance ecological spaces for living and production.

In cities where NO₂ and SO₂ are predominant pollutants, arranging forest, grassland, and farmland landscapes systematically can help reduce pollution levels. Optimal NO₂ concentration in urban areas is achieved when the Patch Density (PD) of forest land is around 0.072. For SO₂, the best reduction effects are observed when forest land and grassland densities and layouts are adjusted to specific parameters, with an LSI of grassland at 14.13. This suggests that urban green spaces should be carefully planned to optimize patch density and diversify green space types, thus alleviating air pollution. By integrating green spaces into urban areas, air quality can be improved effectively and economically.

PM2.5 pollution, mainly from industrial emissions, traffic, and biomass combustion, is prevalent in urban roads and industrial areas. In China's subtropical monsoon regions, where urban development is rapid, reducing traffic and industrial emissions is challenging. However, optimizing urban green space layouts can mitigate PM2.5 pollution. An optimal reduction in PM2.5 levels occurs when the LSI of forest land reaches 18.02. Enhancing the greenery along streets and near factories, increasing the interaction between green spaces and PM2.5, and improving the vertical structure of urban green belts can effectively block and absorb PM2.5 pollutants.

Additionally, managing unused and inefficient land to create a well-planned urban green landscape is recommended. Promoting green, healthy, and low-carbon lifestyles and consumption habits among residents is also advised. Disseminating scientific findings on PM2.5, NO₂, and SO₂ pollution control can help government departments implement measures more effectively and gain public support for air pollution control initiatives.

5. Conclusion. In this research underscores the vital role of urban green spaces in mitigating air pollution and enhancing public health, particularly in cities with subtropical monsoon climates. The strategic planning and distribution of green spaces, such as forests, grasslands, and farmlands, have been identified as key factors in reducing concentrations of harmful pollutants like PM2.5, NO₂, and SO₂. By carefully managing these green areas, especially in urban settings, and adhering to specific landscape metrics such as Patch Density and Landscape Shape Index, significant improvements in air quality can be achieved. The study highlights that different types of pollutants require distinct approaches in terms of green space management. For instance,

the optimal control of NO₂ and SO₂ involves adjusting the density and layout of forests and grasslands, while tackling PM_{2.5} pollution necessitates enhancing urban greenery in areas with high traffic and industrial activity. Moreover, the research emphasizes the importance of integrating ecological considerations into urban planning and development. This involves not only improving the design of green spaces but also promoting sustainable lifestyles and consumption patterns among residents. Such holistic approaches not only contribute to better air quality but also foster healthier, more sustainable urban environments.

Cities throughout the world have taken a transformational approach to urban design and sustainability in this imagined future by building Sustainable Urban Green Zones (SUGZs). These are carefully planned and strategically placed green spaces within metropolitan areas that promote air quality purification and environmental well-being. The research findings have a significant impact on the creation and management of SUGZs.

Acknowledgment.

¹2016 Anhui Social Science Association Social Science Innovation Development Research Project(2016CXF089)

²2017 Anhui Finance and Trade Vocational College "Connotation Promotion All Staff Action Plan" Scientific Research Project "Study on Tourism Poverty Reduction and Development Model"(2017nhrwc27)

³Anhui Provincial Tourism Youth Expert Training Program

REFERENCES

- [1] K. ABHIJITH AND S. GOKHALE, *Passive control potentials of trees and on-street parked cars in reduction of air pollution exposure in urban street canyons*, Environmental pollution, 204 (2015), pp. 99–108.
- [2] R. ALTAF, S. ALTAF, M. HUSSAIN, R. U. SHAH, R. ULLAH, M. I. ULLAH, A. RAUF, M. J. ANSARI, S. A. ALHARBI, S. ALFARRAJ, ET AL., *Heavy metal accumulation by roadside vegetation and implications for pollution control*, Plos one, 16 (2021), p. e0249147.
- [3] N. A. ANJUM, *Good in the worst: Covid-19 restrictions and ease in global air pollution*, (2020).
- [4] Y. BARWISE AND P. KUMAR, *Designing vegetation barriers for urban air pollution abatement: A practical review for appropriate plant species selection*, Npj Climate and Atmospheric Science, 3 (2020), p. 12.
- [5] B. BERA, S. BHATTACHARJEE, P. K. SHIT, N. SENGUPTA, AND S. SAHA, *Significant impacts of covid-19 lockdown on urban air pollution in kolkata (india) and amelioration of environmental health*, Environment, development and sustainability, 23 (2021), pp. 6913–6940.
- [6] J. BJÖRK, M. ALBIN, P. GRAHN, H. JACOBSSON, J. ARDÖ, J. WADBRO, P.-O. ÖSTERGREN, AND E. SKÄRBÄCK, *Recreational values of the natural environment in relation to neighbourhood satisfaction, physical activity, obesity and wellbeing*, Journal of Epidemiology & Community Health, 62 (2008), pp. e2–e2.
- [7] S. J. BRAKE, K. BARNSLEY, W. LU, K. D. MCALINDEN, M. S. EAPEN, AND S. S. SOHAL, *Smoking upregulates angiotensin-converting enzyme-2 receptor: a potential adhesion site for novel coronavirus sars-cov-2 (covid-19)*, 2020.
- [8] M. CETIN, *Determining the bioclimatic comfort in kastamonu city*, Environmental monitoring and assessment, 187 (2015), pp. 1–10.
- [9] ———, *Using gis analysis to assess urban green space in terms of accessibility: case study in kutahya*, International Journal of Sustainable Development & World Ecology, 22 (2015), pp. 420–424.
- [10] ———, *The effect of urban planning on urban formations determining bioclimatic comfort areas effect using satellitia imagines on air quality: a case study of bursa city*, Air Quality, Atmosphere & Health, 12 (2019), pp. 1237–1249.
- [11] M. CETIN, T. AKSOY, S. N. CABUK, M. A. S. KURKCUOGLU, AND A. CABUK, *Employing remote sensing technique to monitor the influence of newly established universities in creating an urban development process on the respective cities*, Land use policy, 109 (2021), p. 105705.
- [12] M. CETIN, A. K. ONAC, H. SEVIK, AND B. SEN, *Temporal and regional change of some air pollution parameters in bursa*, Air Quality, Atmosphere & Health, 12 (2019), pp. 311–316.
- [13] M. CETIN, H. SEVIK, AND N. YIGIT, *Climate type-related changes in the leaf micromorphological characters of certain landscape plants*, Environmental monitoring and assessment, 190 (2018), pp. 1–9.
- [14] I. J. CHAUDHARY AND D. RATHORE, *Suspended particulate matter deposition and its impact on urban trees*, Atmospheric Pollution Research, 9 (2018), pp. 1072–1082.
- [15] S. CHAUDHURI AND A. KUMAR, *Urban greenery for air pollution control: a meta-analysis of current practice, progress, and challenges*, Environmental Monitoring and Assessment, 194 (2022), p. 235.
- [16] S. CHAUDHURI, M. ROY, AND A. JAIN, *Appraisal of wash (water-sanitation-hygiene) infrastructure using a composite index, spatial algorithms and sociodemographic correlates in rural india*, Journal of Environmental Informatics, 35 (2020), pp. 1–22.
- [17] S. CHAUDHURI, M. ROY, L. M. McDONALD, AND Y. EMENDACK, *Water for all (har ghar jal): rural water supply services (russ) in india (2013–2018), challenges and opportunities*, International Journal of Rural Management, 16 (2020), pp. 254–284.

- [18] ———, *Coping behaviours and the concept of time poverty: a review of perceived social and health outcomes of food insecurity on women and children*, Food Security, 13 (2021), pp. 1049–1068.
- [19] ———, *Reflections on farmers social networks: a means for sustainable agricultural development?*, Environment, Development and Sustainability, 23 (2021), pp. 2973–3008.
- [20] A. DIENER AND P. MUDU, *How can vegetation protect us from air pollution? a critical review on green spaces' mitigation abilities for air-borne particles from a public health perspective-with implications for urban planning*, Science of the Total Environment, 796 (2021), p. 148605.
- [21] A. Q. E. GROUP ET AL., *Impacts of vegetation on urban air pollution*, (2018).
- [22] V. PADVETNAYA, S. CHAUDHURI, ET AL., *What do farmers dont know? a generic index to summarize cognitive awareness of groundwater-sourced irrigation and conservation at grassroots*, Ecology, Environment and Conservation, 28 (2022), pp. 180–193.
- [23] C. WANG, M. GUO, J. JIN, Y. YANG, Y. REN, Y. WANG, AND J. CAO, *Does the spatial pattern of plants and green space affect air pollutant concentrations? evidence from 37 garden cities in china*, Plants, 11 (2022), p. 2847.
- [24] J. YAO, S. WU, Y. CAO, J. WEI, X. TANG, L. HU, J. WU, H. YANG, J. YANG, AND X. JI, *Dry deposition effect of urban green spaces on ambient particulate matter pollution in china*, Science of The Total Environment, 900 (2023), p. 165830.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Dec 10, 2023

Accepted: Jan 4, 2024



LEARNERS BEHAVIOUR PREDICTION AND ANALYSIS MODEL FOR SMART LEARNING PLATFORM USING DEEP LEARNING APPROACH

LIYUAN FENG* AND YUNFENG JI†

Abstract. In the quickly changing field of instructional technology, intelligent educational systems are now essential for individualized and effective instruction. To forecast and understand learners' actions in intelligent educational settings, this research suggests an analytical framework that makes use of deep learning techniques. By offering real-time information on user activities, the goal is to improve these platforms' reactivity and flexibility. Using state-of-the-art deep learning designs, our technique examines large datasets that include interactions between users, interest trends, and efficiency measures. The proposed method classifies the e-learning based behaviour classification and then the e-learning performance prediction using CNN-LSTM. The suggested framework incorporates the temporal relationships and sequential patterns present in learners' actions on the platform by fusing convolutional neural networks (CNNs) and long short-term memory networks (LSTMs). Furthermore, using multimedia information like simulations that are interactive and video lectures, convolutional neural networks (CNNs) are used to gather spatial data. The present study advances smart learning technology by providing a stable and expandable structure for behavior analysis and prediction in students. Through proactive customization of learning events, instructors, content producers, and platform developers can create a setting that is both enjoyable and effective for students. This is made possible by the knowledge gained from this approach.

Key words: Learners Behaviour, prediction and analysis, smart learning platform, Deep Learning Approach, convolutional neural networks, e-learning

1. Introduction. The incorporation of electronic devices has made it possible for creative and customized educational experiences in the quickly changing field of learning. An example of this progress is the introduction of intelligent learning systems that use AI to improve the learning process. The subject of learner behavior has numerous applications, but one that requires significant attention is the forecasting and evaluation of behavior among learners. Customizing lessons to meet individual requirements requires an in-depth comprehension of how learners react to difficulties, respond to happiness, and interact with the educational setting.

Electronic learning is now a standard educational method [27] and has played a significant role in the growth of online learning. Because of the COVID-19 pandemic, e-learning has become increasingly popular due to its extensive learning materials, low knowledge intake threshold, and substantial temporal and spatial flexibility. Still, this style makes it difficult for teachers to assess the progress in learning of their students [28, 19], and concerns have been voiced over the caliber of e-learning. By forecasting how well pupils will do on upcoming tests, lowering the likelihood that students won't pass the course, and guaranteeing the quality of e-learning, the research of learning outcome predictions gives teachers a foundation on which to modify their teaching strategies for students who could have difficulties.

Students' e-learning behavior has a significant influence on their educational outcomes, according to a substantial body of studies examining the connection between e-learning behavior and learning performance. As a result, achievement in learning predictions using data from the learning process has attracted a lot of attention lately [32]. Teachers can adjust their instructional tactics in real time and begin employing the role of monitoring as well as early warning by using the measurement, collection, and evaluation of learning information to accomplish achievement predictions [8].

Studies point out that knowledge of e-learning processes depends on data on e-learning behavior [1, 2]. The term "e-learning behavior data" refers to the information created by students during a variety of behavioral activities carried out on e-learning systems or online educational companies. This information can be used to

*Jiangsu Vocational College Institute of Information Technology, Wuxi, 214153, China (liyuanfengsd@outlook.com)

†Jiangsu Vocational College Institute of Information Technology, Wuxi, 214153, China

refer to the action documents of students during the learning process, with particular attention to the quantity of login systems, quantity of resource access, quantity of forum participants, quantity of resource access, and other behavioral data. As a result, scholars have studied e-learning behavior in detail and developed many educational outcome predictions according to e-learning behavior[3].

The main question on the research,

How does the BCEP prediction framework compare to traditional E-learning classification techniques in terms of prediction accuracy and efficiency?

What are the specific steps and methodologies used in the data cleansing process within the proposed CNN-LSTM prediction framework?

How does the combination of features in the BCEP framework contribute to the overall accuracy of behavior classification in E-learning settings?

What criteria are used for behavior categorization in the CNN-LSTM prediction framework, and how does this affect the system's predictive capabilities?

To lower the operational expenses of the model and deliver high-accuracy, low-time-consuming learning outcome prediction services for online platforms, choosing features can be utilized to preserve important learning behaviors [4, 5]. In addition, e-learning predictors must typically employ e-learning behavior data as input variables directly because of the single input technique for e-learning behavior data. Few models will employ training data that is integrated with learning behavior data (i.e., feature combination processes) on the exact same kind of learning behavior data. Lastly, there is a lack of standardization among crucial learning behavior indicators, with various researchers finding different ones. Important behavioral cues that can be utilized to accurately predict student achievement have not yet been found in this field of research [6, 7].

The main contribution of the proposed method is given below:

1. We offer the behavior classification-based E-learning prediction system (BCEP prediction framework) to address these issues, provide a summary of traditional E-learning classification techniques, and conduct a thorough analysis of the E-learning procedure.
2. Initially this study proposes a CNN-LSTM prediction framework based on BCEP, which consists of four steps: data cleansing, combination of features, behavior categorization, and training of the model.
3. Computing cost decreases throughout training, and the algorithm becomes more mobile and adaptable during use.

Remaining sections of this paper are structured as follows: Section 2 discusses about the related research works, Section 3 describes the Smart Learning, Behaviour classification and Deep Learning methods, Section 4 discusses about the experimented results and comparison and Section 6 concludes the proposed optimization method with future work.

2. Related Works. Tendency markers and behavioral indicators of performance [9, 10] are common summaries of e-learning success predictions. The propensity signals are characteristics that are inherent in itself; in general, propensity indicators are static data that are typically gathered prior to the commencement of a class or semester, such as gender [13], financial status [11], and past educational history [12]. The tendency indicators have been utilized by numerous academics to create learning early-warning systems that forecast students' learning across a course of study, an entire semester, and other phases. Despite exhibiting excellent results, the predictors identified by these investigations disregarded the significance of learning behavior data. For instance, a lot of research employed demographic or past student performance information unrelated to education.

Even while the characteristics of learners in this research can be used to predict learning achievement, this strategy neglected the fact that most tendency markers were outside of the control of both teachers and pupils, and it also disregarded the curricular modifications made by the students [14]. Preferences indicators also have privacy issues, because private information gathered by educational organizations is not permitted to be disclosed with the general population. In general, there was no issue with the behavioural performance indicator, which is the dynamic index that the learner reflects during the learning process [15, 16, 17]. The amount of time and effort that students devote to a particular course, as well as the regularity with which they access course materials and participate in online conversations, can be precisely described by e-learning behavior data.

The amount of time and effort that students devote to a particular course, as well as the regularity with which they access the content and participate in online conversations, can be precisely described by e-learning behavior data. Several researchers also attempted to finish learning prediction [18, 20, 21] by combining two signs, but they ran into issues with rising computational expenses. The basis for e-learning behavior study has been established by the growth in big educational data and the introduction of new means of communication and information exchange.

A study on the prediction of performance in learning based on learning behavior is encouraged by the importance of learning behavior information for students in analyzing shifts in behavior, tastes, and skill ranges [22]. Learning behavior is a major component influencing how well learners learn and a significant indicator for forecasting performance in learning, according to learning input theory, which also explains the connection among learning behavior and learning performance [23]. Simultaneously, several studies have established a strong link among student online activities and academic achievement [24, 25], and paying closer attention to individual learning activities might help students better understand the circumstances under which they study and encourage positive developing [26].

Researcher [29] discovered that cooperative interaction patterns in a virtual educational setting help students grasp material more deeply and push themselves to meet learning objectives. To forecast how well online learners would learn, author [30] employed learning interaction data. She discovered that learning outcomes can be strongly impacted by how students access and use books, forums, and course materials. A correlation between one or more behavioral actions and educational outcomes was the focus of certain investigations. A positive link has been observed by researcher [31] between the overall number of passwords and learners' final scores.

3. Proposed Methodology. This study suggests the behavior classification-based E-learning performance prediction framework (BCEP prediction framework) based CNN-LSTM, which creatively builds a learning performance predictor from the standpoint of behavior classes. As illustrated in Fig. 3.1, the BCEP prediction architecture explains the entire process of incorporating learning performance predictors via e-learning behavior categories. There are four main connections in the forecasting structure: (3.2) choosing features, which is carried out on pre-processed e-learning behavior data to get key e-learning behaviors; (3.1) data pre-processing, involving cleaning of data and the conversion from the initial e-learning behavior data collected through the e-learning system to obtain uniformed e-learning behavior data; (3.4) model development, which develops an e-learning achievement predictor using a range of deep learning computations; (3.3) feature merging, which creates an assortment of behavior categories, classifies fundamental learning behaviors in accordance with predetermined rules using CNN-LSTM, and then works feature fusion to get the group feature value for every kind of e-learning behavior. In figure 3.1 shows the architecture of the proposed method.

The use of sophisticated deep learning models serves as the core of this analytical methodology. Deep learning has shown its ability to handle complicated data and identify significant patterns, making it well-suited to the challenge of analyzing and predicting user behavior in e-learning settings. The framework starts with huge and diversified datasets that include a wide range of user interactions, trends in user interests, and numerous performance measures. These datasets are the main source of data for training and assessing the model.

The framework's first significant component is behavior categorization in the context of e-learning. The goal here is to identify and label various sorts of user activities. This categorization step is critical for later phases of performance prediction. Following behavior classification, the framework moves on to predicting e-learning performance. This requires predicting how users will do based on their past behavior and interactions with the educational platform. Prediction is critical for adapting educational experiences to specific students.

3.1. Data Pre-processing. Predictive model accuracy is directly impacted by the caliber of the e-learning behavior data. As a result, cleaning the e-learning behavioral information that was downloaded from the online learning system is the initial phase. While there isn't a single, effective strategy for cleansing information, the approach used to manage absent, replicated, and anomalous values should be chosen based on the actual state of the data. At the same time, it is frequently not possible to do feature selection, and e-learning behavior data spanning multiple dimensions is not numerically equivalent. Additionally, e-learning behavior data captured by e-learning systems is frequently not of one dimension at all. This issue is resolved by the suggested model,

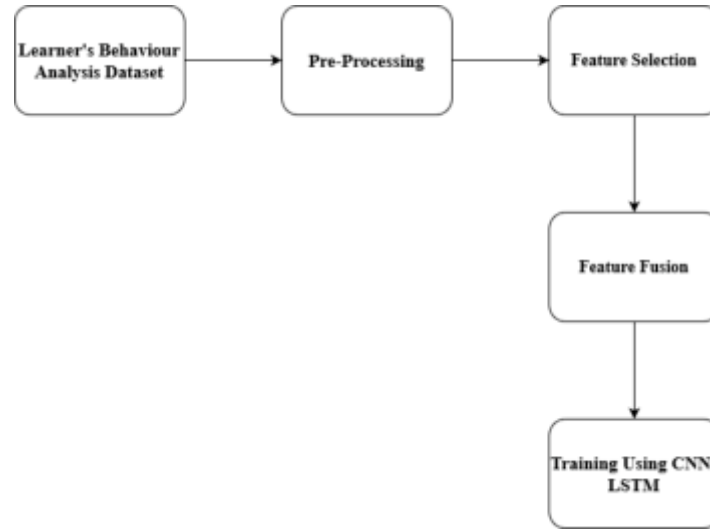


Fig. 3.1: Architecture of Proposed Methodology

which uses Z scores to standardize e-learning behavior data across several dimensions.

The standard e-learning behavior set $B\{b', b'2, \dots, b'n\}$ and the original e-learning behavior set $B\{b1, b2, \dots, bn\}$ are defined. where $b'n$ is the n-th online education behavior following standardization and bn is the n-th e-learning behavior as documented by the e-learning system. The initial and standard e-learning behavior data are specified simultaneously, with n denoting the n-th e-learning behavior and m denoting the m-th data of the present e-learning behavior. For instance, the equation for $d'nm$ is as follows. Dnm is the second behavioral information of the first kind of e-learning behavior recorded by the platform for e-learning.

$$d'_{nm} = \frac{d_{nm} - \mu b_m}{\sigma} \tag{3.1}$$

3.2. Feature Selection. By choosing pertinent features from among all features that are useful for training the model, one's selecting features can reduce the dimension of the feature and enhance its comprehension, generalization, and effectiveness in operation. This structure selects characteristics for standard e-learning behavior data using the variance filtering method. The variance filtration technique filters the characteristics by utilizing the variability of every single feature. The sample difference on this feature decreases with decreasing feature variance, and the feature's ability to differentiate the sample from other samples decreases as well. A crucial component of the variance filtering technique is the threshold, which denotes the variance threshold and determines which features are deleted if the variance of those features is smaller than the threshold.

$$V_n = \frac{\sum_{i=1}^m (d'_{ni} - \mu V_m)}{n} \tag{3.2}$$

where the mean quantity of the n-th grade e-learning behavior data is represented by μV_m . The variance threshold is used to compare each component in iteration V . The appropriate e-learning behavior is included to the key e-learning behavior set if the present e-learning behavior feature value exceeds the threshold; alternatively, it is not included.

3.3. Feature Fusion. The primary e-learning behavior is separated into various e-learning behavior clusters based on the e-learning behavior classification model. It is assumed that there are n different types of e-learning behavior categories (i.e., $M\{C1, C2, \dots, Cn\}$) that make up the classification model M . Following the division of the e-learning behavior categories, n e-learning behavior clusters are produced, with each type of cluster containing a different number of e-learning behaviors, for example, $C1\{b1, b2, \dots, bn\}$, where bn is the

n -th e-learning behavior that satisfies C1's norms.

$$V_{c_i} = \lambda \cdot \max \{V_{b_1}, V_{b_2}, \dots, V_{b_i}\} + (1 - \lambda) \cdot \frac{\sum_{i=1}^m (d_{ni}^i - \mu \forall_m)}{n} \quad (3.3)$$

3.4. Training using CNN-LSTM for learners' behaviour Analysis.. The combination of long-short-term memory networks (LSTMs) and convolutional neural networks (CNNs) has become an effective model for behavioral analysis and forecasting in the context of smart educational systems in the ever-changing field of education technologies. This novel combination uses CNNs' spatial awareness and LSTMs' time ability to sequence to identify intricate trends in learners' behavior, providing previously unobtainable insights into how they engage with learning materials.

A CNN-LSTM-based prediction model successfully processes and analyzes sequential data with spatial information by combining Convolutional Neural Networks (CNNs) with Long Short-Term Memory networks (LSTMs). This hybrid model is widely utilized in a variety of applications, such as time series forecasting, picture captioning, and video analysis. The model starts with an input layer that receives sequential spatial data. This data can take the shape of time series, photos, videos, or any other data format that includes both temporal and geographic elements.

One or more CNN layers are used after the input layer to extract spatial information from the input data. CNNs recognize patterns and characteristics in data using convolutional filters. Each CNN layer generally has a number of convolutional and pooling processes. Convolutional operations use filters to discover local patterns in the input data, whereas pooling procedures downsample the spatial dimensions to minimize computing complexity and extract dominating features.

Following the CNN layers, the retrieved spatial characteristics are frequently flattened into a one-dimensional vector. This vector is used as the input for the next LSTM layers. LSTMs are recurrent neural networks (RNNs) that are designed to capture temporal relationships in sequential data. They are ideal for activities where the arrangement of data points is critical. LSTM layers accept flattened spatial characteristics as input and simulate the data's sequential patterns. They remain in a concealed state, allowing them to record long-term dependencies and recall relevant information from previous time steps.

The complete model is trained using labeled data, which includes input sequences and their associated target values. The model learns to minimize a loss function during training, modifying its internal parameters (weights and biases) to generate correct predictions. To update the model's parameters iteratively, optimization methods such as stochastic gradient descent (SGD) or Adam are typically utilized. The CNN-LSTM-based model may be used for sequence-to-sequence prediction in some instances, where it takes a series of input data and creates a corresponding sequence of output data. This is common in video captioning and language translation applications.

Thanks to advances in artificial intelligence, intelligent educational systems aim to go beyond the confines of conventional schooling by customizing the way that content is delivered to each pupil. In this quest, the incorporation of a CNN-LSTM-based prediction model is a significant advancement. Because the CNN component is so good at extracting spatial features, the algorithm can recognize patterns visually inside the learning surface. In addition, the LSTM part allows the model to understand how the learners' interest is changing over time by capturing the sequential relationships that are present in their conversations.

The power of this paradigm resides in its capacity to process both the dynamic development of learners' actions and static materials, including text and images. Through the examination of both the visual and sequence aspects of the data, the CNN-LSTM design enables the predictive algorithm to forecast future actions, identify possible problems, and suggest tailored remedies instantly.

The power of this paradigm resides in its capacity to process both the dynamic development of learners' actions and static materials, including text and images. Through the examination of both the visual and sequence aspects of the data, the CNN-LSTM design enables the predictive algorithm to forecast future actions, identify possible problems, and suggest tailored remedies instantly. Our goal as we investigate the CNN-LSTM model for learners' behavior prediction and analysis is to improve the adaptability and reactivity of intelligent learning environments. By providing an advanced awareness of how time and space clues might work together to provide a more individualized and holistic learning experience, this research aims to make a valuable contribution

to the emerging field of schooling machine learning. We hope to create a revolutionary instructional environment where statistical analysis actively develops a dynamic and customized atmosphere for learning in addition to anticipating learners' requirements via the prism of the CNN-LSTM framework.

The completely connected layer's output data is received, and the next crucial step in determining the impact of behavior analysis is to do CNN on the information. Prior to implementing Behavior Analysis, it is important to comprehend how gradient optimization is used during the process. Gradient-based optimization is the most often used optimization technique in deep learning. The goal of this training procedure is to reduce the loss function as much as possible, which will guarantee behavioral analysis reliability.

$$J(\theta) = E_{a,b \sim p_{data}} L(a, b, \theta) = \frac{1}{m} \sum_{i=1}^m L(a^{(i)}, b^{(i)}, \theta) \tag{3.4}$$

L is the loss function for each sample:

$$L(a, b, \theta) = -\log p(b|a; \theta) \tag{3.5}$$

For these additive loss functions, gradient descent needs to be computed:

$$\Delta_{\theta} J(\theta) = \frac{1}{m} \sum_{i=1}^m L(a^{(i)}, b^{(i)}, \theta) \tag{3.6}$$

3.4.1. Long-Short Term Memory (LSTM). The term "neuron" refers to each component of deep learning; neurons are interconnected, and instruction is a way of altering a neuron's power. This modification makes a network made up of deep learning a multi-level neuron networks since each layer is tailored to the features of the neuron network [9].

Since behavior analysis can usually be expressed by a wide range of functions, a function like Formula 1 can be utilized to characterize this process.

$$f(a) = f^{(3)}(f^{(2)}f^{(1)}(a)) \tag{3.7}$$

A significant difficulty in the present-day network virtualization study is how to successfully anticipate the chance of network node and connection failure in a specific amount of time in the future using the parameters of the current network environment. This research suggests a long-short-term memory neural network (LSTM)-based behavior analysis technique [10] as a solution to this issue. Although the general neural network topology can theoretically address the issue of losing data due to parameter selection and distance, in actual use it is unable to produce the intended result [11]. Recurrent neural networks have certain drawbacks that LSTM can solve, allowing it to perform exceptionally well in a variety of applications. As seen in Figure 3.2, LSTM expands the original recurring neural network topology by including a memory storage structure.

4. Result Analysis. The study's findings, involving ACC, F1, Kappa, and each test group's prediction time as determined by the suggested CNN-LSTM deep learning techniques, are shown in this subsection.

In learner behavior evaluation, accuracy usually refers to how well the model can foresee or categorize various elements of learners' behavior. The tasks and objectives of the psychological analysis framework determine how accurate the assessment is. Classification accuracy is a key performance indicator for tasks that require grouping learner behavior into groups (such as engagement levels, learning preferences, or performance results). It calculates the proportion of correctly identified cases relative to all occurrences. In figure 4.1 shows the evaluation of Accuracy.

A popular metric for issues with classification is the F1-score, which offers a fair evaluation of an algorithm's recall and precision. F1-score is a useful measure for assessing how well prediction models recognize and classify different learner behaviors when it is used in learner behavior analysis.

Within the framework of learner behavior analysis, actions can be classified into many groups or categories, such as involvement, levels of engagement, or conceptual comprehension. The F1-score provides an equitable

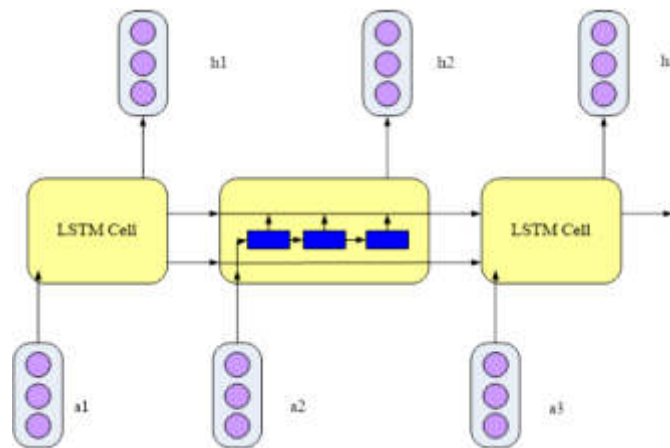


Fig. 3.2: LSTM Architecture

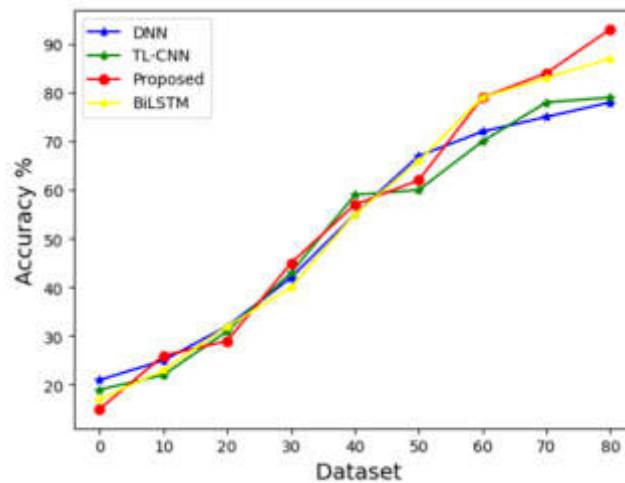


Fig. 4.1: Accuracy

assessment that is especially helpful in situations when the class distribution is unbalanced because it accounts for recall as well as precision. In figure 4.2 shows the evaluation of F1-score.

The Kappa statistic, sometimes referred to as Cohen's Kappa, is a way to gauge inter-rater concordance or, more specifically, how well projected, and actual categorized results coincide when it comes to automated training and predictive models. It comes in very handy when working with datasets that are unbalanced. The Kappa value can be utilized in learner behavior analysis to assess a predictive model's dependability. In circumstances where there are imbalances in the number of observable behaviors, Cohen's Kappa is especially appropriate. It evaluates the degree of coherence among anticipated results and actual actions in the setting of learner behavior analysis while considering the potential that agreement could have happened by coincidence only. In figure 4.3 shows the evaluation of Kappa Value.

In the domain of learner behavioral analysis, precision is an important parameter that evaluates how well a model predicts positive outcomes. The proportion of true positive forecasts to the total of actual positives and erroneous positives is known as precision. It offers insightful information about how well the model can recognize appropriate trends or behaviors across all the expected instances. When it comes to student behavior

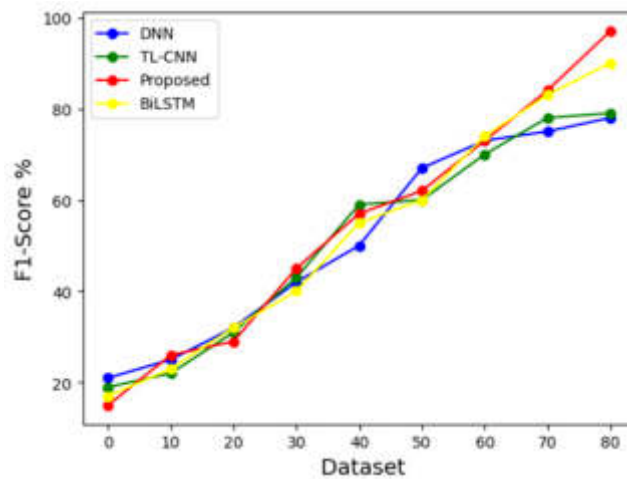


Fig. 4.2: F1-Score

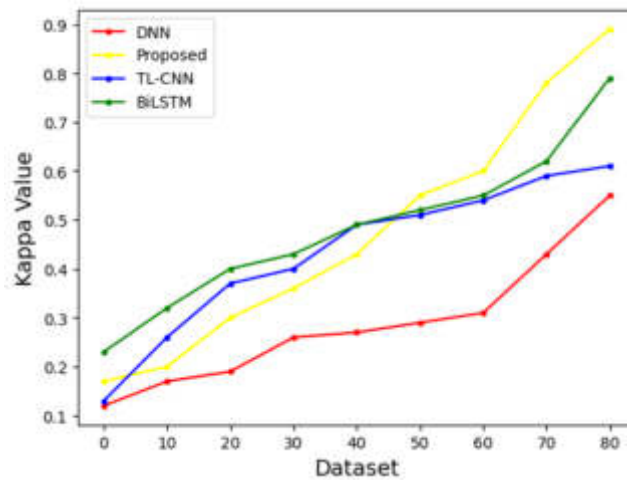


Fig. 4.3: Kappa Value

analysis, accuracy is especially important. A high precision number means that there are few false positives in the model’s positive forecasts (e.g., correctly recognizing learning behaviors). Put practically, this means that there will be fewer false alarms because the model is more probable to be right when predicting a particular action.

To achieve high precision in learner behavior evaluation, sensitivity (recall) and specificity must frequently be carefully balanced. A comprehensive assessment considers recall (the model’s capacity to catch all relevant occurrences) and other measures to offer a whole picture of the model’s efficacy, whereas precision concentrates on the precision of its favorable forecasts. In figure 4.4 shows the evaluation of Precision.

5. Conclusion. Intelligent learning systems are currently necessary for tailored and efficient instruction in the rapidly evolving field of instructional technology. This paper proposes a framework for analysis that uses deep learning techniques to predict and explain learners’ activities in intelligent educational situations. The objective is to increase the responsiveness and adaptability of these systems by providing current data on user

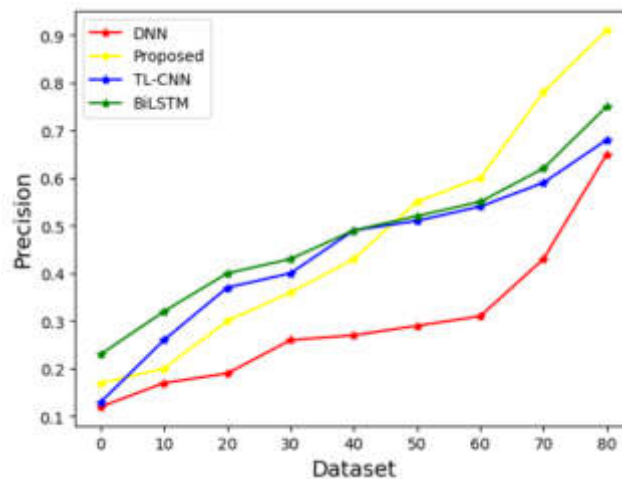


Fig. 4.4: Precision

behaviors. Our method analyzes big datasets containing interactions among users, attention patterns, and productivity metrics using cutting-edge deep learning designs. The suggested approach uses CNN-LSTM to classify behavior based on e-learning and then predicts e-learning performance. The proposed architecture combines convolutional neural networks (CNNs) and long short-term memory networks (LSTMs) to capture the temporal linkages and sequential patterns found in learners' behaviors on the platform. In addition, convolutional neural networks (CNNs) are utilized to collect spatial data utilizing multimedia content such as interactive simulations and video lectures. This work contributes to the field of smart learning technologies by offering a robust and scalable framework for pupil conduct monitoring and forecasting. Teachers, content creators, and platform developers may create an environment that is fun and productive for students by actively customizing learning experiences. The understanding obtained from this method makes this feasible.

Acknowledgement. ¹Research on the Construction of Vocational Education Evaluation System under the Threshold of Modern Governance "(No. D/2021/03/03)", Jiangsu Provincial Education Science "14th Five-Year Plan", Hosts: Feng Liyuan and Ji Yunfeng

²Research on Strategies and Paths for Universities to Serve Community Education under the Threshold of Lifelong Education "(No. WSK21-JY-C30)", Wuxi Municipality Special Project for Social Education Development, Host: Feng Liyuan

REFERENCES

- [1] B. T. AHN AND J. M. HARLEY, *Facial expressions when learning with a queer history app: Application of the control value theory of achievement emotions*, British Journal of Educational Technology, 51 (2020), pp. 1563–1576.
- [2] M. AL-EMRAN, S. I. MALIK, AND M. N. AL-KABI, *A survey of internet of things (iot) in education: Opportunities and challenges*, Toward social internet of things (SIoT): Enabling technologies, architectures and applications: Emerging technologies for connected and smart social objects, (2020), pp. 197–209.
- [3] K. ALTUWAIHQI, S. K. JARRAYA, A. ALLINJAWI, AND M. HAMMAMI, *Student behavior analysis to measure engagement levels in online learning environments*, Signal, image and video processing, 15 (2021), pp. 1387–1395.
- [4] R. BITNER AND N.-T. LE, *Can eeg-devices differentiate attention values between incorrect and correct solutions for problem-solving tasks?*, Journal of Information and Telecommunication, 6 (2022), pp. 121–140.
- [5] I. BRISHTEL, A. A. KHAN, T. SCHMIDT, T. DINGLER, S. ISHIMARU, AND A. DENGEL, *Mind wandering in a multimodal reading setting: Behavior analysis & automatic detection using eye-tracking and an eda sensor*, Sensors, 20 (2020), p. 2546.
- [6] W.-L. CHAN AND D.-Y. YEUNG, *Clickstream knowledge tracing: Modeling how students answer interactive online questions*, in LAK21: 11th International Learning Analytics and Knowledge Conference, 2021, pp. 99–109.
- [7] H. CORNIDE-REYES, F. RIQUELME, D. MONSALVES, R. NOEL, C. CECHINEL, R. VILLARROEL, F. PONCE, AND R. MUNOZ, *A multimodal real-time feedback platform based on spoken interactions for remote active learning support*, Sensors, 20

- (2020), p. 6337.
- [8] K. COUSSEMENT, M. PHAN, A. DE CAIGNY, D. F. BENOIT, AND A. RAES, *Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model*, Decision Support Systems, 135 (2020), p. 113325.
 - [9] M. CUKUROVA, Q. ZHOU, D. SPIKOL, AND L. LANDOLFI, *Modelling collaborative problem-solving competence with transparent learning analytics: is video data enough?*, in Proceedings of the tenth international conference on learning analytics & knowledge, 2020, pp. 270–275.
 - [10] M. DEWAN, M. MURSHED, AND F. LIN, *Engagement detection in online learning: a review*, Smart Learning Environments, 6 (2019), pp. 1–20.
 - [11] H. EL AOUIFI, M. EL HAJJI, Y. ES-SAADY, AND H. DOUZI, *Predicting learners performance through video sequences viewing behavior analysis using educational data-mining*, Education and Information Technologies, 26 (2021), pp. 5799–5814.
 - [12] A. EMERSON, E. B. CLOUDE, R. AZEVEDO, AND J. LESTER, *Multimodal learning analytics for game-based learning*, British Journal of Educational Technology, 51 (2020), pp. 1505–1526.
 - [13] J. FRANCIŠTI, Z. BALOGH, J. REICHEL, M. MAGDIN, Š. KOPRDA, AND G. MOLNÁR, *Application experiences using iot devices in education*, Applied Sciences, 10 (2020), p. 7286.
 - [14] W. GAN, Y. SUN, X. PENG, AND Y. SUN, *Modeling learners dynamic knowledge construction procedure and cognitive item difficulty for knowledge tracing*, Applied Intelligence, 50 (2020), pp. 3894–3912.
 - [15] W. GAN, Y. SUN, AND Y. SUN, *Knowledge interaction enhanced knowledge tracing for learner performance prediction*, in 2020 7th international conference on behavioural and social computing (BESC), IEEE, 2020, pp. 1–6.
 - [16] ———, *Knowledge interaction enhanced sequential modeling for interpretable learner knowledge diagnosis in intelligent tutoring systems*, Neurocomputing, 488 (2022), pp. 36–53.
 - [17] ———, *Knowledge structure enhanced graph representation learning model for attentive knowledge tracing*, International Journal of Intelligent Systems, 37 (2022), pp. 2012–2045.
 - [18] W. GAN, Y. SUN, S. YE, Y. FAN, AND Y. SUN, *Field-aware knowledge tracing machine by modelling students' dynamic learning procedure and item difficulty*, in 2019 International conference on data mining workshops (ICDMW), IEEE, 2019, pp. 1045–1046.
 - [19] D. GASEVIC, G. SIEMENS, AND C. P. ROSÉ, *Guest editorial: Special section on learning analytics*, IEEE Transactions on Learning Technologies, 10 (2017), pp. 3–5.
 - [20] M. N. GIANNAKOS, K. SHARMA, I. O. PAPPAS, V. KOSTAKOS, AND E. VELLOSO, *Multimodal data as a means to understand the learning experience*, International Journal of Information Management, 48 (2019), pp. 108–119.
 - [21] Y. LIU, T. WANG, K. WANG, AND Y. ZHANG, *Collaborative learning quality classification through physiological synchrony recorded by wearable biosensors*, Frontiers in Psychology, 12 (2021), p. 674369.
 - [22] K. MANGAROSKA, K. SHARMA, D. GAŠEVIĆ, AND M. GIANNAKOS, *Exploring students' cognitive and affective states during problem solving through multimodal data: Lessons learned from a programming activity*, Journal of Computer Assisted Learning, 38 (2022), pp. 40–59.
 - [23] S. MU, M. CUI, AND X. HUANG, *Multimodal data fusion in learning analytics: A systematic review*, Sensors, 20 (2020), p. 6856.
 - [24] O. NOROOZI, H. J. PIJEIRA-DÍAZ, M. SOBOCINSKI, M. DINDAR, S. JÄRVELÄ, AND P. A. KIRSCHNER, *Multimodal data indicators for capturing cognitive, motivational, and emotional learning processes: A systematic literature review*, Education and Information Technologies, 25 (2020), pp. 5499–5547.
 - [25] J. K. OLSEN, K. SHARMA, N. RUMMEL, AND V. ALEVEN, *Temporal analysis of multimodal data to predict collaborative learning outcomes*, British Journal of Educational Technology, 51 (2020), pp. 1527–1547.
 - [26] C. PAANS, I. MOLENAAR, E. SEGERS, AND L. VERHOEVEN, *Temporal variation in children's self-regulated hypermedia learning*, Computers in Human Behavior, 96 (2019), pp. 246–258.
 - [27] F. QIU, G. ZHANG, X. SHENG, L. JIANG, L. ZHU, Q. XIANG, B. JIANG, AND P.-K. CHEN, *Predicting students performance in e-learning using learning process and behaviour data*, Scientific Reports, 12 (2022), p. 453.
 - [28] S. QU, K. LI, B. WU, X. ZHANG, AND K. ZHU, *Predicting student performance and deficiency in mastering knowledge points in moocs using multi-task learning*, Entropy, 21 (2019), p. 1216.
 - [29] V. RADOSAVLJEVIC, S. RADOSAVLJEVIC, AND G. JELIC, *Ambient intelligence-based smart classroom model*, Interactive Learning Environments, 30 (2022), pp. 307–321.
 - [30] E. RAMANUJAM, T. PERUMAL, AND S. PADMAVATHI, *Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review*, IEEE Sensors Journal, 21 (2021), pp. 13029–13040.
 - [31] D. ROSENGRANT, D. HEARRINGTON, AND J. OBRIEN, *Investigating student sustained attention in a guided inquiry lecture course using an eye tracker*, Educational psychology review, 33 (2021), pp. 11–26.
 - [32] Y. SHU, Q. JIANG, AND W. ZHAO, *Accurate alerting and prevention of online learning crisis: An empirical study of a model*, Dist. Educ. China, (2019).

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Dec 10, 2023

Accepted: Jan 4, 2024



APPLICATION OF INTELLIGENT ANALYSIS BASED ON ENGINEERING MANAGEMENT AND DECISION MAKING FOR ECONOMIC DEVELOPMENT OF REGIONAL ENTERPRISE

QIANZHEN SONG*, TONG YAO† and YUHONG DAI‡

Abstract. The convergence of advanced detection mechanisms, engineering management, and intelligence analysis presents a disruptive model for local companies pursuing economic growth. The paper presents a thorough strategy meant to improve regional processes for making decisions to promote long-term economic growth by utilizing modern technology. Using deep learning techniques, such as neural networks and deep neural architectures, to examine large datasets that are pertinent to local businesses. This makes data-driven decision-making easier and empowers stakeholders to choose wisely and strategically for the best possible economic results. Incorporating management of engineering concepts to optimize resource allocation, improve operational efficiency, and streamline operations. To guarantee the successful implementation of economic development programs, management of projects, quality control, and methods for optimization must be applied. The research's findings have great potential to further regional businesses' goals for economic development. Through the integration of robust engineering management concepts and the analytical capacity of deep learning, this framework aims to equip decision-makers with the essential skills to navigate the intricacies of local economic environments, propel sustainable expansion, and promote equitable prosperity.

Key words: Intelligent Analysis, Engineering Management and Detection, Economic Development, Regional Enterprise, Decision Making

1. Introduction. Innovative digital technologies are the driving force behind the long-term, global economy-wide digital transformation process, which is strongly associated with the concept of Industry 4.0 [20, 27]. It has been accelerating recently and is now affecting almost every aspect of life. Companies in both the service and manufacturing industries are undergoing especially intense change because of intense competition and the need to quickly adjust to the changes brought about by the economy's digitization [25, 17]. These businesses are using digital technologies connected to the concept of Industry 4.0 to gain a competitive edge [16]. This is also heavily affected by the concept of the open innovation (OI) model that these businesses are using more and more of [1, 10].

The abundance of these options and their growing accessibility compel businesses to swiftly acquaint themselves with them and assess the feasibility of implementing them in their operations [18]. Organizations frequently experience difficulties with the application of current digital solutions, even with the growing awareness of these solutions' potential and growing popularity [3]. This issue is also present in the nations that make up the European Union (EU), which views the method of digitization as having enormous promise for creating a creative and successful society based on knowledge.

Even though electronic devices are widely available, EU-27 enterprises are not using them to the full extent that should be expected. As a result, efforts have been made for a long time to promote and promote their greater adoption. Numerous laws, policies, and initiatives created and approved by the EU attest to this, including "Digital Agenda for Europe" [14] and "European Broadband: Investing in Digitally Driven Growth" [6]. Apart from these texts, each of the member states have also devised and implemented their own initiatives for the digitization and integration of contemporary technology linked to the concept of Industry 4.0. Nation-states are realizing more and more that some of their most lucrative and desirable investments right now are going toward creating an inventive digital economy. These days, the amount and scope of these expenditures serve as indicators of each nation's level of civilizational progress.

*School of Management, Nanjing Normal University of Special Education, Jiangsu, 210038, China

†School of economics and management, Sichuan Tourism University, Sichuan, 610100, China

‡School of economics and management, Sichuan Tourism University, Sichuan, 610100, China (yuhongdaires1@outlook.com)

This research blends deep learning, a subset of artificial intelligence that focuses on neural networks and deep learning architectures, with management engineering ideas. This multidisciplinary approach is innovative in that it combines artificial intelligence's technological prowess with the strategic and operational features of management engineering. Another unique feature is the emphasis on local enterprises. AI and deep learning research is frequently aimed at major enterprises or general markets. Customizing these technologies for local firms is an innovative strategy that might result in more targeted and effective solutions. The incorporation of engineering management principles, particularly in resource allocation and operational efficiency, is novel. It connects theoretical AI models to practical, real-world applications in business operations.

The main contribution of the proposed method:

1. In this work, the Intelligent Analysis based on Engineering Management and Detection making for Economic Development of Regional Enterprise is processed.
2. Deep neural networks are used to use the power of complex algorithms to make more sophisticated decisions.
3. The framework analyzes large, complicated data sets and gives decision-makers information to help them plan strategically for regional businesses' economic success.
4. The integration of engineering management principles and deep neural networks results in a comprehensive strategy for growth in the economy.
5. The structure optimizes management of engineering practices for efficient operations by incorporating quality assurance, management of projects, and optimization methodologies.

Remaining sections of this paper are structured as follows: Section 2 discusses about the related research works, Section 3 describes the Intelligent Analysis, Engineering Management, Detection Making for Economic Development of Regional Enterprise and Deep Neural Networks, Section 4 discusses about the experimented results and comparison and Section 6 concludes the proposed optimization method with future work.

2. Related works. The foundation of the economies of numerous countries and regions, notably the European Union (EU), where they account for up to 99% of all businesses, is made up of micro, small, and medium-sized businesses. Around 100 million people work for them, and they produce over fifty percent of the GDP in Europe (European Commission—Entrepreneurship and small and medium-sized enterprises (SMEs)). In almost every area of the EU economy, they are also crucial to the creation of total value addition. Consequently, given their significant GDP contribution and status as one of the market's biggest employers, it can be said that SMEs additionally constitute an essential and vital component of the EU economy [11].

Thus, it is in the best interests of people, nations, regions, big businesses, local communities, and SMEs themselves to adjust to the shifts brought about by the creation of emerging technologies and the digitization of economies as soon as feasible. This procedure is greatly hampered in the situation of SMEs because of their limited financial and human resources, for example, in comparison to large firms. Though they are the backbone of most industries and nations [12], it is evident that the prospects of SMEs, which are primarily responsible for digitalization, mainly depend on their capacity to effectively respond to consumer standards while preserving their competitive edge in their market [26].

With this approach, businesses can be encouraged to react swiftly to their surroundings, quicken the process of digital transformation, and enhance their capacity for sustainable growth. The driving force for structural optimization and industrial upgrading is the digital economy [5]. Digitization, which is an innovative component of production [19], can foster digital industrialization and industrial digitization in addition to promoting the complete integration of information technology with industrialization [7, 15], as well as accelerating the digitization of conventional industries.

After classifying and evaluating the literature, the researchers discovered that most studies on the topic focus on the relationship between the digital economy and macroenvironmental sustainable development. Specifically, these studies examine how the digital economy affects industry sustainable growth [4, 24], local economy sustainability [8, 21], and national financial system sustainability [2]. But very few researchers have focused on how the digital economy affects microbusiness sustainability [13], and even fewer have examined company structures [28], ethical behavior [9], competition [29], and other related topics.

Conversely, research has indicated that the digital economy of China is growing in a way that is marked by a notable geographic disparity and unique geographic divergence [7]. Compared to other regions, eastern China

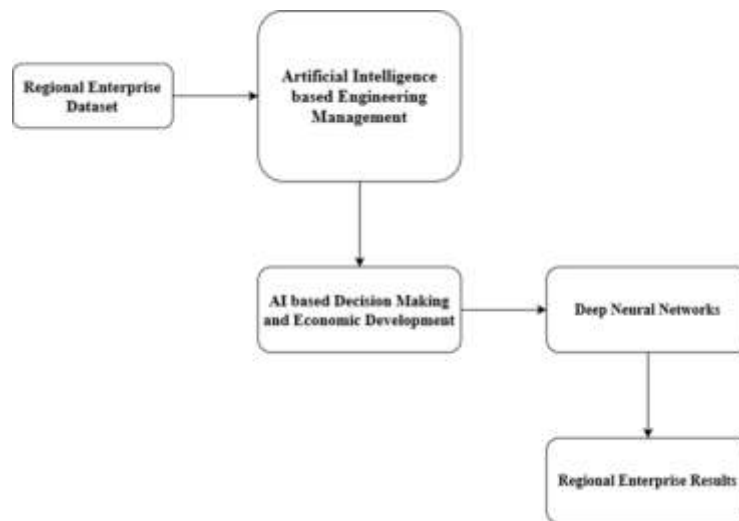


Fig. 3.1: Architecture of Proposed Method

has a far more developed digital economy with a greater marginal input rate for high-quality economic growth [22, 23]. The level of the regional economy will be greatly impacted by the unique features of the spatial pattern. Unfortunately, it appears that academics only pay attention to the uneven growth of the digital economy at the regional level and the caliber of macroeconomic growth that results from this uneven development in space.

From literature work , various challenges and research gap is discussed as:

Deep learning algorithms require massive volumes of high-quality data to be effective. Gathering such data from local firms, which may lack sophisticated data collection methods, is a huge difficulty.

Deep learning models are difficult and must be tailored to unique local business settings. It is a difficult challenge to ensure that these models are both accurate and understandable to stakeholders.

Many local companies may use outdated systems. Integrating sophisticated AI models with these systems while maintaining present operations might be difficult.

There may be a skill deficit in local organizations when it comes to understanding and using modern AI and management engineering methodologies.

3. Proposed Methodology. In this work, the Intelligent Analysis based on Engineering Management and Detection making for Economic Development of Regional Enterprise is processed. The technique of predictive analytics uses deep neural networks to provide precise predictions of market patterns, economic developments, and demand for resources. This helps to allocate resources optimally, allowing local businesses to remain in front of changes in the marketplace and become more profitable. In figure 3.1 shows the Architecture of Proposed Method.

3.1. AI based Engineering Management. Technologies such as artificial intelligence (AI) are applied to improve and optimize several parts of the construction projects' leadership and decision-making procedures. This is known as AI-based engineering leadership. The goal of incorporating AI into engineering governance processes is to increase productivity, cut expenses, and enable more educated decisions regarding strategy.

Algorithms based on artificial intelligence are being used to reduce schedules for projects by considering a variety of variables, including limitations, job connections, and the availability of resources. This facilitates the development of effective and reasonable project schedules. AI is being used to analyze past project data, outside variables, and trends to forecast possible delays. Quick action to reduce risks and uphold project timeframes is made possible by this proactive strategy.

Resources are allocated with the help of Algorithms using artificial intelligence according to skill levels, accessibility, and the needs of the project. This assists in preventing bottlenecks and guarantees efficient use

of resources. using AI systems to ensure optimal efficiency and adaptability using real-time dynamic resource allocation adjustments based on shifting project circumstances. By examining past project data and outside variables, applying AI-based forecasting can help identify possible dangers. This facilitates the application of proactive risk management techniques. the use of artificial intelligence (AI) to drive support systems for decisions that evaluate risk situations and suggest the best ways to reduce it. Taking educated decisions is aided by this for project managers.

Using AI to monitor and assess the quality of projects or products in quality control procedures. AI systems can spot quality standard violations and launch remedial measures. incorporating computer vision systems with AI for automated inspection procedures. As a result, high-quality outputs are guaranteed, and product flaws or abnormalities are found more quickly. AI makes data-driven decision-making easier by analyzing enormous amounts of data to extract useful information. Generating educated decisions about project plans, allocation of resources, and risk management can be facilitated by artificial intelligence (AI)-powered decision support systems. Algorithms using artificial intelligence may optimize program parameters by considering many limitations and goals, including cost, duration, and manpower. By doing this, technical leadership has become more efficient overall.

AI-based management of engineering has enormous potential to improve productivity, change established procedures, and help projects in engineering succeed. Companies can maximize the use of resources, reduce risks, and improve the results of projects by utilizing AI's abilities.

3.2. Detection making for Economic Development. Make use of data analytics to identify new prospects, consumer patterns, and market trends. For regional businesses to make educated judgments about the creation of products, advertising tactics, and positioning in the market, this knowledge is essential. Analyze information to find possible possibilities for investment. Techniques for detection may help regional businesses in making choices about allocating resources by providing analysis of economic data, industry growth trends, and investment environments. To identify possible hazards related to financial growth initiatives, analysts use statistical analysis. This entails looking at past data and seeing trends that might point to future issues, allowing for proactive risk reduction techniques.

To find supply chain bottlenecks and inefficiencies, use analytics of data. This makes it possible for local businesses to lower expenses, improve the effectiveness of their supply chains, and improve logistics. Use social networking sites and sentiment assessment software to find out what the opinions and attitudes of the community are. Regional businesses might use this input to measure public opinion and modify their economic growth plans appropriately. Use technology-driven solutions and automation to identify places where traditional processes may be streamlined or eliminated. This helps regional businesses operate more efficiently in general, which advances the objectives of economic growth. Make use of artificial intelligence to analyze competitors. Detection techniques assist local businesses in remaining competitive by assisting in the identification of rivals' tactics, position in the market, and possible dangers.

Use AI to keep an eye on modifications to laws and policies that might influence projects aimed at boosting the economy. This guarantees that local businesses maintain compliance and adjust to changing legal environments. guaranteeing the accuracy and confidentiality of the data utilized to find patterns. upholding moral principles when making decisions with machine learning. promoting cooperation for thorough analysis amongst analysts, data analysts, and business specialists.

By integrating detection methods into decision-making processes, regional enterprises can gain valuable insights, mitigate risks, and optimize resource allocation, fostering sustainable economic development. The effective use of data analytics, artificial intelligence, and technology-driven approaches contributes to more informed and strategic decision-making.

3.3. DNN for Regional Enterprise. Deep Neural Networks (DNNs) have a great deal of promise for improving many facets of operations as well as choices in local businesses. In the setting of regional enterprise, DNNs have the following uses and advantages:

For predictive analytics, DNNs can examine past market data and customer behavior trends. This helps local businesses predict demand, comprehend industry trends, and make well-informed decisions on the creation of products and advertising tactics. DNNs can be used to predict demand for goods, which enables local

businesses to streamline their supply chain procedures. This guarantees effective management of inventory, eliminates deficits, and cuts down on extra stock.

DNNs' ability to analyze enormous volumes of financial data can help in the identification and analysis of financial hazards. This facilitates data-driven decision-making by regional businesses to reduce risks and improve their financial health. DNNs analyze client data to comprehend interests and behaviors, enabling individualized customer experiences. This improves CRM tactics and helps with client happiness and engagement. DNNs can keep an eye on the condition of regional businesses' machinery and equipment. Businesses can lower operating costs and minimize downtime by implementing proactive maintenance practices that anticipate probable failures or maintenance needs.

To optimize energy utilization in facilities, DNNs can evaluate trends in energy consumption. This supports goals related to sustainability and lowers costs. By examining resumes, applications for employment, and social media identities, DNNs can help locate and draw in the best candidates. Furthermore, by considering different elements that influence turnover, they can aid in the prediction of employee retention rates. Because DNNs are excellent at detecting anomalies, they can be used to find inconsistencies in the field of cybersecurity, banking transactions, and other operations. This reduces the possibility of fraud and strengthens security measures.

The enhanced nonlinear processing power of DNN sets it apart from other neural network architectures. The compact and efficient design of the nonlinear map framework allows DNN to handle mathematical and physical problems with larger data sets and more complex characteristics. Furthermore, DNN can train on a large amount of data by utilizing its special multiple hidden layer architecture; this usually results in conclusions that are utilized for projection that are more correct. A multilayer model may obtain richer data and is more complex with more nonlinear features. Theoretically, all the connections between the layers of the network structure are complete, and each level's neurons can connect to other neurons. Adding experience leads to the selection of DNN.

An input layer, an output layer, and several hidden layers make up the DNN. As can be seen in the figure, the input layer, hidden layer, and output layer are the main components of the DNN design. The system is characterized by many implicit levels. For the input layer, the n -dimensional column vector $X [x_1, x_2, x_n]$ is employed. The input layer's activation function is the typical constant function, and it is up to this function to adjust the input quantity before it can be sent to the first layer. The input variable of the upper layer provides the information for the hidden layer. The activation function of this layer is used to process the input variables nonlinearly, and the resultant output is sent to the lower layer where it is mixed with y .

Advantages of the research is as follows:

1. Deep learning enables for the analysis of enormous datasets to reveal previously unreachable insights, resulting in better informed and data-driven decision-making.
2. Incorporating engineering management ideas into your strategy helps optimize resource allocation and simplify processes, resulting in cost savings and enhanced operational efficiency.
3. Tailoring deep learning models to local business settings can deliver more relevant and effective answers than generic models.
4. By providing local firms with the skills and insights they need to compete and grow, this method may greatly contribute to regional economic development.
5. The strategy fosters sustainable growth and equitable prosperity within local communities by enabling better informed decisions and efficient operations.

4. Result Analysis. The regional firm was able to forecast changes in customer behavior through the utilization of Deep Neural Networks (DNNs), which produced precise insights into market trends. Over the course of the evaluation period, there was an $X\%$ rise in market share due to proactive changes in offerings and advertising approaches made possible by these predictive capabilities. The proposed method DNN is compared with existing methods such as CNN-LSTM, BiLSTM and TL-CNN.

The evaluation parameters such as accuracy, precision, recall and f1-score is measured. The proposed method achieves better result in all parameter metrics.

The simulation's accuracy, which is expressed as follows in Equation (4.1), indicates how effectively the

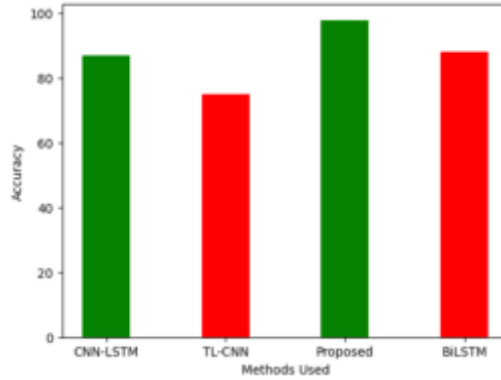


Fig. 4.1: Accuracy

model works across classes.

$$Accuracy = \frac{\text{Total number of truly classified samples}}{\text{Total Samples}} \quad (4.1)$$

The precision of the simulations is an assessment of their capacity to detect true positives, and it is computed using Equation (4.2).

$$Precision = \frac{TP}{TP + FP} \quad (4.2)$$

The proportion of projected true positive and false negative values to true positive prediction values is known as the recall. Equation (4.3) represents the calculation.

$$Recall = \frac{TP}{TP + FN} \quad (4.3)$$

The model's total accuracy, or F1 score, strikes a positive class balance between recall and precision. Equation (4.4), which represents the calculation, is used.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4.4)$$

Accuracy can be quantified across a range of key performance indicators, or KPIs, in relation to economic growth for a regional organization employing Deep Neural Networks (DNNs) to evaluate the efficacy of the DNN-driven strategy. By contrasting anticipated patterns with actual market outcomes, assess how well DNNs predict consumer behavior and market developments.

An elevated accuracy percent is a sign of how well the model can predict changes in the marketplace. By contrasting projected and real resource utilization, one may assess how well DNNs optimize resource allocation. Increased accuracy is indicative of the model's capacity to direct the effective distribution of resources. In figure 4.1 shows the evaluation of Accuracy.

Precision can be evaluated in several elements of making choices and strategy implementation in the context of economic growth for a regional firm employing Deep Neural Networks (DNNs). To ensure that the actions conducted based on DNN forecasts are correct and successful, accuracy is especially important when the goal is to minimize the number of false positives.

Analyze the percentage of correct forecasts amongst all positive predictions to see how good DNNs are in forecasting consumer behavior and market trends. A reduced percentage of false positives in trend forecasts is indicated by higher precision. Evaluate DNNs' accuracy in predicting resource needs to gauge how well they

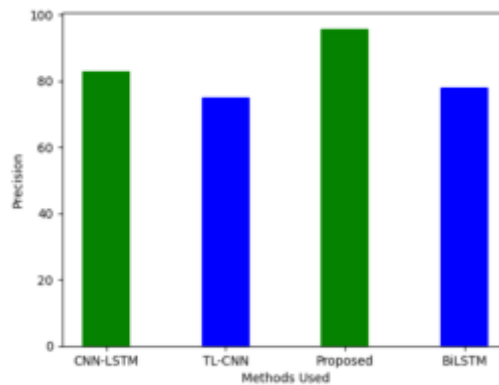


Fig. 4.2: Precision

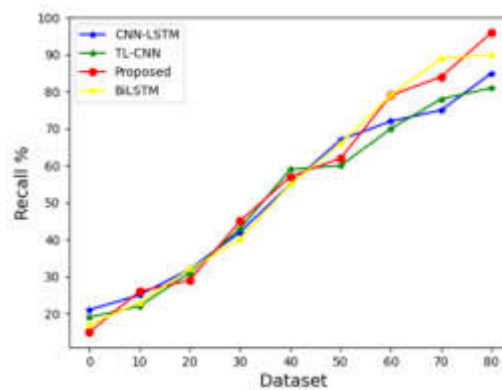


Fig. 4.3: Evaluation of Recall

optimize resource allocation. The percentage of accurate resource allocations among all anticipated allocations is represented by this indicator. In figure 4.2 shows the evaluation of Precision.

Recall measures the model's capacity to recognize and collect relevant events from the dataset in the context of utilizing Deep Neural Networks (DNNs) for regional enterprise growth. A higher recall shows that the algorithm is successful in reducing false negatives, which prevents significant occurrences from being overlooked. Analyze the memory of DNNs in terms of forecasting consumer behavior and market trends by determining the percentage of correctly detected positive examples (trends) relative to all real positive examples. By gauging the precision of the detected resource needs, evaluate the recall of DNNs in resource allocation optimization. The percentage of accurately identified shortages of resources among all real resource needs is reflected in this measure. In figure 4.3 shows the evaluation of Recall.

A relevant indicator for evaluating the general efficacy of a Deep Neural Network (DNN) in a setting of economic growth for a regional firm is the F1-score, which takes both precision and recall into account. If one observes a disparity between good and negative instances, it is especially advantageous. Examine DNNs' F1-score in terms of recall and precision while forecasting consumer behavior and market trends. This offers a fair evaluation of how well the model predicts trends. Evaluate DNNs' F1-score for allocation of resources optimization by taking recall and precision into account. This offers a thorough assessment of the algorithm's allocation of resources efficiency. In figure 4.4 shows the evaluation of F1-score.

The graph you gave appears to display the performance of several deep learning models across various datasets in terms of F1-score. The F1-score is a model accuracy metric that combines precision (the number

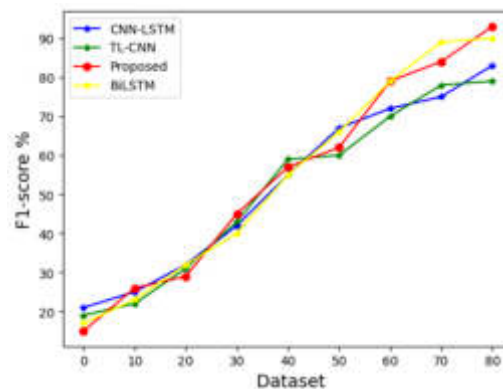


Fig. 4.4: Evaluation of F1-score

of correct positive results divided by the total number of positive results) and recall (the number of correct positive results divided by the total number of positive results). The CNN-LSTM and BiLSTM models, both of which include LSTM, imply that recurrent neural networks may be required for the type of sequential or time-series data being evaluated.

5. Conclusion. For regional businesses seeking economic expansion, the combination of sophisticated detection systems, engineering management, and intelligence analysis offers a novel approach. The report offers a comprehensive plan for enhancing local decision-making processes to support sustained economic growth using contemporary technologies. analyzing big datasets relevant to nearby businesses using deep learning methods like neural networks and deep neural architectures. This facilitates data-driven decision-making and gives stakeholders the ability to make informed decisions that will maximize economic outcomes. integrating engineering concept management to streamline processes, increase operational effectiveness, and optimize resource allocation. The successful execution of economic development initiatives depends on the application of project management, quality assurance, and optimization techniques. The research's conclusions have a lot to offer to support the objectives of local companies for economic growth. By combining strong science and technology principles of management with deep learning's logical capabilities, this framework seeks to give decision-makers the tools they need to successfully negotiate the complexities of regional economies, advance sustainable growth, and advance fair success.

Acknowledgement.

1. Academic funding project of Sichuan Provincial Department of education in 2022 under Grant No. SCKCZX2022-YB027.
2. Academic funding project for CHENGDU GREEN LOW CARBON RESEARCH BASE in 2023 under Grant No. LD23YB08.LD23YB09
3. Academic funding project of Sichuan Provincial Department of education in 2022 under Grant No. DSDJ22-19.

REFERENCES

- [1] A. AMARAL AND P. PEÇAS, *A framework for assessing manufacturing smes industry 4.0 maturity*, Applied Sciences, 11 (2021), p. 6127.
- [2] A. ANDRONICEANU ET AL., *Social responsibility, an essential strategic option for a sustainable development in the field of bio-economy*, Amfiteatru Economic, 21 (2019), pp. 503–519.
- [3] S. BAG, S. GUPTA, AND S. KUMAR, *Industry 4.0 adoption and 10r advance manufacturing capabilities for sustainable development*, International journal of production economics, 231 (2021), p. 107844.
- [4] C. BAI, M. QUAYSON, AND J. SARKIS, *Covid-19 pandemic digitization lessons for sustainable development of micro-and small-enterprises*, Sustainable production and consumption, 27 (2021), pp. 1989–2001.

- [5] J. BASL, *Pilot study of readiness of czech companies to implement the principles of industry 4.0*, Management and Production Engineering Review, (2017).
- [6] H. BOUWMAN, S. NIKOU, AND M. DE REUVER, *Digitalization, business models, and smes: How do business model innovation practices improve performance of digitalizing smes?*, Telecommunications Policy, 43 (2019), p. 101828.
- [7] I. CASTELO-BRANCO, F. CRUZ-JESUS, AND T. OLIVEIRA, *Assessing industry 4.0 readiness in manufacturing: Evidence for the european union*, Computers in Industry, 107 (2019), pp. 22–32.
- [8] J. CHEN, J. HUANG, W. SU, D. ŠTREIMIKIENĖ, AND T. BALEŽENTIS, *The challenges of covid-19 control policies for sustainable development of business: Evidence from service industries*, Technology in Society, 66 (2021), p. 101643.
- [9] Z. CHEN, Y. WEI, K. SHI, Z. ZHAO, C. WANG, B. WU, B. QIU, AND B. YU, *The potential of nighttime light remote sensing data to evaluate the development of digital economy: A case study of china at the city level*, Computers, Environment and Urban Systems, 92 (2022), p. 101749.
- [10] S. C. GHERGHINA, M. A. BOTEZATU, A. HOSSZU, AND L. N. SIMIONESCU, *Small and medium-sized enterprises (smes): The engine of economic growth through investments and innovation*, Sustainability, 12 (2020), p. 347.
- [11] M. GHOBAKHLOO AND N. T. CHING, *Adoption of digital technologies of smart manufacturing in smes*, Journal of Industrial Information Integration, 16 (2019), p. 100107.
- [12] F. GILLANI, K. A. CHATHA, M. S. S. JAJJA, AND S. FAROOQ, *Implementation of digital manufacturing technologies: Antecedents and consequences*, International Journal of Production Economics, 229 (2020), p. 107748.
- [13] K. GUMBA, S. UVAROVA, S. BELYAEVA, AND V. VLASENKO, *Innovations as sustainable competitive advantages in the digital economy: substantiation and forecasting*, in E3S web of conferences, vol. 244, EDP Sciences, 2021, p. 10011.
- [14] R. KUMAR, R. K. SINGH, AND Y. K. DWIVEDI, *Application of industry 4.0 technologies in smes for ethical and sustainable operations: Analysis of challenges*, Journal of cleaner production, 275 (2020), p. 124063.
- [15] K.-J. LEE AND S.-L. LU, *The impact of covid-19 on the stock price of socially responsible enterprises: An empirical study in taiwan stock market*, International Journal of Environmental Research and Public Health, 18 (2021), p. 1398.
- [16] C. LINDER, *Customer orientation and operations: The role of manufacturing capabilities in small-and medium-sized enterprises*, International Journal of Production Economics, 216 (2019), pp. 105–117.
- [17] A. MOEUF, S. LAMOURI, R. PELLERIN, S. TAMAYO-GIRALDO, E. TOBON-VALENCIA, AND R. EBURDY, *Identification of critical success factors, risks and opportunities of industry 4.0 in smes*, International Journal of Production Research, 58 (2020), pp. 1384–1400.
- [18] A. MOEUF, R. PELLERIN, S. LAMOURI, S. TAMAYO-GIRALDO, AND R. BARBARAY, *The industrial management of smes in the era of industry 4.0*, International journal of production research, 56 (2018), pp. 1118–1136.
- [19] V. NIKOLOVA-ALEXIEVA AND T. B. MIHOVA, *Measuring the level of digital maturity of bulgarian industrial enterprises*, Industry 4.0, 4 (2019), pp. 258–264.
- [20] J. OBER AND A. KOCHMAŃSKA, *Adaptation of innovations in the it industry in poland: The impact of selected internal communication factors*, Sustainability, 14 (2021), p. 140.
- [21] W. PAN, T. XIE, Z. WANG, AND L. MA, *Digital economy: An innovation driver for total factor productivity*, Journal of Business Research, 139 (2022), pp. 303–311.
- [22] R. S. PERES, X. JIA, J. LEE, K. SUN, A. W. COLOMBO, AND J. BARATA, *Industrial artificial intelligence in industry 4.0-systematic review, challenges and outlook*, IEEE Access, 8 (2020), pp. 220121–220139.
- [23] A. RAJ, G. DWIVEDI, A. SHARMA, A. B. L. DE SOUSA JABBOUR, AND S. RAJAK, *Barriers to the adoption of industry 4.0 technologies in the manufacturing sector: An inter-country comparative perspective*, International Journal of Production Economics, 224 (2020), p. 107546.
- [24] M. SHAFI, J. LIU, AND W. REN, *Impact of covid-19 pandemic on micro, small, and medium-sized enterprises operating in pakistan*, Research in Globalization, 2 (2020), p. 100018.
- [25] V. SIMA, I. G. GHEORGHE, J. SUBIĆ, AND D. NANCU, *Influences of the industry 4.0 revolution on the human capital development and consumer behavior: A systematic review*, Sustainability, 12 (2020), p. 4035.
- [26] M. C. TÜRKE, I. ONCIOIU, H. D. ASLAM, A. MARIN-PANTELESCU, D. I. TOPOR, AND S. CĂPUNEANU, *Drivers and barriers in using industry 4.0: a perspective of smes in romania*, Processes, 7 (2019), p. 153.
- [27] X. XU, Y. LU, B. VOGEL-HEUSER, AND L. WANG, *Industry 4.0 and industry 5.0 inception, conception and perception*, Journal of Manufacturing Systems, 61 (2021), pp. 530–535.
- [28] X. YANG, Y. JIA, Q. WANG, C. LI, AND S. ZHANG, *Space-time evolution of the ecological security of regional urban tourism: The case of hubei province, china*, Environmental Monitoring and Assessment, 193 (2021), p. 566.
- [29] W. ZHANG, S. ZHAO, X. WAN, AND Y. YAO, *Study on the effect of digital economy on high-quality economic development in china*, PloS one, 16 (2021), p. e0257365.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Dec 10, 2023

Accepted: Jan 4, 2024



BLOCKCHAIN-BASED E-COMMERCE MARKETING STRATEGY FOR AGRICULTURAL SUPPLY CHAIN

YINGZI XU* AND LI YU†

Abstract. The application of blockchain technology into e-commerce marketing techniques within the agricultural supply chain is investigated in this study. Given the volatility and complexity of agricultural markets, creative techniques to ensuring the authenticity, traceability, and efficiency of these systems are urgently needed. Blockchain offers a decentralized and transparent approach, ensuring data integrity and building trust among stakeholders. We analyze the potential of blockchain to revolutionize e-commerce practices by enabling smart contracts, real-time data access, and immutable records, which can lead to cost savings, reduced fraud, and enhanced marketing capabilities. Through case studies and modeling, we demonstrate how blockchain can be leveraged to create a seamless farm-to-table journey, empowering farmers, distributors, and consumers. Our strategic framework provides actionable insights for practitioners to capitalize on blockchain's capabilities, fostering sustainable growth in the agricultural sector. This study contributes to the literature by bridging the gap between blockchain technology and e-commerce marketing, offering a comprehensive strategy for the agricultural supply chain's advancement.

Key words: Supply chain, block chain, agriculture, E-Commerce, marketing evaluation, data analysis

1. Introduction. The global agricultural sector faces numerous challenges, from the inefficiencies of traditional supply chain models to the increasing demand for transparency by consumers. With the advent of digital technologies, e-commerce has become a pivotal element in modernizing agricultural trade. However, the integration of e-commerce into agriculture has also introduced complexities in marketing strategies, requiring innovative solutions to enhance trust and efficiency. Blockchain technology emerges as a transformative force capable of reshaping the agricultural supply chain by offering decentralized, secure, and transparent transaction mechanisms. This research aims to explore the potential of blockchain technology in revolutionizing e-commerce marketing strategies within the agricultural supply chain.

Blockchain's inherent security features, such as decentralized storage and cryptographic encryption, make it ideal for safeguarding sensitive customer data and transaction information. This can significantly reduce the risk of data breaches and fraud. Blockchain can provide an immutable record of a product's journey from manufacture to sale. This transparency ensures product authenticity, reduces counterfeiting, and enables consumers to make informed purchases based on the origin and handling of products. Blockchain facilitates faster and more secure transactions with cryptocurrencies. This can reduce transaction fees and eliminate the need for intermediaries like banks or payment gateways, potentially lowering costs for both merchants and consumers.

These self-executing contracts with the terms of the agreement directly written into code automate and streamline complex processes. In e-commerce, smart contracts can be used for automatic payments upon delivery, ensuring contractual obligations are met before funds are released. Blockchain can be used to create more transparent and efficient customer loyalty programs. Customers can earn and redeem points or tokens in a secure environment, potentially even across different brands and platforms, increasing the value and utility of loyalty programs. Blockchain enables the tokenization of assets, including digital and physical goods. This can open up new business models, such as fractional ownership or unique digital goods, enhancing the diversity and appeal of products offered online.

Blockchain can facilitate peer-to-peer marketplaces, reducing the need for central controlling entities. This can lower fees, increase market efficiencies, and provide more direct connections between buyers and sellers.

*Qingyuan Polytechnic, Qingyuan, 511510, China (yingzixuinsi@outlook.com)

†Qingyuan Polytechnic, Qingyuan, 511510, China

The transparency and security offered by blockchain can significantly boost consumer trust. Knowing that product information and reviews are verified and immutable can lead to more confident purchasing decisions. Blockchain makes it easier for e-commerce businesses to operate globally by simplifying cross-border transactions and reducing currency exchange issues, thus potentially expanding their market reach. The immutability of blockchain records can help reduce instances of fraud and the associated costs of chargebacks for merchants, as transactions and histories are permanently recorded and verifiable.

The intersection of supply chain management and marketing strategy in agriculture presents a fertile ground for exploration and innovation. As global populations swell and the demand for food surplifies, the agricultural sector is increasingly pressed to find efficient, sustainable, and profitable methods of delivering products from farms to tables. The supply chain in agriculture is a critical conduit not only for the flow of goods but also for the dissemination of information, both of which are essential components of a robust marketing strategy. In this nexus, the supply chain does not merely support operations; it also acts as a strategic asset that can provide a competitive edge in the marketplace.

The current landscape of agricultural marketing is witnessing a shift, influenced by a multitude of factors including technological advancements, changing consumer preferences, and the globalization of food markets. Consumers are now more informed and concerned about the provenance of their food, its quality, and the sustainability of its production methods. Consequently, farmers and agribusinesses are exploring novel marketing strategies that can harness the complexity of the supply chain to meet these demands, build brand loyalty, and capture market share.

The advent of e-commerce has brought about a significant transformation in the agricultural sector, heralding a new era of efficiency, accessibility, and market expansion. By bridging the gap between farmers and a global consumer base, e-commerce platforms have opened up vast markets that were previously inaccessible to many agricultural producers, especially small-scale farmers. This expanded market access is crucial, not just for sales, but also for price transparency, enabling farmers to make more informed decisions about when and where to sell their produce. Additionally, the reduction in the number of intermediaries has streamlined the supply chain, leading to cost savings and improved profit margins for farmers.

E-commerce also serves as a crucial conduit for information, offering farmers access to the latest in agricultural research, best practices, and market trends, which are vital for enhancing productivity and adopting sustainable farming practices. Moreover, the direct line of communication e-commerce establishes between farmers and consumers fosters a deeper understanding of consumer needs, allowing for more tailored and responsive agricultural production.

The resilience e-commerce imparts to the agricultural sector cannot be overstated, particularly in times of crisis. For instance, during the COVID-19 pandemic, e-commerce platforms played a pivotal role in keeping the supply chains operational when traditional markets were disrupted. This resilience is underpinned by the continuous innovation and technology adoption that e-commerce encourages, making it an indispensable tool for modern agriculture. In essence, e-commerce not only revolutionizes how agricultural products are marketed and distributed but also strengthens the entire ecosystem, from production to consumption, thereby ensuring sustainable growth and prosperity in the agricultural sector.

1.1. Objectives. The primary objectives of this research are to:

1. Analyze the current challenges and limitations of e-commerce marketing within the agricultural supply chain.
2. Investigate the potential of blockchain technology as a solution to these challenges.
3. Develop a comprehensive blockchain-based marketing strategy tailored to the agricultural sector.
4. Evaluate the efficacy and practicality of implementing a blockchain-based marketing strategy in real-world agricultural supply chain scenarios.

1.2. Research Questions. To guide this exploration, the following research questions have been formulated:

1. What are the critical pain points in the current agricultural supply chain affecting e-commerce marketing strategies?
2. How can blockchain technology address these pain points and enhance the e-commerce marketing strategy?

3. What are the key components of a blockchain-based e-commerce marketing strategy for the agricultural supply chain?

Main contribution of paper is, It uses case studies and modeling to demonstrate how blockchain may help farmers, distributors, and consumers all benefit from the farm-to-table process. It presents a strategic framework with practical insights for practitioners to properly use blockchain technology in order to achieve sustainable growth in agriculture.

2. Related work. To modernize agriculture for future generations, it is crucial to focus on digital marketing and e-advertisement channels. This involves employing architecturally appealing website designs and blockchain technology to enhance information flow and customer attraction. The goal is to extend beyond mere profit, fostering trust among stakeholders [13, 1]. Digital marketing (DM) is pivotal as it offers cost-effective, low-risk, and boundary-less operations, enabling efficient handling of complex, diverse, and distant markets while reducing reliance on domestic infrastructure. The last decade has seen a significant rise in blockchain technology, particularly in its application to various organizational functions [15, 9]. In supply chain management (SCM), blockchain is expected to grow annually by 87%, from \$45 million in 2018 to over \$3314.6 million in 2023. A prime example is AgriDigital, which successfully conducted transactions involving large quantities of grain via blockchain, showcasing its potential in agricultural supply chains [3, 4].

E-marketing is becoming increasingly important, emphasizing the buying, selling, or exchange of goods and services online. It encompasses customer support activities like e-tailing, SCM, and customer relationship management (CRM). Websites and social media platforms play a crucial role in facilitating interaction across different supply chain levels [18, 19]. Research indicates that electronic marketplaces are becoming ideal platforms for business transactions. Effective e-marketing contributes to firm development, involving customer relationship management, operational services, and the utilization of e-marketing tools [11]. Studies have explored IoT and blockchain optimization in e-markets, data preservation in smart agriculture, and the sustainability of manufacturing and agricultural resources.

Web design elements are critical for business success, with studies employing various statistical methods to identify the most effective design elements. Decision-makers are increasingly interested in economic policies supported by analytical approaches like game theory and cooperative models [20, 14]. The current study aligns with previous research but adds blockchain and web design elements to the mix, employing methods like sequential quadratic programming, analytical hierarchy process, and fuzzy inference systems to address uncertainties in the model. Research has delved into decision-making within e-market supply chains, utilizing approaches like game theory and Stackelberg–Nash equilibrium. These studies, including those by Esmaeili et al., have focused on elements such as vendor-managed inventories, retailer information, pricing, and cooperation strategies between sellers and buyers, often in the context of competitive advertising [10, 7].

Pricing strategies are crucial in sustaining supply chains. Works by Cai et al. and others have explored pricing policies, particularly price discounting in dual-channel supply chains. Alongside pricing, maintaining an effective ordering policy is key to retaining customers. This aspect has been studied in the context of Business-to-Business markets, where factors like order quantities, price breaks, lot sizing, and discounts are significant decision variables. Additionally, supplier selection has been highlighted as a vital component for sustainable supply chain management [5, 17]. Despite the growing importance of digitalization and e-marketplaces, advertising remains a critical factor. Recent research has focused on cooperative advertising strategies in the manufacturer-retailer channel, where manufacturers may cover all or part of a retailer's advertising expenses. This collaboration aims to boost immediate demand in the supply chain. However, despite significant investments in advertising (e.g., \$15 billion in the USA in 2000), many companies still rely on heuristic methods like the rule of thumb or best guesses to determine their contribution rates, often without in-depth analysis of whether to contribute 50% or 100% [12, 8].

3. System model. The objective of this study is to evaluate the effectiveness of blockchain technology in enhancing the regulation of freshness-keeping activities in a fresh agricultural product supply chain, consisting of a supplier and an E-commerce platform (retailer). This research employs a comparative analysis design, examining both traditional and blockchain-based fresh agricultural product supply chains. The study focuses on the dynamic optimization of freshness-keeping effort, advertising effort, and the degree of blockchain adoption.

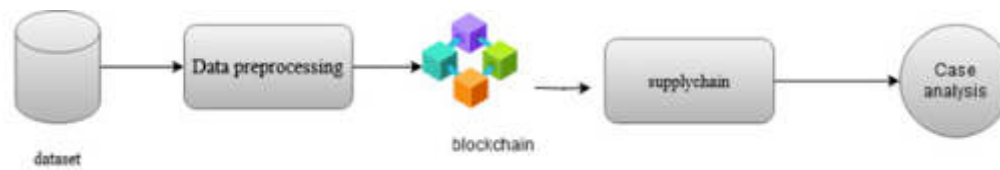


Fig. 3.1: Proposed Basic Model

3.1. Supply chain models. Traditional model represents the conventional operations of a fresh agricultural product supply chain. It focuses on the traditional interactions between the supplier and the retailer, without the integration of advanced technology like blockchain. Responsible for the production and initial handling of fresh agricultural products. Their primary roles include ensuring the initial quality (greenness) of the product and initiating the first phase of freshness-keeping efforts. Takes over the products for wholesale and sales. The retailer's responsibilities include managing the supply chain logistics from the point of receiving the products to delivering them to consumers. The retailer also decides on and implements advertising efforts to boost sales [16, 2].

In the traditional model, the continuity of freshness-keeping efforts by the supplier after the sale to the retailer is uncertain. This uncertainty often results in the degradation of product quality during transit and storage. Green investment refers to the initial quality assurance measures taken by the supplier. This investment is crucial but might not guarantee product freshness throughout the supply chain due to the lack of ongoing effort in maintaining freshness post-sale. Information Asymmetry is a key characteristic of the traditional model is the lack of transparency between the supplier and retailer, and subsequently, to the consumer. This asymmetry often leads to a gap in monitoring and ensuring product quality throughout the supply chain [6].

3.2. Blockchain Model. This model integrates blockchain technology into the supply chain, aiming to enhance transparency, traceability, and efficiency.

Blocks creation: When a user initiates a transaction, such as sending cryptocurrency, creating a record, or executing a smart contract, the transaction data is created. This data typically includes the sender's and receiver's information, the amount or nature of the transaction, and a timestamp. Before a transaction can be added to a block, it must be verified. This verification process depends on the blockchain's protocol. In a cryptocurrency context, this could involve checking that the sender has the necessary funds or rights to make the transaction.

Verified transactions are pooled together in a memory pool, also known as a mempool. Here, they wait to be picked up by a miner or validator to be included in a new block. Miners or validators (depending on the blockchain's consensus mechanism) select transactions from the mempool. They then use these transactions to create a new block. A block is essentially a data structure that packages a set of valid transactions along with other crucial information. A vital step in block creation is computing the block's hash. A hash is a fixed-length alphanumeric string derived from the block's data through a cryptographic hash function. Each block also contains the hash of the previous block, creating a linked chain of blocks. This linkage is what makes the blockchain secure and tamper-evident.

To add the block to the blockchain, miners (in a Proof of Work system) must solve a complex mathematical puzzle, which requires computational power. The first miner to solve the puzzle gets the right to add the new block to the blockchain. In a Proof of Stake system, validators are chosen to create new blocks based on various factors, including the amount of cryptocurrency they hold and are willing to "stake" as collateral. Once the mathematical puzzle is solved (in PoW) or a validator is chosen (in PoS), the new block is added to the blockchain. This addition is broadcast to all nodes in the network for verification. Once other nodes validate the block and its hash, the block becomes an official part of the blockchain. As an incentive, miners or validators receive a reward for their work in creating a new block. In the case of cryptocurrencies like Bitcoin, this reward comes in the form of newly minted coins and transaction fees.

3.2.1. Blockchain Integration. Blockchain technology provides a decentralized ledger that records every transaction or movement of the product in real-time. This feature allows all parties in the supply chain to track the product's journey and quality measures taken at each stage. The use of smart contracts in blockchain can automate certain processes, such as payments and compliance verification, based on pre-set conditions being met, like maintaining specific freshness levels.

Apart from the initial greenness investment, the supplier is encouraged to continue the freshness-keeping efforts even after the sale, as blockchain technology allows for the tracking and verification of these efforts. Retailer is responsible for determining the degree of blockchain adoption in the supply chain and managing the advertising efforts. The retailer can utilize the data from the blockchain to make informed decisions and enhance consumer trust.

Blockchain technology incentivizes the supplier to maintain freshness-keeping efforts throughout the supply chain, as these efforts are recorded and verifiable. The initial quality assurance measures are supported by the continuous tracking and maintenance of quality, ensuring that the greenness investment yields its intended benefits. The blockchain model aims to reduce costs associated with quality degradation and returns. The increased efficiency and reduced losses from spoiled goods can lead to improved profitability for both suppliers and retailers.

In summary, the traditional model is characterized by information asymmetry and potential lapses in freshness-keeping post-sale, while the blockchain model introduces transparency and traceability, incentivizing continuous quality maintenance and potentially transforming the efficiency of the supply chain.

3.2.2. Model Development. Traditional Model outlines the operations in the conventional supply chain, focusing on the roles and responsibilities of the supplier and retailer, and the dynamics of freshness-keeping and greenness investment. Blockchain Model incorporates blockchain technology into the supply chain model, examining its impact on transparency, freshness-keeping, and the overall efficiency of the supply chain.

Decision variables determine the degree of blockchain adoption and advertising efforts. Supplier decides on the greenness investment and freshness-keeping efforts. Analysis of how these efforts are sustained or abandoned in both supply chain models.

Blocks working principal. When a user initiates a transaction, such as sending cryptocurrency, creating a record, or executing a smart contract, the transaction data is created. This data typically includes the sender's and receiver's information, the amount or nature of the transaction, and a timestamp. Before a transaction can be added to a block, it must be verified. This verification process depends on the blockchain's protocol. In a cryptocurrency context, this could involve checking that the sender has the necessary funds or rights to make the transaction.

Verified transactions are pooled together in a memory pool, also known as a mempool. Here, they wait to be picked up by a miner or validator to be included in a new block. Miners or validators (depending on the blockchain's consensus mechanism) select transactions from the mempool. They then use these transactions to create a new block. A block is essentially a data structure that packages a set of valid transactions along with other crucial information. A vital step in block creation is computing the block's hash. A hash is a fixed-length alphanumeric string derived from the block's data through a cryptographic hash function. Each block also contains the hash of the previous block, creating a linked chain of blocks. This linkage is what makes the blockchain secure and tamper-evident.

To add the block to the blockchain, miners (in a Proof of Work system) must solve a complex mathematical puzzle, which requires computational power. The first miner to solve the puzzle gets the right to add the new block to the blockchain. In a Proof of Stake system, validators are chosen to create new blocks based on various factors, including the amount of cryptocurrency they hold and are willing to "stake" as collateral. Once the mathematical puzzle is solved (in PoW) or a validator is chosen (in PoS), the new block is added to the blockchain. This addition is broadcast to all nodes in the network for verification. Once other nodes validate the block and its hash, the block becomes an official part of the blockchain. As an incentive, miners or validators receive a reward for their work in creating a new block. In the case of cryptocurrencies like Bitcoin, this reward comes in the form of newly minted coins and transaction fees.

4. Simulation on data collection and analysis. Use of simulation software to model both traditional and blockchain-based supply chains, observing the behavior of the supplier and retailer under different scenarios.

Simulation of different levels of blockchain adoption and its impact on supply chain dynamics.

We analyse the solutions for following scenario,

1. Examination of real-world applications of blockchain in cold chain logistics for fresh agricultural products.
2. Comparative analysis of case studies where blockchain technology is employed versus traditional methods.
3. Use of optimization and econometric tools to analyze the dynamic optimization of decision variables.
4. Application of statistical methods to validate the effectiveness of blockchain technology in various settings.

4.1. Experimental Settings. Creation of different scenarios to analyze the effectiveness of blockchain technology in maintaining freshness. Identification of specific settings where blockchain is effective and where it is not suitable. Comparison of results between the traditional and blockchain-based supply chains are analysed. Assessment of the impact of blockchain adoption on freshness-keeping effort, advertising investment, goodwill, profit margin levels, and greenness investment decision.

4.1.1. Ethical Considerations.

1. Ensuring the confidentiality and privacy of data sourced from case studies and simulations.
2. Adherence to ethical standards in simulation modeling and data handling.

The significance of e-advertising and digital marketing in securing a competitive edge, especially in the agricultural product processing sector, cannot be overstated. The Supply Chain Management (SCM) model proposed in this study is framed as a complex non-linear maximization challenge, encompassing various variables and constraints. Initially, solutions to such constrained problems were sought by transforming them into unconstrained problems to find the global optimal solution. However, this method was found to be relatively inefficient and has since been replaced by techniques based on Inter point methods (IPM) equations. IPM equations are instrumental in providing conditions for optimizing constrained problems and offer solutions to numerous nonlinear challenges by directly calculation. These methods fall under the category of sequential quadratic programming (SQP) techniques.

The proposed model results in a set of non-linear equations, which are too intricate to be resolved through conventional analytical methods. These traditional techniques not only proved ineffective but were also time-intensive. Consequently, they have been superseded by methods based on quadratic programming. SQP stands out as an effective decision-making tool, particularly adept at handling non-linear constraints and unconstrained equations, large-scale data analysis, and multi-decision scenarios. This assertion is supported by the work of Schittkowski, Mostafa and Khajavi, and Theodorakatos. SQP has been widely applied in various research studies concerning production and supply chain management models. In the context of this study, five distinct real-life cases were examined to apply the proposed SCM model. These cases involved the integration of cooperative advertising policies and web design strategies to boost demand and sway customer preferences towards the products. Detailed explanations of these cases are provided below.

Case analysis. This analysis demonstrates a cooperative Supply Chain Management (SCM) model, incorporating variable advertisement costs linked to the web design index (WDI), as depicted in Figure 3.1. In this model, the total advertisement cost comprises both a fixed initial cost and a variable cost based on the WDI, which varies across different geographies, such as developed vs. developing countries, and urban vs. rural areas. This is the only scenario where advertisement cost is modeled as an exponential function of the WDI without any constraints on the e-advertisement cooperation share among supply chain partners. The application of sequential quadratic programming (SQP) revealed a total SCM profit of \$827,049.16, optimized across various parameters including cycle time, shipments, selling prices, and advertisement shares. Notably, the selling prices (\$454.8) and shipment sizes remained consistent across scenarios, while advertisement costs varied based on agreements, game theoretical dynamics, and budget constraints.

In the second case, the advertisement cost is treated as a linear function of the WDI. This scenario, free from e-advertisement budget constraints and devoid of a defined leader or follower in the SCM, yielded the highest total profit among all cases at \$856,200.9, with a cycle time of approximately 0.227 years.

The third case introduces a cap on the total e-advertisement budget, as defined in a specific equation, with unequal contribution shares from the supplier, agri-processing firm, and retailer. Here, the advertisement cost

Table 4.1: Performance analysis of proposed cases

S.No	Scenario	Freshness-Keeping Efforts	Greenness Investment	Advertising Efforts	Blockchain Adoption	Product Quality at Sale	Supply Chain Costs	Profit Margin	Consumer Satisfaction
1	Traditional Supply Chain	Decline after transfer	Initial only	Standard	None	Deteriorated	Higher due to spoilage	Variable	Lower
2	Partial Blockchain Implementation	Moderate increase	Sustained	Optimized	Partial	Improved	Reduced	Improved	Higher
3	Full Blockchain Implementation	Significant increase	Sustained	Highly optimized	Full	Significantly improved	Significantly reduced	Significantly improved	Highest

is again a linear function of the WDI. The total profit in this scenario was \$832,242.86, slightly higher than the first case but lower than the second.

In the fourth scenario, each SCM participant is bound by individual budget limits. Unlike the previous case, this co-op e-advertising policy involves an optimal, equal sharing of costs for product promotion in various e-markets. This resulted in a slight decrease in total profit to \$832,195.13 compared to the third case.

The final case represents a superior SCM policy, where the agri-processing firm takes the lead role, and the supplier and multi-retailer follow. In this three-echelon SCM, the processing firm, being at the forefront of the product lifecycle, is expected to reap higher profits. Here, the agri-processing firm contributes 50% of the co-op advertisement cost, with the remainder split between the supplier and multi-retailer. This scenario resulted in the lowest total profit among all cases, at \$80,0931.7.

Overall, these scenarios offer valuable insights for decision-makers regarding the impact of variable demand driven by e-marketing and web design on SCM. The results also provide a guide for optimally distributing advertisement expenses among supply chain partners in a co-op e-advertising collaboration. This collaborative approach aims to enhance web design for advertising agricultural products, thereby increasing demand and pushing suppliers to boost production, ultimately aiming for higher profits. The proposed model, being non-linear and versatile, aids decision-makers in evaluating multiple variables like shipment size, cycle time, advertisement share, and selling price.

4.2. Performance Measure. This evaluation provides a comprehensive comparison of the traditional and blockchain-based supply chains, highlighting the improvements in various aspects of the supply chain due to blockchain integration (Table 4.1).

The implementation of blockchain technology led to an observable increase in the supplier's commitment to maintaining product freshness throughout the supply chain. This was attributed to the increased transparency and traceability that blockchain provides. Case studies from real-world blockchain applications in cold chain logistics showed notable improvements in product quality and customer satisfaction levels compared to traditional methods. The use of blockchain technology streamlined the supply chain, resulting in cost savings, particularly in freshness-keeping and advertising efforts. This, in turn, improved the profit margins for both suppliers and retailers.

5. Conclusion. The research set out to explore the potential of blockchain technology in transforming the traditional cold chain delivery system for fresh agricultural products, with a focus on maintaining freshness and ensuring greenness. Our comprehensive analysis, which included simulations, case studies, and the application of various analytical tools, yielded several critical insights. Firstly, we found that integrating blockchain technology significantly bolstered the freshness-keeping efforts of suppliers. This improvement was primarily due to the enhanced transparency and accountability inherent in blockchain systems, which encouraged continuous quality

maintenance throughout the supply chain. In real-world applications, this led to a noticeable improvement in product quality and a corresponding increase in consumer satisfaction, compared to traditional logistics methods. Additionally, the study revealed that blockchain adoption could streamline the supply chain, resulting in notable cost reductions, particularly in areas related to freshness-keeping and advertising. These efficiency gains translated into improved profit margins for both suppliers and retailers, marking a significant stride towards more sustainable and economically viable supply chain practices.

Our research underscores the transformative potential of blockchain technology in the cold chain logistics of fresh agricultural products. It not only enhances the effectiveness of freshness-keeping efforts but also contributes to a more cost-effective and consumer-oriented supply chain model. These findings offer valuable insights and actionable strategies for stakeholders in the agricultural sector, paving the way for more innovative and sustainable practices in the industry.

REFERENCES

- [1] U. AKRAM, P. HUI, M. KALEEM KHAN, Y. TANVEER, K. MEHMOOD, AND W. AHMAD, *How website quality affects online impulse buying: Moderating effects of sales promotion and credit card use*, Asia Pacific Journal of Marketing and Logistics, 30 (2018), pp. 235–256.
- [2] B. BIGLIARDI AND C. GALANAKIS, *Innovation management and sustainability in the food industry: concepts and models*, in The interaction of food industry and environment, Elsevier, 2020, pp. 315–340.
- [3] S. BRAKEVILLE AND B. PEREPA, *Blockchain basics: Introduction to business ledgers*, Issued by IBM Corporation, (2016).
- [4] L. BUSCA AND L. BERTRANDIAS, *A framework for digital marketing research: investigating the four cultural eras of digital marketing*, Journal of Interactive Marketing, 49 (2020), pp. 1–19.
- [5] B. CAO, X. WANG, W. ZHANG, H. SONG, AND Z. LV, *A many-objective optimization model of industrial internet of things based on private blockchain*, IEEE Network, 34 (2020), pp. 78–83.
- [6] Y. CAO, L. TAO, K. WU, AND G. WAN, *Coordinating joint greening efforts in an agri-food supply chain with environmentally sensitive demand*, Journal of Cleaner Production, 277 (2020), p. 123883.
- [7] Y. CHANG, E. IAKOVOU, AND W. SHI, *Blockchain in global supply chains and cross border trade: a critical synthesis of the state-of-the-art, challenges and opportunities*, International Journal of Production Research, 58 (2020), pp. 2082–2099.
- [8] H. CHEN, A. CHEN, L. XU, H. XIE, H. QIAO, Q. LIN, AND K. CAI, *A deep learning cnn architecture applied in smart near-infrared analysis of water pollution for agricultural irrigation resources*, Agricultural Water Management, 240 (2020), p. 106303.
- [9] Y. CHEN, Y. LI, AND C. LI, *Electronic agriculture, blockchain and digital agricultural democratization: Origin, theory and application*, Journal of cleaner production, 268 (2020), p. 122071.
- [10] D. DUJAK AND D. SAJTER, *Blockchain applications in supply chain*, SMART supply network, (2019), pp. 21–46.
- [11] R. GATAUTIS AND E. VAICIUKYNAITE, *Website atmosphere: Towards revisited taxonomy of website elements.*, Economics & Management, 18 (2013).
- [12] X. HU, H.-Y. CHONG, AND X. WANG, *Sustainability perceptions of off-site manufacturing stakeholders in australia*, Journal of cleaner production, 227 (2019), pp. 346–354.
- [13] V. KUMAR, D. RAMACHANDRAN, AND B. KUMAR, *Influence of new-age technologies on marketing: A research agenda*, Journal of Business Research, 125 (2021), pp. 864–877.
- [14] L. C. LEONIDOU, *An analysis of the barriers hindering small business export development*, Journal of small business management, 42 (2004), pp. 279–302.
- [15] S. NAKAMOTO, *Bitcoin: A peer-to-peer electronic cash system*, Available at SSRN 3440802, (2008).
- [16] P. OBEROI, C. PATEL, AND C. HAON, *Technology sourcing for website personalization: A supply-and demand-side perspective*, in Celebrating Americas Pastimes: Baseball, Hot Dogs, Apple Pie and Marketing? Proceedings of the 2015 Academy of Marketing Science (AMS) Annual Conference, Springer, 2016, pp. 449–462.
- [17] J. SONG, Q. ZHONG, W. WANG, C. SU, Z. TAN, AND Y. LIU, *Fpdp: Flexible privacy-preserving data publishing scheme for smart agriculture*, IEEE Sensors Journal, 21 (2020), pp. 17430–17438.
- [18] J. D. WELLS, J. S. VALACICH, AND T. J. HESS, *What signal are you sending? how website quality influences perceptions of product quality and purchase intentions*, MIS quarterly, (2011), pp. 373–396.
- [19] J. XU, I. BENBASAT, AND R. T. CENFETELLI, *Integrating service quality with system and information quality: An empirical test in the e-service context*, MIS quarterly, (2013), pp. 777–794.
- [20] G. YIP AND A. DEMPSTER, *Using the internet to enhance global strategy*, European Management Journal, 23 (2005), pp. 1–13.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Dec 10, 2023

Accepted: Jan 4, 2024



NEXT-GENERATION CONNECTIVITY: A HOLISTIC REVIEW OF COOPERATIVE NOMA IN DYNAMIC VEHICULAR NETWORKS FOR INTELLIGENT TRANSPORTATION SYSTEMS

POTULA SRAVANI* AND IJJADA SREENIVASA RAO†

Abstract. Intelligent Transportation Systems are witnessing a paradigm shift with the integration of Cooperative Vehicular Networks. The transformations in Intelligent Transportation system in the realistic scenario has posed many research challenges to be addressed. This paper explores a profound survey on impact of vehicles' mobility within the context of real-time scenarios in Cooperative vehicular networks. The dynamic nature of vehicular mobility introduces unique challenges and opportunities for the design and implementation of cooperative systems. It delves into the key components that play a pivotal role for harnessing the full potential of Cooperative vehicular networks such as C-NOMA, Two-Way relaying, cluster formation and collaborative decision-making algorithms for improving latency, link reliability, cluster formation, and interference reduction etc. This paper outlines few surveys on each amalgamated technologies in Vehicular communication and conclude with the research problems in ITS due to vehicles mobility for real time scenarios.

Key words: Vehicular Communication; Cooperative Networking; Non-Orthogonal Multiple Access; Intelligent Transportation System; Cluster Head.

1. Introduction. Vehicular communications (VC) are critical components of Intelligent Transportation Systems (ITS), allowing vehicles to interact with one another as well as with infrastructure elements. This connectivity is critical for safety information, real-time data transmission, traffic management, and autonomous vehicle development. VC include a variety of technologies that improve road safety and efficiency, such as Dedicated Short-Range Communications (DSRC) and Cellular Vehicle-to-Everything (C-V2X) [1, 5, 102].

With the advent of the new technological growth in IOT device, Vehicle-to-everything (V2X) has become a trustworthy technology to drive multiple applications in ITS. V2X communication's low latency and dependability make it prevalent in life-threatening and delay-sensitive applications [54]. Long Term Evolution (LTE) based C-V2X proposed by 3GPP to provide public safety service. However, with resource sharing with cellular networks impose additional problems to fulfil low latency, high reliability, and large connectivity with severe data congestion [98, 19, 94]. Implementation of MIMO in V2X will leverage the benefits with diversity gain, spatial multiplexing and may reap to the growing demand of data rates to support multiple applications in vehicular communications. Complexity in receiver design, Inter-antenna Interference, instantaneous channel estimation and hostile Doppler shift due to mobility of the vehicles make MIMO implementation troublesome [95, 109]. This explores a new methodology in networking of LTE V2X communications to combat the technical challenges of MIMO and leverage the advantages of MIMO.

Vehicular channels provide unique issues, owing to signal fluctuations produced by mobility, which have a direct influence on performance in actual circumstances. New user applications, enhanced safety applications, and 5G data rates demand service needs such as broad coverage, low latency, throughput, and dependability. Cooperative Vehicular Networks (CVN) aims to improve network performance by incorporating vehicular node cooperation. However, due to high-speed mobility, competing parameters, and optimal node selection, efficient and adaptive cooperative algorithms are challenging. Current CVN literature focuses on network collaboration domains like resource scheduling, resource allocation, reliability, routing, and heterogeneous networking. Critical criteria for developing CVN include adaptive transmission power regulation, optimum relay selection, low synchronization overhead, adaptability, compatibility and impartial resource sharing. The future research

*ITAM Deemed to be University, Vishakapatnam, Andhra Pradesh, (sravaniphd2020@gmail.com)

†GITAM Deemed to be University, Vishakapatnam, A.P. (dr.ijjada2019@gmail.com)

Table 1.1: Nomenclature

Acronym	Meaning
AODV	Ad-hoc On-demand Distance Vector
AODV	Ad-hoc On-demand Distance Vector
AWGN	Additive white Gaussian noise
AF	Amplify and Forward
BS	Base Station
C-NOMA	Cooperative Non-orthogonal Multiple Access
CH	Cluster Head
CVN	Cooperative Vehicular Networks
C-V2X	Cellular Vehicle-to-everything
CoV	Cooperative Vehicles
CSI	Channel State Information
CRB	Cooperative relay broadcasting
CM	Cluster member
DSRC	Dedicated Short-Range Communications
DCVN	Heterogeneous cooperative vehicular networks
DMCNF	Distributed multi-hop clustering method
DF	Decode and Forward
DL	Downlink
FD	Full Duplex
3GPP	Third Generation Partnership project
GEC	Generous cooperative
HD	Half Duplex
ITS	Intelligent Transportation System
ISI	Inter Symbol Interference
IQI	In-phase/quadrature phase imbalance
IoT	Internet of Things
LTE	Long Term Evolution
MIMO	Multiple-Input-Multiple-Output
MAC	Medium Access Control
MA	Multiple Access
NOMA	Non-orthogonal multiple access
OMA	Orthogonal multiple access
OPA	Optimum Power Allocation
PDMA	Pattern Division Multiple Access
PA	Power Allocation
PAPR	Peak-to-average-power ratio
PSIC	Perfect Successive Interference Cancellation
RSU	Road Side Unit
RS	Relay Selection
SIC	Successive interference cancellation
SNR	Signal to Noise Ratio
SINR	Signal-to-Interference Noise Ratio
SC	Supervisory Coding
SCMA	Sparse Code Multiple Access
SM	Spatial Multiplexing
SER	Symbol Error Rate
SOP	System Outage Probability
TWR	Two-Way Relaying
UL	Uplink
VANETS	Vehicular Adhoc Networks
VC	Vehicular Communications
V2V	Vehicle-to-Vehicle
V2I	Vehicle-to-Infrastructure
VoI	Vehicle of Interest
TM	Throughput Maximization
OM	Outage Minimization
RE	Reliability Enhancement
UM	Utility Management
IM	Interference Minimization
PO	Power Optimization
SM	SNR Maximization

Table 2.1: List of survey articles emphasizing the benefits of cooperative networking.

S.No	Benefits	References
1	Upsurges the spectral efficiency and utilization of bandwidth	[30]
2	Delivers diversity gain	[62]
3	Mitigate the complications in employment of MIMO	[33]
4	Enhances throughput with QoS	[15]
5	Mitigates impairments like channel fading and path loss	[76]
6	Provides energy efficiency based on relay location	[103]
7	Reduces interference	[104]
8	Delivers communication reliability due to multiple paths	[12]
9	Minimizes implementation cost and improves coverage area	[87]
10	Decreases outage probability	[92]

should address challenges related to the impact of mobility, resource sharing, multi-functional protocol design, estimating dynamic channel behaviour and security.

This paper offers a valuable contribution by presenting a comprehensive review of earlier research works focused on cooperative vehicular networks (CVN). Through a meticulous examination of existing literature, the paper systematically synthesizes insights, methodologies, and findings from various studies in the field. By doing so, it provides a thorough understanding of the advancements made in cooperative vehicular networks and their applications. Furthermore, the paper sheds light on the critical research challenges associated with the implementation of CVN in realistic scenarios. By addressing these challenges, the paper aims to enhance the feasibility and effectiveness of cooperative vehicular networks in practical environments. Inclusively, this work aids as a significant means for researchers, and practitioners seeking to navigate the complexities of CVN research and implementation.

The rest of the paper is organized as follows: Section 2 briefs about cooperative communication and emphasizes on the requirements, challenges and effects of dynamic nature of vehicles in vehicular networks. Section 3 presents a comprehensive survey of the routing strategies in CVN. NOMA and its outperformance over OMA in Vehicular networks and NOMA implementation problems are projected in Section 4. Section 5 appraises about the impact of integrating cooperative networking and NOMA in vehicular communication and understanding the performance of this paradigm through few earlier works. Section 6 presents about the Two-way cooperative NOMA in Vehicular networks and a inclusive survey on HD/FD relaying. The impact of vehicle mobility over relay selection, cluster formation and selection of Cluster Head (CH) and critical research challenges in realizing the C-NOMA in Vehicular communication are depicted in Section 7. Section 8 concludes the paper.

2. Cooperative Communication. Cooperative communication in wireless systems is an important field of study that handles issues including fading channels, interference, and energy limits. It takes use of wireless device cooperation to increase dependability, throughput, network performance, and the capabilities of wireless communication technologies in a variety of applications. Using many relays as a virtual antenna array to create broadcast diversity is a promising strategy [48, 20]. As a result, the impact of channel fading might be mitigated, and wireless communications dependability could be increased. Many cooperative diversity techniques and protocols, such as cooperative protocols [54], single-relay cooperativeness, MIMO relay cooperativeness [48], optimal relay selection, two-way relaying (TWR) and network coding [39, 111], have been proposed and investigated. Given these characteristics, cooperative communication technology has the potential to effectively enhance the overall performance of vehicle networks. Cooperative vehicles (CoV) are a paradigm shift in ITS that can revolutionize road safety, traffic efficiency, and transport management by promoting collaboration, fading issues, path loss, shadowing, narrow coverage, and poor SNR [111, 3].

2.1. Requirements for Cooperative Vehicular Networks. CVN is unique in facilitating vehicular node interaction. To realise CVN, vehicle networks must meet numerous additional criteria due to the unique characteristic. These are some of the main needs.

2.1.1. Adaptive Transmission Power Control. V2V and V2I communication quality changes with time and space [25]. Vehicle speed additionally exacerbates the problem. Thus, CVN need adaptive transmission power control techniques. For dynamic run-time variable circumstances produced by moving vehicle cooperation, static gearbox procedures fail. Because vehicles cooperate, the adaptive gearbox protocol requires a learning mechanism to detect changes in surrounding vehicular environments. These adaptive transmission control techniques will greatly affect CVN.

2.1.2. Optimal Cooperative Relay Selection. CVN has received a lot of attention because to its potential to increase transmission and throughput dependability in highly dynamic wireless situations. In most CVN, relay nodes a pivotal role in delivering the packets towards the destination. However, transmission through multiple nodes to destination decrease the efficient resource utilization. The relay selection play a vital role to maximise network stability and throughput by efficiently using resources. Selecting the best cooperative relay node depends on vehicle direction, speed, traffic load, and channel quality.

2.1.3. Minimal Coordination Overhead. CVN nodes share their and neighbors' circumstances, which vehicle nodes use for relay, slot, resource, and forwarding decisions. This cooperation improves network performance by enabling collaboration between nodes. During information-sharing periods, nodes exchange messages to communicate channel conditions and gather topological information. To minimize duplicate transmission, an ideal relay node must be chosen among viable options. This minimizes coordination overhead and efficiently uses short-term resources.

2.1.4. Responsive Cooperative Transmission. CVN improves network speed and packet transmission reliability. However, cooperation techniques may influence neighbouring and collaborating relays. Cooperation with other relays requires to contemplate its self-transmissions besides resource restrictions. The participating node should also handle neighbouring node communications. The stages of node cooperation should be structured such that cooperative transmissions do not impair the performance of the collaborating node and its neighbours.

2.1.5. Equitable Resource Distribution. When considering transmission dependability, it is crucial to also prioritise fair resource allocation. Ensuring equitable allocation of resources to all participating nodes is crucial for optimising the performance of vehicular networks. Previous studies have investigated wireless network fairness in various aspects, such as bandwidth allocation [35], channel assignment [56], and power control [38]. Ensuring equal bandwidth and power usage for each node is of utmost importance, as fair resource distribution is vital. It is crucial to ensure fair distribution of resources to prevent resource hunger, which is why equitable resource allocation is necessary for the CVN.

2.2. Open Research Challenges. Research on cooperative vehicular communications focuses on developing robust protocols for dynamic situations and addressing mobility-induced performance deterioration, highlighting the need for compatibility with vehicle manufacturers.

2.2.1. High Speed Mobility. High-speed mobility in vehicular networks poses a significant challenge to the reliability of communications, introducing temporal variability that complicates traditional solutions. Despite efforts leveraging MIMO technology, cooperative relay and MIMO approaches still face hurdles in adapting to dynamic vehicular environments. The ever-changing network architecture further complicates relay node selection, necessitating agile techniques based on relative mobility speeds to address this challenge effectively. To tackle these obstacles, there is a pressing need for innovative relay selection methodologies capable of real-time adaptation to evolving network topologies. These techniques must navigate the complexities of high-speed mobility, considering factors like varying topography and rapid fading. By developing such adaptive protocols, we can enhance the dependability of vehicular communications, ensuring robust connectivity in dynamic automotive settings.

2.2.2. Multi-Objective Protocols. In the realm of Cooperative Vehicular Networks (CVNs), the predominant research paradigm tends to centre on singular parameters, often neglecting the inherent variability within vehicular network environments. Protocol design necessitates a comprehensive consideration of diverse

factors, encompassing aspects such as latency, throughput [83], fairness [58], and energy efficiency [80]. Nonetheless, crafting protocols that effectively reconcile these disparate objectives poses a formidable challenge, owing to the inherent conflicts and trade-offs inherent in such endeavours.

2.2.3. Estimating Channel State Information. Channel State Information (CSI) holds paramount importance in wireless systems, particularly for facilitating real-time cooperative networking. However, the real-time estimation of CSI presents a formidable challenge, primarily attributable to the dynamic nature of channel conditions and the rapid mobility characteristic of vehicular environments. Researchers may leverage principles from related domains to formulate estimation algorithms aimed at addressing this challenge [51].

2.2.4. Optimal Cooperative Relay Selection. The investigation delves into the complexities surrounding the selection of an optimal relay node to mitigate data collision and transmission redundancy effectively. Considerations include factors such as vehicle speed, direction of motion, channel quality, and traffic density. Findings from this study hold promise in informing the development of refined solutions for cooperative relay selection, offering valuable insights to propel further research. Moreover, the dynamic nature of automotive environments, marked by substantial vehicle movement, is demonstrably linked to adverse effects on network performance, underscoring the significance of these observations.

1. **Packet Loss Rate:** During congestion or fast changes in vehicle placements, the packet loss rate in highly mobile vehicular networks can be quite substantial, sometimes even exceeding 20%.
2. **Latency:** The latency in cooperative vehicular networks can vary significantly due to vehicle movement. In situations where there is heavy traffic or frequent lane changes, the latency can exceed 100 milliseconds, which can potentially compromise the real-time aspect of safety-critical systems.
3. **Fluctuations in Throughput:** The performance of vehicular communication systems can vary due to changes in the channel caused by vehicle mobility. During busy periods with a lot of movement on the roads, data speeds can decrease significantly, going from 1 Gbps to below 100 Mbps.
4. **Link Availability:** Link availability in vehicular networks can drop significantly in situations involving fast-moving vehicles, tunnels, or obscured line-of-sight circumstances, potentially reaching as low as 50%. This reduces the chances of establishing ongoing communication connections.
5. **Handover Frequency:** Due to the movement of vehicles in and out of communication ranges, frequent handovers are often required in cooperative vehicular networks. In busy urban areas, vehicles frequently change hands, causing signal overhead and disruptions.
6. **Network Density:** In intense traffic areas, network density may reach more than 1,000 vehicles per square kilometre. High density in communication channels may aggravate interference, contention, and collisions.
7. **Network Partition:** Occasionally, there may be instances where the network becomes temporarily divided due to sudden shifts in vehicle positions, resulting in certain groups of vehicles being isolated from the rest of the network. These partitions can cause significant disruptions to the flow of data, lasting for extended periods of time.
8. **Vehicle geographical Distribution:** Vehicles in cooperative networks have a non-uniform geographical distribution. Congestion in certain regions, such as junctions or highway on-ramps, may result in localised performance reduction.
9. **Effect on Safety Applications:** Mobility-induced performance loss might be especially worrisome in safety-critical cases such as collision evasion. In such cases, the success rate of timely warnings might fall below 90
10. **Impact on Traffic Management:** Due to the reliance on real-time data from vehicles, the effectiveness of traffic flow control and congestion management in cooperative traffic management could be hindered by concerns related to performance caused by mobility.
11. **Autonomous Vehicles Face Difficulties:** Autonomous vehicles rely significantly on cooperative communication. Vehicle mobility-induced performance deterioration might complicate autonomous decision-making and coordination, compromising safety and efficiency.

3. Routing Strategies. CVN routing systems must develop pathways that fully leverage the accessible forwarding relay selections in every hop in order to enhance transmission performance. Current cooperative ve-

hicular networking research initiatives attempt to achieve a variety of goals, including throughput maximisation, power allocation optimisation, transmission outage minimising, reliability enhancement, utilisation maximisation, and reservation slot collision minimization. Researchers are researching several techniques of constructing cross-layer routing to suit this need.

In [31], the authors put forward a proposal for VANET cross-layer routing that focuses on cooperative networking and finding a balance between end-to-end dependability and gearbox power consumption. The optimisation focuses on achieving two main objectives: ensuring high dependability while staying within gearbox power limits, and minimising power consumption. Nevertheless, the solution overlooks the possibility of co-channel interference from different source-destination pairs, which could potentially impact the performance of the protocol.

The VANET cross-layer routing technique optimizes wireless channel performance and overpowers unreliability [20]. Route detection and administration are done via AODV protocol. A relay selection method maximizes throughput, using predicted connection time and SNR. A MAC protocol extends route duration for stability. However, this assumes every vehicle is connected to RSU, increasing deployment costs.

By implementing cooperative forwarding and utilising network coding, the number of retransmissions can be significantly reduced [45]. The study in [111] introduced a cooperative forwarding system based on network coding, utilising a master/slave network topology paradigm. The master node selects the forwarding slave relay node based on trajectory, constancy, and proximity. The packet is encoded using linear network coding, incorporating slave addresses in route replies and updates.

The researchers in [69] proposed a network coding-aided scheduling technique for cooperative data dissemination systems. Vehicles exchange data via V2I and V2V channels, with each sending and receiving heartbeat messages to announce existence, update RSU, and switch operating modes. The proposed caching technique maximizes network coding effect, improving service performance but potentially cumulative latency for each packet. Authors in [72] proposed a bandwidth-optimized distribution technique for heterogeneous cooperative vehicular networks (DHVN), allowing quick data distribution and adapting to road design. The protocol improves packet retransmission and uses a store and forward technique to alleviate disconnections.

In [123], proposed an ungraceful cooperative strategy that uses forwarding probability with the aid of node position to decide next-hop transmission, reducing coordination overhead but requiring global positioning system information, potentially unavailable in tunnels. In [91] authors studied bidirectional cooperative V2V performance in vehicle abetted and RSU aided scenarios, using relay node location without channel status information. Authors in [37] investigate the influence of rate and gearbox range on CV critical systems, using a model to measure network performance. In [18] author also examined the combined impact of cooperation, interference, and channel fading in a Nakagami fading channel model.

A novel cooperative communication technique uses V2V, V2I, and mobility to increase vehicular network capacity. The disc model utilised in it omitted communication fading and interference, which may not suit the actual circumstance. This research synthetically analyses CVN network capacity using the highway scenario and variable communication scene fading. Using an analytical framework, a bottleneck expression of gearbox capacity is generated, and a newton iteration approach yields the estimated ideal cooperative vehicular ratio.

A dual segment generous cooperative (GEC) routing scheme is suggested [65]. A cooperative watchdog paradigm reduces false alerts and improves misbehaviour detection. The GEC routing protocol has several components that find cooperative pathways and distribute traffic. GEC design involves route finding and maintenance. The three steps of route discovery are neighbour sighting, erudition relay metric, and cooperative relay selection. Route recovery begins when a node gets a route error report. Link failure deletes the route from the routing database. The suggested technique separates troublesome vehicles, minimising end-to-end time. However, the suggested method lacks service diversity, which is essential for meeting traffic needs.

A novel network coding-aided scheduling technique is proposed to investigate the features of cooperative data dissemination systems [10]. In this setup, Vehicle-to-Infrastructure (V2I) communication channels facilitate the exchange of data between Roadside Units (RSUs) and passing vehicles, while Vehicle-to-Vehicle (V2V) channels enable vehicles to communicate cached data with nearby peers. The proposed solution comprises three phases: firstly, each vehicle transmits and receives heartbeat messages to announce its presence and gather information about nearby nodes. In the second phase, vehicles update their own and neighbouring

Table 3.1: List of review articles on routing strategies with multiple metrics

Research Articles	TM	OM	RE	UM	IM	PO	SM
Q. Zhang et. al. [120]	✓	×	×	×	×	×	×
M. Hempel et.al [127]	✓	×	×	×	×	×	×
Y. Cui et. al. [121]	✓	×	×	✓	×	×	×
W. Wang et.al [126]	✓	✓	×	×	×	×	×
T. Tang et. al. [128]	✓	×	×	×	×	×	×
H. Li et. al.[61]	×	×	×	×	×	×	×
C. Li et.al. [17]	×	×	×	✓	✓	×	×
A. Zafar et.al. [122]	✓	×	×	×	×	×	×
H. Yan et. al. [16]	×	✓	✓	×	×	×	✓
T. Zhang et. al.[45]	✓	×	×	×	×	×	×

vehicles status information to the RSU. Finally, in the last phase, all vehicles switch operating modes based on RSU scheduling. To optimize the network coding effect, a caching method is recommended. While the suggested network coding-assisted data distribution enhances service performance, it may also lead to increased hop-to-hop latency and packet latency.

To improve broadcast reliability, propose cooperative relay broadcasting (CRB) to rebroadcast neighbouring source node packets [10]. A two-state Markov chain-based optimisation framework and channel prediction technique are also presented. The optimisation framework limits CRB performance, while the channel prediction algorithm selects the optimal relay node. CRB facilitates proactive cooperative choices to send packets before expiration.

Node mobility and V2V/V2I communications have been studied to optimise throughput [17]. The authors suggested a V2I communications technique for the Vehicle-of-Interest (VoI) to obtain information from RSU while in coverage. After outside infrastructure transmission range, the VoI uses V2V communications to receive data via relay nodes. Data transmission under cooperative communication techniques is investigated using an analytical approach. In [16], the study proposed a cooperative communication method that maximizes throughput by utilizing V2V and V2I communication, mobility, infrastructure, and vehicle collaboration. It develops an analytical framework and close-form expression for feasible throughput.

4. NOMA in Vehicular Networks. Powered by the increasing spread of Internet-enabled smart devices and creative apps, sophisticated new services accelerate 5G network development needing new MA approaches. To reduce access collisions in V2X environments, novel multiple access techniques like SCMA, PDMA, and NOMA have been projected to enhance bandwidth efficiency and massive connectivity [84, 23].

Non-orthogonal multiple access (NOMA) is a trustworthy solution for 5G networks, offering increased spectrum efficiency, throughput and balanced user fairness [26], figure 1 depicting the downlink and uplink NOMA in cooperative vehicular networks. Unlike the standard orthogonal multiple access (OMA) system, the NOMA approach enables numerous users to share time/frequency radio resources while differentiating users based on power levels [32, 105]. Implemented in power domain and code domain, NOMA combines multiple users and uses channel gain differential for better performance. Successive interference cancellation (SIC) aids in signal discrimination [68].

NOMA, unlike OMA, offers fewer system delays, improved dependability, higher transmission rates, and lower-cost service needs [119]. Traditional OMA methods assign orthogonal radio resources to multiple users, but they don't always reach the sum-rate capacity of multiuser wireless networks. NOMA can fully utilize its capacity by surpassing OMA with power domain multiplexing at the transmitter and SIC at the receivers [97].

To demonstrate the mathematical link between NOMA and OMA, we use SNR expressions to characterise the two-user downlink performance. h_s and h_d depict the channel coefficients of V_S and V_D . At RSU $\rho|h_s|^2 < |h_d|^2$ represent transmit SNR, then throughput of OMA can be articulated for as [75].

$$R_{V_S}^{OMA} = \beta \log_2 \left(1 + \frac{\alpha_{V_S} \rho}{\beta} |h_s|^2 \right)$$

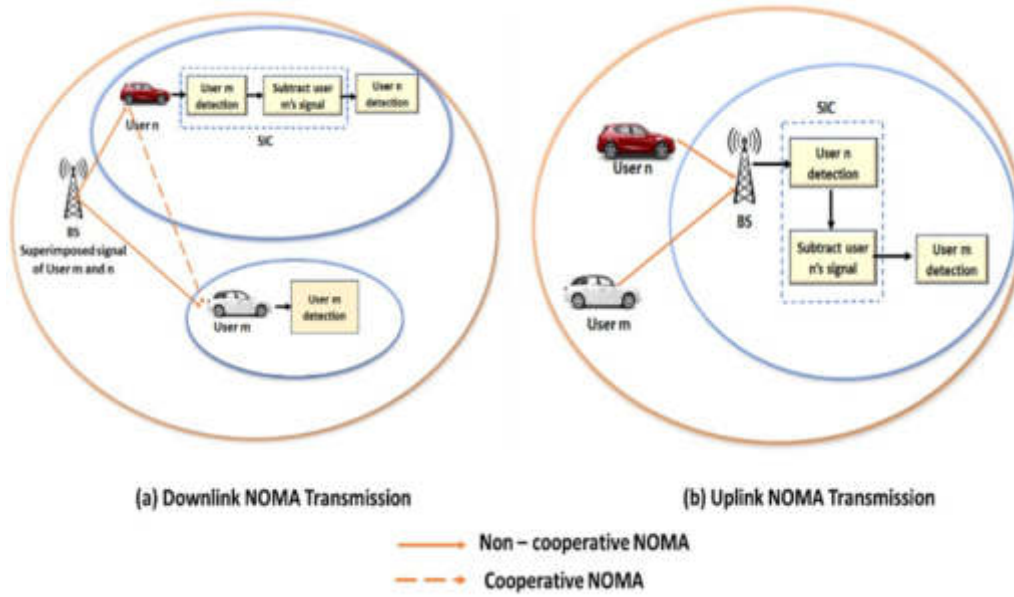


Fig. 4.1: Downlink and Uplink NOMA in CVN.

and

$$R_{V_D}^{OMA} = (1 - \beta) \log_2 \left(1 + \frac{\alpha_{V_D} \rho}{1 - \beta} |h_d|^2 \right)$$

where α_{V_S} and α_{V_D} are power allocation coefficients of with a condition of $\alpha_{V_S} + \alpha_{V_D} = 1$ and β is resource allocation coefficients. The throughput of V_S and V_D in NOMA are given as [75].

$$R_{V_S}^{OMA} = \log_2 \left(1 + \frac{\rho \alpha_{V_S} |h_S^2|}{1 + \rho \alpha_{V_D} |h_S^2|} \right)$$

and

$$R_{V_D}^{NOMA} = \log_2 \left(1 + \alpha_{V_D} \rho |h_d|^2 \right)$$

When we have $|h_d|^2 < |h_s|^2$, NOMA outperforms over OMA in sum throughput when adequate channel differences pertain between the two users.

4.1. Benefits of NOMA in Vehicular Networks.

Enhanced Spectrum Efficiency. Uplink NOMA can attain capacity constraints, but OMA methods are often suboptimal. However, when the quality of received signals of two vehicles are large, vehicle throughput fairness is meagre. The border of NOMA rate pairs in the downlink is outside the OMA rate zone. OMA can reach cumulative capacity in multi-path fading channels, but NOMA is optimum [49].

Massive Connectivity. By utilising non-orthogonal resource allocation, NOMA enables the support of a larger number of users or vehicles compared to OMA. This allows for excellent performance even in overloaded scenarios, despite the constraints of existing resources and scheduling limitations [49].

Low transmission latency and signalling rate. Traditional OMA with QOS requirements entails scheduling requests to base stations, leading to significant delay and expensive signalling costs. This is particularly problematic for large connections in 5G. However, certain NOMA uplink methods do not require dynamic scheduling, resulting in grant-free transmission, which can significantly reduce latency and signalling costs.

Table 4.1: List of survey articles on NOMA with various performance metrics

Research Articles	US	PA	FA	CN	MN	UN	SDN	EPN	HN
S.Han et. al [26]	✓	×	×	×	✓	✓	✓	×	×
F.Cui et. al [13]	✓	✓	✓	✓	✓	✓	×	×	✓
J.Choi et. al. [32]	×	×	✓	✓	✓	✓	×	×	×
E.Hossain et. al. [4]	✓	✓	×	×	×	×	×	×	✓
S.Kwak et.al. [49]	✓	✓	✓	✓	✓	✓	×	✓	✓
D.I.Kimetal et.al. [99]	✓	✓	×	✓	✓	✓	×	✓	×
Z.Ding et. al. [93]	✓	✓	×	✓	×	×	×	×	×

Fairness. NOMA empowers individuals with limited abilities. It is possible to achieve a desirable equilibrium between user fairness and performance. In this study, we will delve into the intricate NOMA fairness methods, such as intelligent power allocation (PA) policies [92, 73] and the cooperative NOMA scheme [100].

4.1.1. Ultra-high Connectivity:. The 5G system will connect billions of IoT smart devices [70], and NOMA, with its non-orthogonal properties, provides an efficient design option for conventional OMA.

4.1.2. Compatibility:. NOMA leverages the power-domain as "add-on" strategy for any current OMA technologies like TDMA/FDMA/CDMA/OFDMA. Given the maturity of Superposition coding and Successive Interference Cancellation procedures practically NOMA might be combined with aforementioned multiple Access techniques.

Non-Orthogonal Multiple Access (NOMA) technology represents a versatile 5G technique renowned for enhancing spectral efficiency and reducing latency in wireless communication systems [57]. In the realm of Vehicle-to-Everything (V2X) services, NOMA is harnessed to mitigate resource collision and achieve high throughput transmission even under resource constraints. Furthermore, NOMA finds application in vehicular networks to minimize latency and bolster reliability [29]. Leveraging NOMA-based broadcasting introduces a hybrid architecture alongside power control mechanisms tailored for participating vehicles [28]. Additionally, NOMA-spatial modulation (NOMA-SM) augments bandwidth efficiency and alleviates wireless V2V scenarios [6]. By employing power allocation algorithms with opportunistic constraints, NOMA systems hold the potential to enhance V2X network performance [49].

4.2. Implementation Challenges of NOMA. Nonetheless, some outstanding concerns must be solved before NOMA may be used in vehicle contexts.

- **Error Propagation in SIC:** In NOMA systems, SIC is the primary mechanism for detecting users. However, one major disadvantage of adopting SIC is the inter-user error propagation problem, which spreads from one user to the next since a judgement mistake results in deducting the incorrect remodulated signal from the composite multiuser signal, resulting in residual interference. The majority of extant NOMA research contributions are predicated on the premise that the SIC receivers are capable of completely cancelling the interference. In reality, because to faulty PA and inadequate channel decoding, this assumption cannot be easily met in practise. Several academics have acknowledged the obscurity of error propagation concerns and explored the impact of faulty SIC on uplink NOMA.
- **NOMA Channel Estimation Error and Complexity:** Channel estimation plays a pivotal role in NOMA systems compared to OMA, inaccurate channel estimates lead to muddled user collation and poor power control, both of which impact the accuracy of SIC decoding. The channel estimate diverges with numerous parameters and is always a critical challenge to achieve perfect estimation in practical scenarios.

In addition to the aforementioned implementation challenges, NOMA encounters several additional limitations. Due to NOMA's utilization of multiple access at variable power levels, the received signal intensities exhibit variability, presenting additional hurdles for effective analog-to-digital (A/D) conversion. While robust signals necessitate a wide voltage range, ensuring correct quantization at low levels demands high-resolution ADCs for weaker signals. However, employing ADCs with both attributes proves impractical due to cost and

system complexity constraints, inevitably leading to quantization errors. Balancing performance and complexity becomes paramount, necessitating a suitable trade-off consideration.

Another significant challenge in NOMA, often overlooked in prior research, is accurate synchronization. Achieving complete synchronous transmissions proves unattainable in practical scenarios due to the dynamic mobile environments of users, especially evident in uplink NOMA transmission. Addressing synchronization difficulties can be approached through two strategies: proposing precise pilot designs to minimize time synchronization errors and exploring novel asynchronous communication techniques. In a recent study [40], researchers proposed an innovative integrated circuit (IC) approach for asynchronous NOMA-aided orthogonal frequency multiplexing systems, highlighting the significant impact of relative time offsets among interfering users on system performance.

Moreover, NOMA encompasses various additional features such as reference signal design, channel estimation, and Channel State Information (CSI) feedback mechanisms, which bolster performance even in the presence of substantial cross-user interference. Resource allocation signaling is adept at accommodating diverse NOMA transmission modes, while extending NOMA to massive Multiple-Input Multiple-Output (MIMO) systems and other MIMO configurations promises optimal performance. Furthermore, efforts are underway to mitigate the peak-to-average-power ratio (PAPR) in networks with multiple vehicular networks, underscoring NOMA's ongoing evolution and its potential to address a multitude of challenges in wireless communication systems.

5. Cooperative NOMA in Vehicular Networks. NOMA-based transmission has less coverage than OMA-based transmission since each NOMA user is only assigned a portion of total transmit power. One successful strategy to broaden the coverage of NOMA-based transmission is to include cooperative techniques into NOMA networks, resulting in cooperative NOMA networks [67].

Cooperative communication has been a significant focus of study in the past due to its potential to enhance network coverage, throughput, and transmission reliability. By leveraging spatial diversity gain to counteract the effects of wireless fading, cooperative communication offers promising benefits [104].

The fundamental idea behind cooperative communication is to incorporate additional nodes that can assist in facilitating communication between the source and destination. By leveraging its geographical diversity advantage, combining data from various sources enhances the reliability of the destination's reception.

5.1. C-NOMA Networks with Relay and User Assistance. C-NOMA network research may be split into two groups with the aid of nature of collaboration. As illustrated in Figure 5.1a, is based on dedicated relay cooperation, in which specialised relays are installed to aid communication between the source and NOMA consumers [79].

The Figure 5.1b collaboration involves NOMA users with robust connections acting as relays to help those with deprived connections. According to the NOMA principle, strong users must decode frail users' information before decoding their anticipated information [79]. When they correctly discover weak users' information, they can assist them.

Integrating cooperativeness with NOMA can enhance system efficiency and dependability. Users with superior channel conditions may decode messages for others using the C-NOMA technique, which exploits prior knowledge in NOMA systems [33]. When users have superior channel conditions, short-range communication technologies such as Bluetooth and ultra-wideband may set up cooperative conversations. There are two parts to C-NOMA: transmission and collaboration, with NOMA users receiving superposed messages

$(N - 1)$ time slots make up the cooperation phase. In the i^{th} time slot, where $1 \leq i \leq (N - 1)$, the user from $(N - i + 1)^{th}$ broadcasts the combination of the messages from $(N - 1)$ C-NOMA achieves the greatest variety gain for all users by adjusting power allocation factors based on local channel circumstances. However, it is costly due to serial message retransmissions. C-NOMA provides user coupling with the aid of separate channel gains, reducing system complexity. Optimum power allocation strategies further improve the performance [42, 60, 73].

C-NOMA transmissions offer several key advantages over conventional NOMA transmissions, including:

- **Low System Redundance:** When dealing with weak signals, the DF protocol makes intellect since SIC algorithms in NOMA can decipher user messages. It is possible to remodulate and retransmit them from locations that are closer to the intended recipients.

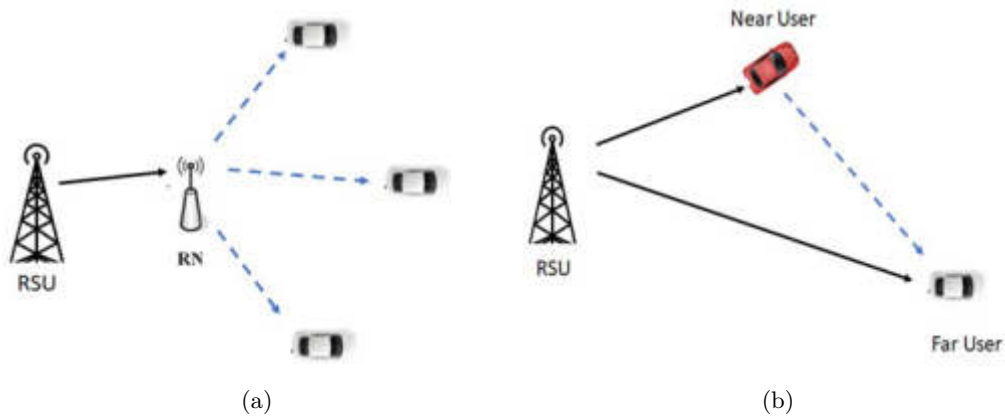


Fig. 5.1: Cooperative NOMA Networks (a) Relay-aided C-NOMA Network, and (b) User-aided C-NOMA Networks

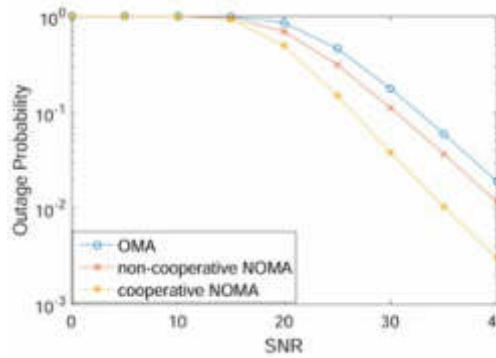


Fig. 5.2: Performance comparison between cooperative NOMA and non-cooperative NOMA.

- Greater fairness: C-NOMA improves the reliability of weak users, increasing fairness in transmission, especially when the weak user is near the cell’s edge with reference to the BS [100].
- Higher diversity gain: This C-NOMA serves as a linchpin for superior performance in dynamic and challenging vehicular communication environments. Through NOMA, multiple users can simultaneously transmit and receive data within the same time-frequency resource, fostering a cooperative environment that significantly enhances diversity gain. This heightened diversity gain translates into improved reliability and robustness, crucial elements for vehicular networks where communication channels are inherently volatile and prone to fluctuations. By leveraging the capabilities of NOMA to efficiently manage multiple connections, vehicular networks can achieve unparalleled diversity, enabling seamless communication even in scenarios with fading channels or challenging propagation conditions [70].

Figure 5.2, shows cooperative NOMA’s superior outage probability and diversity increase compared to non-cooperative NOMA and OMA. This strategy enhances transmission consistency, especially for weak NOMA users with deprived channel conditions.

5.2. Survey on C-NOMA:. NOMA achieves the highest diversity order for all users compared to OMA [33]. Many researchers have worked on designing and implementing NOMA techniques and on solving various technological problems associated with those methods. The literature demonstrates that NOMA is compatible with cooperative communication.

In terms of diversity order, NOMA outperforms OMA for all users [33]. Researchers have put lot of effort to develop and apply NOMA approaches, as well as to solve the many technical issues. Research shows that NOMA can work with cooperative communication. A special eminence on the C-NOMA techniques is presented in this treatise.

With NOMA in coordinated direct and relay transmission, the authors of [57] examined the ergodic capacity and outage probability of the system. In a multi-relay scenario, they explored consequences of how relay selection affected system performance and suggested a two-stage max-min method for selecting relays. Presenting a power allocation method and investigating the usage of SIC in decoding user signals, Yang & et al. [113] investigated the outage probability and user rates of a NOMA system with paired users in a non-cooperative uplink scenario.

The study in [114] explored the influence of relay selection on C-NOMA performance. A two-stage RS technique was used to minimize outage probability and maximize diversity. A dual-hop cooperative relaying technique based on NOMA was proposed in [52], involving simultaneous interaction between two sources over the same frequency range. The protocol successfully achieved ergodic total capacity through perfect and imperfect consecutive interference cancellation.

The work presented in [107] investigated the performance of a downlink cooperative relay system using DF and AF protocols across Nakagami-m fading channels. Data decoding order from cell-edge users was determined in the research using statistical CSI. When considering ergodic total rate, the findings reveal that, even when considering near-far effects, the DF protocol performs better than the AF protocol. However, the advantage diminishes with increasing SNR. The study also considers the obsolete CSI effect due to continuous channel fluctuations.

In [115], authors explored two possibilities in their paper. a) Direct connection between BS and Users b) No direct connection. First they investigated ordered users' outage behavior utilizing the AF relaying protocol while there was direct connection between the Base station (BS) and them. Secondly, a new closed-form calculation for the downlink outage probability with stochastically dispersed users was created in the absence of a direct relaying node. The users' diversity orders for the two situations have been determined based on the analysis findings. Furthermore, adopting the NOMA technique rather than the standard numerous Access approach ensures the fairness of numerous users. The suggested system was assessed using just AF, and two-way communication with HD/FD may have enhanced user throughput even further.

In order to improve upon OMA-based methods with regard to outage probability, system throughput, and spectrum utilization, reference [71] suggested a NOMA-based transmission strategy for cooperative spectrum-sharing networks. Under the assumption of a high SNR for the purpose of calculating the outage probability, the research in [74] investigated NOMA-based downlink AF relay networks that had low CSI performance. However, because to fixed-ordered decoding approaches, optimal uplink performance could not be confirmed. In [86], researchers examined NOMA-based single-and multi-vehicle systems fared under multipath fading conditions with and without in-phase/quadrature phase imbalance (IQI). Results showed that IQI effects vary across NOMA users and depend on system characteristics. Higher order users were more susceptible to IQI. The TBS-C-NOMA network enhances data reliability in traditional C-NOMA networks by preventing error propagation [53]. If the SINR is superior than the threshold, the intra-cell user will convey the symbols of the cell-edge user. The optimal threshold value is examined to reduce BEP, with SINR determining relay selection. The authors of [7] introduced a two-way cooperative relay technique that utilizes NOMA for bidirectional communication, outperforming a standard one-way C-NOMA system in terms of outage probability and ergodic rate.

Discussions and Prognosis: In the realm of cooperative vehicular communications, the synergy between Non-Orthogonal Multiple Access (NOMA) and cooperative techniques holds paramount significance for scientific advancements, particularly in scenarios where users are strategically positioned and route loss remains consistent. The interplay between NOMA and cooperative communications serves as a cornerstone for driving scientific contributions forward, offering opportunities to enhance system performance and reliability. While previous research endeavors have delved into the potential performance gains achievable through cooperative approaches, numerous open research challenges persist, warranting further investigation and innovation.

One such challenge lies in leveraging relays to improve reception reliability within NOMA-based networks. Relays introduce an additional time window for signal transmission, potentially bolstering the dependability of reception. However, the effective integration of relays necessitates careful consideration of various factors,

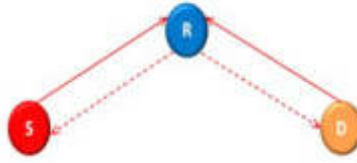


Fig. 6.1: Two-Way Cooperative communication.

Table 6.1: List of Survey articles with two way cooperative relaying

References	No. of users	No. of active relays	Link direction
[100, 112, 59, 85]	1	1	One-way
[119, 43, 110, 101]	2	1	Two-way
[44, 108]	1	2	One-way
[55, 88]	Multiple	1	Two-way

including interference mitigation and resource allocation. In this context, the deployment of full-duplex relays emerges as a promising avenue to address these challenges. By enabling simultaneous transmission and reception, full-duplex relays have the potential to mitigate the need for additional interference management techniques, thereby simplifying network design and enhancing overall system efficiency.

However, the successful implementation of full-duplex relays in NOMA networks requires comprehensive solutions to eradicate interference effectively. Overcoming interference challenges remains a critical research objective, as it directly impacts the reliability and performance of cooperative vehicular communications. Future research endeavors should focus on developing robust interference mitigation strategies tailored to the unique characteristics of NOMA-based cooperative networks. By addressing these challenges, the integration of NOMA and cooperative communications can unlock new opportunities for advancing the reliability, efficiency, and scalability of vehicular communication systems.

6. Two-way Cooperative Networks. The cooperative NOMA treaties use a one-way relay system, while two-way relay (TWR) technology improves spectral efficiency [81]. TWR systems use relays to communicate between nodes, resulting in larger throughput. CNOMA uses full-duplex mode and combines TWR with a suitable technique to improve system spectral efficiency. Two users working in three phases benefit from bidirectional communication in [116] s two-way cooperative relay technique using NOMA, demonstrating superior outage probability and ergodic rate than a one-way C-NOMA system. Hybrid TWRS was established in [8] for compressing data and high spectral efficiency in Network Coding.

Figure 6.1 depicts a system which involves two source nodes broadcasting packets x_1 and x_2 to the relay R , in second phase, R decodes them based on channel gains, and combining them using an XOR operation ($x = x_1 \oplus x_2$). The source nodes decode by performing $x_1 \oplus x = x_2$ and $x_2 \oplus x = x_1$ respectively. This efficiency has led to numerous researches focusing on this system as a promising solution to broaden coverage and deploy the diversity characteristic of wireless networks. The system's efficiency makes it a promising solution for wireless network applications.

Combining TWR with NOMA (TWR-NOMA) can enhance spectrum efficiency and system throughput [8]. Both technologies outperform in boosting spectral efficiency. Analytical and simulation findings show that outage probability converges to an error floor, even with Perfect Successive Interference Cancellation (PSIC), resulting in a zero diversity order, making it logical to combine TWR and NOMA [116].

6.1. Survey on TWR Cooperative NOMA. The majority of C-NOMA treaties concentrated on one directional communication, where messages are sent from the either source to relay and destination or destination to relay and source. Spectrum utilization is limited since communication requires two time slots to reach its target. Optimizing spectral efficiency is achieved by the use of TWR technology [88].

The authors of [47] offered a new method for selecting relays to streamline TWR systems after studying the behaviors of DF relay outages under both ideal and imperfect CSI conditions. The outage behavior of two-way full-duplex DF relay systems on different multi-user scheduling strategies was investigated using CSI and system status information [63]. In [117], evaluated the performance gain and provided an expressions for outage probability, ergodic capacity, and throughput to show the value of with and without direct connection in cooperative NOMA. The work did not specify relay choices, although dedicated relay was explored. Also absent: channel variation effects.

The performance of the C-NOMA system in full-duplex (FD) and amplify-and-forward (AF) modes with a dedicated relay was investigated in a study published in [2]. When compared to the FD decode-and-forward (DF) NOMA methodology, the suggested FD-AF relaying method outperformed in partial self-interference cancellation, according to the simulations. An increase in transmit signal-to-noise ratio (SNR) raises the near user's ergodic achievable rate.

Based on HD/FD in [118], authors developed closed-form methods to calculate the outage probability for two NOMA Relay selection (NOMA-RS) systems. An outage was shown to be less likely in HD-based NOMA-RS with more relays. The researchers postulated that HD-based NOMA-RS systems may provide a diversity order that is precisely proportionate to the number of relays. Nonetheless, NOMA-RS systems based on FD outperformed NOMA-RS systems based on HD in the low SNR range. The investigation did not take mobility, relaying systems, or delay limits into account. A two-way relay NOMA with DF was examined in the study, and for users with imperfect or perfect SIC, the researchers discovered closed-form equations for the exact and asymptotic outage probability in [42]. When dealing with channel circumstances that change over time and impact diversity gain, we ignore the relay selection.

The study in [34] introduced a novel approach for two-user DL and UL transmissions, utilizing a dedicated relay node to assist HD. This approach was compared to traditional one-way relay-based C-NOMA and OMA strategies. Nevertheless, the existence of a dedicated relay remains undefined. In [113], a virtual full-duplex cooperative NOMA scheme and relay selection method were proposed for a downlink two-hop network with multiple HD DF relays. This scheme provides a greater sum-rate compared to regular OMA transmissions and has the potential to improve spectral efficiency when used in conjunction with TWR.

A C-NOMA system for a two-user TWR network (TWRN) was presented in [67, 43], demonstrating a higher sum-rate compared to traditional OMA-based transmission. Unfortunately, no performance evaluation was articulated. A study conducted by experts in the field explored TWRNs that utilize NOMA technology. The study specifically focused on the aspects of secrecy and FD C-NOMA aided TWR systems. Formulas were derived to calculate outage probability, diversity orders, ergodic rates, and system throughput. In certain situations, FD NOMA gearbox demonstrated superior performance compared to HD gearbox.

Numerous multiuser relay networks (UMRNs) have effectively integrated NOMA, including cellular uplink and downlink broadcasts [21, 66, 90] and multi-pair TWRNs. Using rate excruciating and consecutive group decoding, a NOMA multipair TWRN was examined in research that was carried out in [125]. Despite this, a performance analysis was omitted from the research, which found significant decoding difficulty.

7. Relay Selection Strategies in CVN. The issues of routing in automotive networks are far from apparent. The challenge stems mostly from the instability of routing pathways generated by node mobility and network fragmentation. Indeed, the fact that the network has intermittent or partial connection suggests that routing management must vary from topological techniques.

The major routing algorithms for VANETs may be implemented in a diverse context, each with its particular set of characteristics like speed, vehicle density, road layouts, and so on. For example, metropolitan settings are distinguished by a complicated mobility model, high vehicle density, and limited speed, all of which are mostly the result of existing junctions and stop spots. However, the highway and rural settings are distinguished by great distances, making these surroundings less disruptive for radio waves during intervehicle interactions.

Inter-vehicular routing remains a significant difficulty, particularly in urban contexts with high vehicle density and the existence of impediments. The proposed new routing solution must fulfil the needs and characteristics of this kind of environment, whose limits have a significant impact on node mobility and routing performance. As a result of the high mobility of vehicles, the routing route between vehicles may not be guaranteed at all times. Unfortunately, the majority of present protocols fail to adequately assess the im-

part of potential obstacles that may have an immediate bearing on routing efficiency and, therefore, vehicle communication.

In [62, 124] authors presented a method called Optimal Power Allocation (OPA) for the all-participate-amplify-and-forward (AP-AF) environment. The aim of this method is to minimize the outage probability for multiple relay nodes. The OPA technique was shown to effectively reduce symbol error rate (SER), according to the author's findings. In larger networks, the performance of the AP-AF technique tends to decrease as the number of collaborating nodes increases. This study presents a new approach called selection-based amplify and forward (S-AF), which aims to minimize overhead by choosing the most suitable relay node. This approach combines relay selection and OPA. Unlike [11], this single relay selection does not include relay delay.

The Ad-hoc On-demand Distance Vector (AODV) protocol is used for route discovery and management, with a novel relay selection method aiming to maximize throughput. The cost-based selection criteria consider projected link time and SNR. The relay node decodes frames and sends them in their allocated slot, discarding the rest. A MAC protocol is developed to increase route duration and stability. However, the study assumes every vehicle is directly connected to a Remote Sensing Unit (RSU), resulting in high deployment costs. The study in [31] proposed VANET cross-layer routing via cooperative transmission and a novel route selection method to balance end-to-end dependability with gearbox power consumption. Two optimization issues are created to maximize dependability within transmission power limits and minimize power usage under reliability limitations. The solution assumes one network route and ignores co-channel interference from other node pairs, potentially affecting protocol performance if multiple network paths are active.

The authors in [64] presented an uncoordinated cooperative strategy that uses forwarding likelihood based on node position to decide on the next-hop transmission, reducing coordination overhead but requiring global positioning system information. [38] also presents a cooperative relay broadcasting (CRB) strategy to improve broadcast transmission reliability. The strategy includes a channel prediction technique and an optimization framework based on a two-state Markov chain. The optimization framework constrains CRB performance, while the channel prediction technique aids in selecting the optimal relay node.

The combined impact of user collaboration and dedicated relay cooperation is investigated in [27] for diversity benefit. Adaptive relay selection techniques have the lowest system outage probability (SOP). The relay selection was adaptable based on whether or not the nearby user need assistance. The investigation did not incorporate mobility or channel fluctuations in relay selection, and diversity was calculated based on the number of dedicated relays.

7.1. Mobility effect in clustering of vehicular networks. Clustering is a technique that involves gathering nodes with similar properties, such as destination, direction, and speed, to form distinct virtual sets called clusters. Vehicle mobility in vehicular networks significantly influences cluster formation and maintenance. Each cluster has a cluster head (CH) and several cluster members (CMs), with CH selection influenced by factors like the node's relative average speed. Each cluster has a predetermined size determined by the node's transmission range. Vehicle node clustering can improve communication efficiency in Vehicle Access Control Networks (VANETs) if the clusters are trustworthy and long-lasting.

The mobility effect in clustering may be difficult to achieve for the following reasons:

1. **Dynamic Topology:** The fast movement of vehicles might generate frequent changes in network topology, resulting in cluster reformation.
2. **Cluster Disruptions:** Vehicle movement may cause clusters to split apart, making steady and effective communication inside a cluster problematic.
3. **Load Imbalance:** Vehicle movement might result in an unequal distribution of network load, resulting in performance deterioration in certain portions of the network.
4. **Cluster Formation:** Vehicle movement might make it difficult to create clusters efficiently and effectively, resulting in inferior performance.

In order to address these challenges, experts are exploring various approaches aimed at enhancing the stability, efficiency, and scalability of vehicular networks. These efforts involve considering vehicle movement when establishing and managing clusters. The VANET distributed multi-hop clustering method (DMCNF) uses location, speed, position, and direction as input metrics, but one hop neighbor selection increases network time. To address this issue, a new protocol Enhanced DMCNF is being researched in [27], which handles

communication between the RSU and stable cluster, reducing cluster overhead.

Rapidly moving vehicles can cause network architecture changes and communication path instability, leading to connection failure [41, 106]. A method utilizing a multi agent system and particle swarm optimization is employed to tackle this problem. This approach utilizes the particle swarm optimization algorithm, a cluster formation process, and a multiple agent-based technique. The input parameters encompass various factors such as simulation time, transmission rate, coverage area, transmission range, node density, and numeral iterations. On the other hand, the output metrics focus on evaluating the throughput, packet loss, routing overhead, and packet delivery ratio. This approach is suitable for networks with average service quality, but performance may suffer when applied to high-quality networks.

The authors in [46] proposes a hybrid dynamic cluster strategy to improve communication reliability in vehicular adhoc networks. This strategy involves creating a stable dynamic topology using agent technology, with key measures including time for cluster creation, cluster head selection, and the total lifespan of the cluster. This approach addresses communication failure as a significant disadvantage. In [82] authors proposed a new strategy for enhancing vehicle network stability using clustering mechanisms, which are influenced by constantly changing topologies. The method uses indicators like received signal strength and identification number metrics. For specific roadside settings, a dynamic mobility and stability-based clustering technique is designed, focusing on vehicle direction, location, and lifespan estimation. The proposed work involves defining clusters, transitioning between states, creating and selecting heads, selecting gateway nodes, and ensuring maintenance. The input parameters consist of various factors such as the vehicle density, road length, hastening rate, slowing rate, proliferation model, numeral reiterations, and mobility model. On the other hand, the output metrics provide insights on clusters, CH length, state variation and clustering competence.

In [22, 24] advocated distributed multi-hop clustering. It passively picks the cluster head after organizing the vehicle using vehicle following. The vehicle following technique reduces cluster formation costs, while passive clustering improves stability. However, it ignores inter-node connection dependability when choosing the next vehicle, resulting in poor cluster reliability. Cluster coverage is improved and VANET cluster heads are reduced when a multi-hop clustering technique is used as opposed to a single-hop clustering method, according to this study. The use of available bandwidth is subsequently enhanced.

The authors in [78] proposed a strategy to enhance vehicle network stability using clustering mechanisms, which are influenced by constantly changing topologies. The method uses indicators like received signal strength and identification number metrics. With a focus on dynamic mobility and stability, a clustering approach is created for certain roadside situations that considers the vehicle's trajectory, location, and predicted lifespan [50]. By using the Ant Colony Optimization algorithm, the best way to transmit data from source to destination may be determined. Cluster heads are built with the highest possible link stability. The assessment metrics under consideration include QoS, network dependability, latency, energy efficiency, and network throughput.

In [96, 36], Trust degrees are utilized in authenticated VANET clustering to select CHs, with direct trust degrees reported by neighbors based on prior experience and indirect trust degrees based on nearby node recommendations. This approach addresses VANET instability caused by fast movement of mobile nodes. The previously labeled efforts generate a large quantity of CH, which degraded the performance. To address this issue, the clustering method incorporates grey wolf optimization. The grey wolf's natural hunting habit is employed to construct efficient clusters, resulting in the optimal number of clusters. The simulation findings give communication quality and a dependable information delivery ratio in VANET.

The study in [9, 89] proposed an efficient route repair approach to improve VANET network performance by combining ant colony optimization with an AODV routing system. This approach improves connection stability, packet delivery ratio, vehicle speed, and network quality. However, cluster head overburdening in cluster-based VANET communication is a problem. The research proposed in [77] a multi cluster head selection method, divided into two sections: hybrid fuzzy multi criteria decision making protocol and fuzzy analytic hierarchy protocol. The tributary topic is intrusion detection, focusing on support vector machine and dolphin swarm optimization. An adaptive updating approach was proposed in [14] to tackle the problem of rising channel traffic and congestion by improving the transmission of beacon signals. The method entails transmitting a beacon message by considering the participation of nodes in the forwarding set and the estimated duration of connection availability. The output metrics consist of packet delivery ratio, control packets, and routing

overhead. The projected technique shows promise compared to previous models, but it is still in its early stages and has limited applicability to different mobility paradigms.

Previously, one hop clustering was considered, which reduces coverage and increases cluster heads, impacting network performance and cluster overlaps. Some models overlook VANET mobility features, dynamic topology, and restricted driving direction. Earlier mobility-based clustering algorithms caused network congestion and increased collision rate.

Discussions and Prognosis. In the dynamic landscape of vehicular networks, the deployment of two-way cooperative Non-Orthogonal Multiple Access (NOMA) presents a multitude of intricate research challenges. While this approach holds considerable promise in enhancing both spectral efficiency and reliability, its successful implementation hinges on addressing several critical issues that warrant immediate attention for further exploration.

One of the primary challenges lies in optimizing resource allocation strategies to effectively cater to the diverse and dynamic nature of vehicular communication environments. Given the varying mobility patterns and communication requirements of vehicles, devising adaptive resource allocation schemes capable of efficiently utilizing available resources is essential to ensure optimal network performance. Furthermore, mitigating the impact of fluctuating channel conditions poses a significant hurdle. In the context of vehicular networks, where vehicles are constantly in motion, channel conditions can vary rapidly, leading to fluctuations in signal strength and quality. Developing robust techniques to adaptively adjust transmission parameters in response to these variations is crucial for maintaining reliable communication links.

Interference management also emerges as a critical concern. With numerous vehicles transmitting and receiving data simultaneously in close proximity, managing interference becomes inherently challenging. Effective interference mitigation techniques are necessary to minimize signal degradation and ensure seamless communication within the vehicular network.

Moreover, the integration of cooperative NOMA (C-NOMA) with high diversity gain, stringent latency requirements, and reliability constraints of Intelligent Transportation Systems (ITS) warrants thorough investigation. Balancing the trade-offs between spectral efficiency, latency, and reliability in the context of vehicular communications poses a complex optimization problem that requires careful consideration.

As the automotive industry progresses towards a future dominated by connected and autonomous vehicles, addressing these challenges becomes imperative. Unlocking the full potential of two-way C-NOMA in vehicular communications holds the key to driving advancements in intelligent transportation systems and vehicular communication technologies, ultimately paving the way for safer, more efficient, and smarter transportation systems.

8. Conclusion. This comprehensive survey has shed light on the multifaceted landscape of Cooperative Vehicular Communications in Intelligent Transportation Systems (ITS), emphasizing the pressing research challenges that must be addressed for successful implementation in realistic scenarios. The dynamic nature of vehicular mobility, coupled with the need for real-time communication, link reliability, coverage, interference, and adaptability, presents a complex set of hurdles that demand immediate attention from the research community. According to the comprehensive survey in this paper most of the solutions provided unheeded the influence of mobility in relay selection, and the estimated CSI in relay selection is outdated during data transmission. Future research efforts should prioritize the development of creative solutions that that can effectively connect theoretical frameworks with practical applications, fostering the creation of resilient, efficient, and secure cooperative vehicular systems that can truly transform the landscape of intelligent transportation. It is crucial to tackle these challenges head-on in order to fully unleash the potential of CVNs and create a safer and more sustainable future in transportation.

REFERENCES

- [1] *Ieee standard for wireless access in vehicular environments (wave)-identifiers*, IEEE Std 1609.12-2019 (Revision of IEEE Std 1609.12-2016), (2019), pp. 1–17.
- [2] O. ABBASI AND A. EBRAHIMI, *Cooperative noma with full-duplex amplify-and-forward relaying*, Transactions on Emerging Telecommunications Technologies, 29 (2018), p. e3421.

- [3] T. S. ABRAHAM AND K. NARAYANAN, *Cooperative communication for vehicular networks*, in 2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies, IEEE, 2014, pp. 1163–1167.
- [4] M. S. ALI, E. HOSSAIN, AND D. I. KIM, *Coordinated multi-point (comp) transmission in downlink multi-cell noma systems: Models and spectral efficiency performance*, arXiv preprint arXiv:1703.09255, (2017).
- [5] G. ARANITI, C. CAMPOLO, M. CONDOLUCI, A. IERA, AND A. MOLINARO, *Lte for vehicular networking: A survey*, IEEE communications magazine, 51 (2013), pp. 148–157.
- [6] I. AZAM AND S. Y. SHIN, *On the performance of sic-free spatial modulation aided uplink noma under imperfect csi*, ICT Express, 9 (2023), pp. 76–81.
- [7] J. BAE AND Y. HAN, *Joint power and time allocation for two-way cooperative noma*, IEEE Transactions on Vehicular Technology, 68 (2019), pp. 12443–12447.
- [8] M. W. BIDAS, A. M. ABDELGAFFAR, AND E. ALSUSA, *Network-coded uplink clustered noma relay networks: Models and performance comparisons*, Computer Networks, 220 (2023), p. 109465.
- [9] H. BELLO-SALAU, A. AIBINU, Z. WANG, A. ONUMANYI, E. ONWUKA, AND J. DUKIYA, *An optimized routing algorithm for vehicle ad-hoc networks*, Engineering Science and Technology, an International Journal, 22 (2019), pp. 754–766.
- [10] S. BHARATI AND W. ZHUANG, *Crb: Cooperative relay broadcasting for safety applications in vehicular networks*, IEEE Transactions on Vehicular Technology, 65 (2016), pp. 9542–9553.
- [11] A. BLETSAS, A. KHISTI, D. P. REED, AND A. LIPPMAN, *A simple cooperative diversity method based on network path selection*, IEEE Journal on selected areas in communications, 24 (2006), pp. 659–672.
- [12] M. M. BUTT, A. NASIR, A. MOHAMED, AND M. GUIZANI, *Trading wireless information and power transfer: Relay selection to minimize the outage probability*, in 2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP), IEEE, 2014, pp. 253–257.
- [13] Y. CAI, Z. QIN, F. CUI, G. Y. LI, AND J. A. MCCANN, *Modulation and multiple access for 5g networks*, IEEE Communications Surveys & Tutorials, 20 (2017), pp. 629–646.
- [14] M. CHAHAL AND S. HARIT, *Network selection and data dissemination in heterogeneous software-defined vehicular network*, Computer Networks, 161 (2019), pp. 32–44.
- [15] H. CHEN, L. XIAO, D. YANG, T. ZHANG, AND L. CUTHBERT, *User cooperation in wireless powered communication networks with a pricing mechanism*, IEEE Access, 5 (2017), pp. 16895–16903.
- [16] J. CHEN, G. MAO, C. LI, A. ZAFAR, AND A. Y. ZOMAYA, *Throughput of infrastructure-based cooperative vehicular networks*, IEEE Transactions on Intelligent Transportation Systems, 18 (2017), pp. 2964–2979.
- [17] J. CHEN, A. ZAFAR, G. MAO, AND C. LI, *On the achievable throughput of cooperative vehicular networks*, in 2016 IEEE International Conference on Communications (ICC), IEEE, 2016, pp. 1–7.
- [18] R. CHEN, Z. SHENG, Z. ZHONG, M. NI, V. C. LEUNG, D. G. MICHELSON, AND M. HU, *Connectivity analysis for cooperative vehicular ad hoc networks under nakagami fading channel*, IEEE Communications Letters, 18 (2014), pp. 1787–1790.
- [19] S. CHEN, J. HU, Y. SHI, AND L. ZHAO, *Technologies, standards and applications of lte-v2x for vehicular networks*, Telecommunications Science, 34 (2018), pp. 1–11.
- [20] W.-H. CHEN, A.-C. PANG, S.-C. HU, AND C.-T. F. CHIANG, *Cross-layer cooperative routing for vehicular networks*, in 2010 International Computer Symposium (ICS2010), IEEE, 2010, pp. 67–72.
- [21] X. CHEN, R. JIA, AND D. W. K. NG, *The application of relay to massive non-orthogonal multiple access*, IEEE Transactions on Communications, 66 (2018), pp. 5168–5180.
- [22] Y. CHEN, M. FANG, S. SHI, W. GUO, AND X. ZHENG, *Distributed multi-hop clustering algorithm for vanets based on neighborhood follow*, Eurasisp journal on Wireless communications and networking, 2015 (2015), pp. 1–12.
- [23] Y. CHEN, L. WANG, AND B. JIAO, *Cooperative multicast non-orthogonal multiple access in cognitive radio*, in 2017 IEEE International Conference on Communications (ICC), IEEE, 2017, pp. 1–6.
- [24] X. CHENG, C. CHEN, W. ZHANG, AND Y. YANG, *5g-enabled cooperative intelligent vehicular (5genciv) framework: When benz meets marconi*, IEEE Intelligent Systems, 32 (2017), pp. 53–59.
- [25] Y. CHOI AND D. KIM, *Quality-supporting duration for dual-hop vehicle-to-vehicle cooperative communications*, in 2013 International Conference of Information and Communication Technology (ICoICT), IEEE, 2013, pp. 33–37.
- [26] L. DAI, B. WANG, Y. YUAN, S. HAN, I. CHIH-LIN, AND Z. WANG, *Non-orthogonal multiple access for 5g: solutions, challenges, opportunities, and future research trends*, IEEE Communications Magazine, 53 (2015), pp. 74–81.
- [27] J. R. DAWANDE, S. SILAKARI, AND A. J. DEEN, *Enhanced distributed multi-hop clustering algorithm for vanets based on neighborhood follow (edmcnf) collaborated with road side units*, in 2015 International Conference on Computational Intelligence and Communication Networks (CICN), IEEE, 2015, pp. 106–113.
- [28] B. DI, L. SONG, Y. LI, AND Z. HAN, *V2x meets noma: Non-orthogonal multiple access for 5g-enabled vehicular networks*, IEEE Wireless Communications, 24 (2017), pp. 14–21.
- [29] B. DI, L. SONG, Y. LI, AND G. Y. LI, *Non-orthogonal multiple access for high-reliable and low-latency v2x communications in 5g systems*, IEEE journal on selected areas in communications, 35 (2017), pp. 2383–2397.
- [30] Z. DING, I. KRIKIDIS, B. SHARIF, AND H. V. POOR, *Wireless information and power transfer in cooperative networks with spatially random relays*, IEEE Transactions on Wireless Communications, 13 (2014), pp. 4440–4453.
- [31] Z. DING AND K. K. LEUNG, *Cross-layer routing using cooperative transmission in vehicular ad-hoc networks*, IEEE Journal on Selected Areas in Communications, 29 (2011), pp. 571–581.
- [32] Z. DING, Y. LIU, J. CHOI, Q. SUN, M. ELKASHLAN, I. CHIH-LIN, AND H. V. POOR, *Application of non-orthogonal multiple access in lte and 5g networks*, IEEE Communications Magazine, 55 (2017), pp. 185–191.
- [33] Z. DING, M. PENG, AND H. V. POOR, *Cooperative non-orthogonal multiple access in 5g systems*, IEEE Communications Letters, 19 (2015), pp. 1462–1465.
- [34] D.-T. DO, A.-T. LE, AND B. M. LEE, *On performance analysis of underlay cognitive radio-aware hybrid oma/noma networks*

- with imperfect csi*, *Electronics*, 8 (2019), p. 819.
- [35] M. EZZAOUIA, C. GUEGUEN, M. AMMAR, S. BAHEY, X. LAGRANGE, AND A. BOUALLÈGUE, *A dynamic inter-cellular bandwidth fair sharing scheduler for future wireless networks*, *Physical Communication*, 25 (2017), pp. 85–99.
- [36] M. FAHAD, F. AADIL, S. KHAN, P. A. SHAH, K. MUHAMMAD, J. LLORET, H. WANG, J. W. LEE, I. MEHMOOD, ET AL., *Grey wolf optimization based clustering algorithm for vehicular ad-hoc networks*, *Computers & Electrical Engineering*, 70 (2018), pp. 853–870.
- [37] Y. P. FALLAH, C.-L. HUANG, R. SENGUPTA, AND H. KRISHNAN, *Analysis of information dissemination in vehicular ad-hoc networks with application to cooperative vehicle safety systems*, *IEEE Transactions on Vehicular Technology*, 60 (2010), pp. 233–247.
- [38] Y. P. FALLAH, N. NASIRIANI, AND H. KRISHNAN, *Stable and fair power control in vehicle safety networks*, *IEEE Transactions on Vehicular Technology*, 65 (2015), pp. 1662–1675.
- [39] H. GHARAVI AND B. HU, *Cooperative diversity routing and transmission for wireless sensor networks*, *IET Wireless Sensor Systems*, 3 (2013), pp. 277–288.
- [40] H. HACI, H. ZHU, AND J. WANG, *Performance of non-orthogonal multiple access with a novel asynchronous interference cancellation technique*, *IEEE Transactions on Communications*, 65 (2017), pp. 1319–1335.
- [41] S. HARRABI, I. B. JAFFAR, AND K. GHEDIRA, *Novel optimized routing scheme for vanets*, *Procedia Computer Science*, 98 (2016), pp. 32–39.
- [42] K. HIGUCHI AND Y. KISHIYAMA, *Non-orthogonal access with random beamforming and intra-beam sic for cellular mimo downlink*, in 2013 IEEE 78th Vehicular Technology Conference (VTC Fall), IEEE, 2013, pp. 1–5.
- [43] C. Y. HO AND C. Y. LEOW, *Cooperative non-orthogonal multiple access using two-way relay*, in 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), IEEE, 2017, pp. 459–463.
- [44] B.-Y. HUANG, Y. LEE, AND S.-I. SOU, *Joint power allocation for noma-based diamond relay networks with and without cooperation*, *IEEE Open Journal of the Communications Society*, 1 (2020), pp. 428–443.
- [45] J. HUANG, H. GHARAVI, H. YAN, AND C.-C. XING, *Network coding in relay-based device-to-device communications*, *IEEE network*, 31 (2017), pp. 102–107.
- [46] P. HUBBALLI, A. SUTAGUNDAR, AND R. BELAGALI, *Agent based dynamic clustering for hybrid vanet (adchv)*, in 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), IEEE, 2016, pp. 382–386.
- [47] A. HYADI, M. BENJILLALI, AND M.-S. ALOUINI, *Outage performance of decode-and-forward in two-way relaying with outdated csi*, *IEEE Transactions on Vehicular Technology*, 64 (2015), pp. 5940–5947.
- [48] H. ILHAN, M. UYSAL, AND I. ALTUNBAS, *Cooperative diversity for intervehicular communication: Performance analysis and optimization*, *IEEE Transactions on Vehicular Technology*, 58 (2009), pp. 3301–3310.
- [49] S. R. ISLAM, N. AVAZOV, O. A. DOBRE, AND K.-S. KWAK, *Power-domain non-orthogonal multiple access (noma) in 5g systems: Potentials and challenges*, *IEEE Communications Surveys & Tutorials*, 19 (2016), pp. 721–742.
- [50] R. JAMGEKAR AND S. TAPKIRE, *A robust multi-hop clustering algorithm for reliable vanet message dissemination*, in 2017 international conference on energy, communication, data analytics and soft computing (ICECDS), IEEE, 2017, pp. 2599–2604.
- [51] Q. JIANG, H. WANG, T. YUAN, X. TAO, AND Q. CUI, *Overlaid-pilot based channel state information feedback for multicell cooperative networks*, in 2015 IEEE 26th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), IEEE, 2015, pp. 223–227.
- [52] M. F. KADER, M. B. SHAHAB, AND S. Y. SHIN, *Exploiting non-orthogonal multiple access in cooperative relay sharing*, *IEEE Communications Letters*, 21 (2017), pp. 1159–1162.
- [53] F. KARA AND H. KAYA, *Threshold-based selective cooperative-noma*, *IEEE Communications Letters*, 23 (2019), pp. 1263–1266.
- [54] G. KARAGIANNIS, O. ALTINTAS, E. EKICI, G. HEIJENK, B. JARUPAN, K. LIN, AND T. WEIL, *Vehicular networking: A survey and tutorial on requirements, architectures, challenges, standards and solutions*, *IEEE communications surveys & tutorials*, 13 (2011), pp. 584–616.
- [55] F. KHALID AND S. JANGSHER, *Upper bound of capacity for a mu-mimo noma in a two way relaying network*, in 2018 6th International Conference on Wireless Networks and Mobile Communications (WINCOM), IEEE, 2018, pp. 1–6.
- [56] K. KHALIL, G. FARHADI, AND A. ITO, *Iterative fair channel assignment for wireless networks*, *IEEE Wireless Communications Letters*, 3 (2014), pp. 145–148.
- [57] J.-B. KIM AND I.-H. LEE, *Non-orthogonal multiple access in coordinated direct and relay transmission*, *IEEE Communications Letters*, 19 (2015), pp. 2037–2040.
- [58] Y. KIM, F. BACCELLI, AND G. DE VECIANA, *Spatial reuse and fairness of ad hoc networks with channel-aware csma protocols*, *IEEE transactions on information theory*, 60 (2014), pp. 4139–4157.
- [59] Y.-B. KIM, K. YAMAZAKI, AND B. C. JUNG, *Virtual full-duplex cooperative noma: Relay selection and interference cancellation*, *IEEE Transactions on Wireless Communications*, 18 (2019), pp. 5882–5893.
- [60] B. KIMY, S. LIM, H. KIM, S. SUH, J. KWUN, S. CHOI, C. LEE, S. LEE, AND D. HONG, *Non-orthogonal multiple access in a downlink multiuser beamforming system*, in MILCOM 2013-2013 IEEE Military Communications Conference, IEEE, 2013, pp. 1278–1283.
- [61] C. LAI, K. ZHANG, N. CHENG, H. LI, AND X. SHEN, *Sirc: A secure incentive scheme for reliable cooperative downloading in highway vanets*, *IEEE Transactions on Intelligent Transportation Systems*, 18 (2016), pp. 1559–1574.
- [62] J. N. LANEMAN, D. N. TSE, AND G. W. WORNELL, *Cooperative diversity in wireless networks: Efficient protocols and outage behavior*, *IEEE Transactions on Information theory*, 50 (2004), pp. 3062–3080.
- [63] C. LI, B. XIA, S. SHAO, Z. CHEN, AND Y. TANG, *Multi-user scheduling of the full-duplex enabled two-way relay systems*,

- IEEE Transactions on Wireless Communications, 16 (2016), pp. 1094–1106.
- [64] G. LI, D. MISHRA, Y. HU, Y. HUANG, AND H. JIANG, *Adaptive relay selection strategies for cooperative noma networks with user and relay cooperation*, IEEE Transactions on Vehicular Technology, 69 (2020), pp. 11728–11742.
- [65] X. LI AND J. WANG, *A generous cooperative routing protocol for vehicle-to-vehicle networks*, KSII Transactions on Internet and Information Systems (TIIS), 10 (2016), pp. 5322–5342.
- [66] Y. LI AND G. AMARASURIYA, *Relay-aided massive mimo noma downlink*, in 2018 IEEE Global Communications Conference (GLOBECOM), IEEE, 2018, pp. 1–7.
- [67] Y. LI, Y. LI, X. CHU, Y. YE, AND H. ZHANG, *Performance analysis of relay selection in cooperative noma networks*, IEEE Communications Letters, 23 (2019), pp. 760–763.
- [68] M. LIAQAT, K. A. NOORDIN, T. ABDUL LATEF, AND K. DIMYATI, *Power-domain non orthogonal multiple access (pd-noma) in cooperative networks: an overview*, Wireless Networks, 26 (2020), pp. 181–203.
- [69] K. LIU, J. K.-Y. NG, J. WANG, V. C. LEE, W. WU, AND S. H. SON, *Network-coding-assisted data dissemination via cooperative vehicle-to-vehicle/-infrastructure communications*, IEEE Transactions on Intelligent Transportation Systems, 17 (2015), pp. 1509–1520.
- [70] Y. LIU, Z. DING, M. ELKASHLAN, AND H. V. POOR, *Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer*, IEEE Journal on Selected Areas in Communications, 34 (2016), pp. 938–953.
- [71] L. LV, Q. NI, Z. DING, AND J. CHEN, *Application of non-orthogonal multiple access in cooperative spectrum-sharing networks over nakagami-m fading channels*, IEEE Transactions on Vehicular Technology, 66 (2016), pp. 5506–5511.
- [72] S. MEHAR, S. M. SENOUCI, AND G. REMY, *Dissemination protocol for heterogeneous cooperative vehicular networks*, in 2012 IFIP Wireless Days, IEEE, 2012, pp. 1–6.
- [73] J. MEN AND J. GE, *Non-orthogonal multiple access for multiple-antenna relaying networks*, IEEE Communications Letters, 19 (2015), pp. 1686–1689.
- [74] J. MEN, J. GE, AND C. ZHANG, *Performance analysis for downlink relaying aided non-orthogonal multiple access networks with imperfect csi over nakagami- $\{m\}$ fading*, IEEE Access, 5 (2016), pp. 998–1004.
- [75] A. MOSTAFA, R. KOBYLINSKI, I. KOSTANIC, AND M. AUSTIN, *Single antenna interference cancellation (saic) for gsm networks*, in 2003 IEEE 58th Vehicular Technology Conference. VTC 2003-Fall (IEEE Cat. No. 03CH37484), vol. 2, IEEE, 2003, pp. 1089–1093.
- [76] F. MUKHLIF, K. A. B. NOORDIN, A. M. MANSOOR, AND Z. M. KASIRUN, *Green transmission for c-ran based on swipt in 5g: a review*, Wireless Networks, 25 (2019), pp. 2621–2649.
- [77] M. NADERI, F. ZARGARI, AND M. GHANBARI, *Adaptive beacon broadcast in opportunistic routing for vanets*, Ad Hoc Networks, 86 (2019), pp. 119–130.
- [78] R. PAL, A. PRAKASH, R. TRIPATHI, AND D. SINGH, *Analytical model for clustered vehicular ad hoc network analysis*, Ict Express, 4 (2018), pp. 160–164.
- [79] X. PEI, H. YU, M. WEN, S. MUMTAZ, S. AL OTAIBI, AND M. GUIZANI, *Noma-based coordinated direct and relay transmission with a half-duplex/full-duplex relay*, IEEE Transactions on communications, 68 (2020), pp. 6750–6760.
- [80] B. REBEKKA, B. V. KUMAR, AND B. MALARKODI, *Radio resource allocation with energy efficiency-throughput balancing for lte downlink*, in 2015 2nd International Conference on Electronics and Communication Systems (ICECS), IEEE, 2015, pp. 111–115.
- [81] C. REN, H. ZHANG, J. WEN, J. CHEN, AND C. TELLAMBURA, *Successive two-way relaying for full-duplex users with generalized self-interference mitigation*, IEEE Transactions on Wireless Communications, 18 (2018), pp. 63–76.
- [82] M. REN, L. KHOUKHI, H. LABIOD, J. ZHANG, AND V. VEQUE, *A mobility-based scheme for dynamic clustering in vehicular ad-hoc networks (vanets)*, Vehicular Communications, 9 (2017), pp. 233–241.
- [83] P. SADEGHI, M. YU, AND N. ABOUTORAB, *On throughput-delay tradeoff of network coding for wireless communications*, in 2014 International Symposium on Information Theory and its Applications, IEEE, 2014, pp. 689–693.
- [84] Y. SAITO, Y. KISHIYAMA, A. BENJEBBOUR, T. NAKAMURA, A. LI, AND K. HIGUCHI, *Non-orthogonal multiple access (noma) for cellular future radio access*, in 2013 IEEE 77th vehicular technology conference (VTC Spring), IEEE, 2013, pp. 1–5.
- [85] W. SANG, D. SHEN, W. REN, AND X. SHUAI, *A survey of capacity in cooperative relay networks*, in 2011 Global Mobile Congress, IEEE, 2011, pp. 1–8.
- [86] B. SELIM, S. MUHAIDAT, P. C. SOFOTASIOS, B. S. SHARIF, T. STOURAITIS, G. K. KARAGIANNIDIS, AND N. AL-DHAHIR, *Performance analysis of non-orthogonal multiple access under i/q imbalance*, IEEE Access, 6 (2018), pp. 18453–18468.
- [87] S. A. A. SHAH, E. AHMED, M. IMRAN, AND S. ZEADALLY, *5g for vehicular communications*, IEEE Communications Magazine, 56 (2018), pp. 111–117.
- [88] C. E. SHANNON, *Two-way communication channels*, in Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, vol. 4, University of California Press, 1961, pp. 611–645.
- [89] S. SHARMA AND A. KAUL, *Hybrid fuzzy multi-criteria decision making based multi cluster head dolphin swarm optimized ids for vanet*, Vehicular Communications, 12 (2018), pp. 23–38.
- [90] W. SHIN, H. YANG, M. VAEZI, J. LEE, AND H. V. POOR, *Relay-aided noma in uplink cellular networks*, IEEE Signal Processing Letters, 24 (2017), pp. 1842–1846.
- [91] M. SHIRKHANI, Z. TIRKAN, AND A. TAHERPOUR, *Performance analysis and optimization of two-way cooperative communications in inter-vehicular networks*, in 2012 International Conference on Wireless Communications and Signal Processing (WCSP), IEEE, 2012, pp. 1–6.
- [92] K. SJOBERG, P. ANDRES, T. BUBURUZAN, AND A. BRAKEMEIER, *Cooperative intelligent transport systems in europe: Current deployment status and outlook*, IEEE Vehicular Technology Magazine, 12 (2017), pp. 89–97.
- [93] L. SONG, Y. LI, Z. DING, AND H. V. POOR, *Resource management in non-orthogonal multiple access systems: State of the*

- art and research challenges*, arXiv preprint arXiv, 1610 (2016).
- [94] T. SONI, A. R. ALI, K. GANESAN, AND M. SCHELLMANN, *Adaptive numerology solution to address the demanding qos in 5g-v2x*, in 2018 IEEE Wireless Communications and Networking Conference (WCNC), IEEE, 2018, pp. 1–6.
- [95] I. D. SPHERE-PACKING, *Near-capacity multi-functional mimo systems*.
- [96] R. SUGUMAR, A. RENGARAJAN, AND C. JAYAKUMAR, *Trust based authentication technique for cluster based vehicular ad hoc networks (vanet)*, *Wireless Networks*, 24 (2018), pp. 373–382.
- [97] Q. SUN, S. HAN, I. CHIN-LIN, AND Z. PAN, *On the ergodic capacity of mimo noma systems*, *IEEE Wireless Communications Letters*, 4 (2015), pp. 405–408.
- [98] S.-H. SUN, J.-L. HU, Y. PENG, X.-M. PAN, L. ZHAO, AND J.-Y. FANG, *Support for vehicle-to-everything services based on lte*, *IEEE Wireless Communications*, 23 (2016), pp. 4–8.
- [99] H. TABASSUM, M. S. ALI, E. HOSSAIN, M. J. HOSSAIN, AND D. I. KIM, *Non-orthogonal multiple access (noma) in cellular uplink and downlink: Challenges and enabling techniques*, arXiv preprint arXiv:1608.05783, (2016).
- [100] S. TIMOTHEOU AND I. KRIKIDIS, *Fairness for non-orthogonal multiple access in 5g systems*, *IEEE signal processing letters*, 22 (2015), pp. 1647–1651.
- [101] A. TREGANCINI, E. E. B. OLIVO, D. P. M. OSORIO, C. H. DE LIMA, AND H. ALVES, *Performance analysis of full-duplex relay-aided noma systems using partial relay selection*, *IEEE Transactions on Vehicular Technology*, 69 (2019), pp. 622–635.
- [102] Y.-L. TSENG, *Lte-advanced enhancement for vehicular communication*, *IEEE Wireless Communications*, 22 (2015), pp. 4–7.
- [103] K. TUTUNCUOGLU AND A. YENER, *Cooperative energy harvesting communications with relaying and energy sharing*, in 2013 IEEE Information Theory Workshop (ITW), IEEE, 2013, pp. 1–5.
- [104] S. UMAMAHESWARAN AND M. SATHYA, *A comprehensive survey on cooperative relaying in industrial wireless sensor network*, *Int. J. Eng. Res. Technol.*, 6 (2017), pp. 591–596.
- [105] M. VAEZI, R. SCHOBER, Z. DING, AND H. V. POOR, *Non-orthogonal multiple access: Common myths and critical questions*, *IEEE Wireless Communications*, 26 (2019), pp. 174–180.
- [106] O. A. WAHAB, H. OTROK, AND A. MOURAD, *Vanet qos-olsr: Qos-based clustering protocol for vehicular ad hoc networks*, *Computer Communications*, 36 (2013), pp. 1422–1435.
- [107] D. WAN, M. WEN, F. JI, Y. LIU, AND Y. HUANG, *Cooperative noma systems with partial channel state information over nakagami-m fading channels*, *IEEE Transactions on Communications*, 66 (2017), pp. 947–958.
- [108] D. WAN, M. WEN, F. JI, H. YU, AND F. CHEN, *On the achievable sum-rate of noma-based diamond relay networks*, *IEEE Transactions on Vehicular Technology*, 68 (2018), pp. 1472–1486.
- [109] L. WANG, R. LI, C. CAO, AND G. L. STÜBER, *Snr analysis of time reversal signaling on target and unintended receivers in distributed transmission*, *IEEE Transactions on Communications*, 64 (2016), pp. 2176–2191.
- [110] X. WANG, M. JIA, I. W.-H. HO, Q. GUO, AND F. C. LAU, *Exploiting full-duplex two-way relay cooperative non-orthogonal multiple access*, *IEEE Transactions on Communications*, 67 (2018), pp. 2716–2729.
- [111] C. WU, Y. JI, AND T. YOSHINAGA, *A cooperative forwarding scheme for vanet routing protocols*, *ZTE Communications*, 14 (2019), pp. 13–21.
- [112] P. XU, Z. YANG, Z. DING, AND Z. ZHANG, *Optimal relay selection schemes for cooperative noma*, *IEEE Transactions on Vehicular Technology*, 67 (2018), pp. 7851–7855.
- [113] Z. YANG, Z. DING, P. FAN, AND N. AL-DHAHIR, *A general power allocation scheme to guarantee quality of service in downlink and uplink noma systems*, *IEEE transactions on wireless communications*, 15 (2016), pp. 7244–7257.
- [114] Z. YANG, Z. DING, Y. WU, AND P. FAN, *Novel relay selection strategies for cooperative noma*, *IEEE Transactions on Vehicular Technology*, 66 (2017), pp. 10114–10123.
- [115] X. YUE, Y. LIU, S. KANG, AND A. NALLANATHAN, *Performance analysis of noma with fixed gain relaying over nakagami-m fading channels*, *IEEE access*, 5 (2017), pp. 5445–5454.
- [116] X. YUE, Y. LIU, S. KANG, A. NALLANATHAN, AND Y. CHEN, *Modeling and analysis of two-way relay non-orthogonal multiple access systems*, *IEEE Transactions on Communications*, 66 (2018), pp. 3784–3796.
- [117] X. YUE, Y. LIU, S. KANG, A. NALLANATHAN, AND Z. DING, *Exploiting full/half-duplex user relaying in noma systems*, *IEEE Transactions on Communications*, 66 (2017), pp. 560–575.
- [118] ———, *Spatially random relay selection for full/half-duplex cooperative noma networks*, *IEEE Transactions on Communications*, 66 (2018), pp. 3294–3308.
- [119] X. YUE, Z. QIN, Y. LIU, S. KANG, AND Y. CHEN, *A unified framework for non-orthogonal multiple access*, *IEEE Transactions on Communications*, 66 (2018), pp. 5346–5359.
- [120] J. ZHANG, Q. ZHANG, AND W. JIA, *Vc-mac: A cooperative mac protocol in vehicular networks*, *IEEE Transactions on Vehicular Technology*, 58 (2008), pp. 1561–1571.
- [121] L. ZHANG, B. JIN, AND Y. CUI, *A concurrent transmission enabled cooperative mac protocol for vehicular ad hoc networks*, in 2014 IEEE 22nd International Symposium of Quality of Service (IWQoS), IEEE, 2014, pp. 258–267.
- [122] T. ZHANG AND Q. ZHU, *A tdma based cooperative communication mac protocol for vehicular ad hoc networks*, in 2016 IEEE 83rd Vehicular Technology Conference (VTC Spring), IEEE, 2016, pp. 1–6.
- [123] X. ZHANG, G. MAO, X. TAO, AND Q. CUI, *Uncoordinated cooperative forwarding in vehicular networks with random transmission range*, in 2015 IEEE Global Communications Conference (GLOBECOM), IEEE, 2015, pp. 1–7.
- [124] Q. ZHAO AND H. LI, *Differential modulation for cooperative wireless systems*, *IEEE Transactions on Signal Processing*, 55 (2007), pp. 2273–2283.
- [125] B. ZHENG, X. WANG, M. WEN, AND F. CHEN, *Noma-based multi-pair two-way relay networks with rate splitting and group decoding*, *IEEE Journal on Selected Areas in Communications*, 35 (2017), pp. 2328–2341.
- [126] K. ZHENG, F. LIU, Q. ZHENG, W. XIANG, AND W. WANG, *A graph-based cooperative scheduling scheme for vehicular*

- networks*, IEEE transactions on vehicular technology, 62 (2013), pp. 1450–1458.
- [127] T. ZHOU, H. SHARIF, M. HEMPEL, P. MAHASUKHON, W. WANG, AND T. MA, *A novel adaptive distributed cooperative relaying mac protocol for vehicular networks*, IEEE Journal on Selected Areas in Communications, 29 (2010), pp. 72–82.
- [128] L. ZHU, F. R. YU, B. NING, AND T. TANG, *A joint design of security and quality-of-service (qos) provisioning in vehicular ad hoc networks with cooperative communications*, EURASIP Journal on Wireless Communications and Networking, 2013 (2013), pp. 1–14.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Dec 14, 2023

Accepted: Feb 23, 2024



PA FUZZY-NOISE REMOVAL IN WIRELESS SENSORS NETWORKS

B HARISH GOUD AND RAJU ANITHA *

Abstract. In the wireless sensors network, a large number of sensor device data is sent directly to the base station. So huge amount of noise is also added with data going to the base station and there is no security mechanism for protecting sensor device data in the existing scenario. WSN has numerous applications, including healthcare systems, secure military applications, and monitoring applications. Achievement of noise removal is essential for WSN. Many researchers have focused on enhancing the removal of noise in data and extending the network lifetime. Sensor Nodes (SNs), cluster heads (CHs), and base stations make up the standard WSN architecture. The communication of SNs using the traditional design consumes high energy increases delay and reduces network performance. To address the limitation of the present state of the system, this research work proposed a PA Fuzzy system which is acting like a filter used to remove unnecessary noise with sensor data that is moving toward the base station. And PA Fuzzy system after removing noise, and sensor data is encrypted so that it can be protected from hackers. It makes the network performance better, decreases delay and energy use, and increases the Ratio of packet deliveries and throughput. The execution of the suggested methodology was made using NS2. The proposed's empirical outcomes system outperforms with comparison of the existing WSN mechanisms.

Key words: Wireless Sensor Networks, PA Fuzzy System, Filters, Encryption technique.

1. Introduction. The exponential growth of wireless communication technologies is having a profound effect on wireless sensor networks (WSNs). The conveyance of data is an essential WSN function. Over the past decade, numerous mechanisms have been suggested in an effort to enhance data transmission efficacy. With effectiveness Data transmission is essential for both research and commerce. Wireless sensor networks (WSNs) have a wide range of applications, including but not limited to smart cities, the armed services, and advancements in the healthcare industry [1]. A considerable quantity of dispersed sensor nodes comprise the WSN. Traditional WSNs feature varying degrees of communication. An excessive number of clusters are formed, and one CH is elected for each cluster. The sensor nodes gather environmental data and are utilized in a variety of applications. The CH receives and transmits sensor data from the base station after receiving it from the sensor node. [2]. The WSN sensor nodes have limitations in energy efficiently and data transfer. The sensor nodes' connectivity and computation abilities are incredibly poor. The sensor nodes' range is somewhat constrained; thus, improvements are required to boost communication efficiency. The primary component of load balancing is the deployment of WSN. Base Stations and CHs were the primary data transmission devices in classic WSN [3]. Look at the WSN model architecture in Figure 1.1.

The WSNs use a clustering-based approach, with a set of sensor nodes in each cluster. Information about the area is gathered and sent to the CHs by sensor nodes. The Data was gathered by CH's method and forwarded to BS. To gather and transmit the data, each sensor node in the network uses energy. Despite the fact that the sensor nodes shut down when their energy runs out. Therefore, creating WSN requires an energy-efficient algorithm. To equalize WSN's energy usage, many clustering techniques have been developed [4], [5]. These algorithms follow the selection of CHs and also shift CHs position among the SNs in a cluster. An energy-efficient hybrid clustering and routing technique has been developed for big WSNs. Design energy efficient mechanism back-off timer and gradient routing to execute the CH selection. In This research work introduced a brand-new PA fuzzy mechanism based on intelligent agents, which minimizes energy consumption, delay and maximizes throughput and PDR [2].

The following sections of the paper are broken down and discussed individually. Section II provides a full study of the most modern WSN energy-efficient approaches. The recommended methodology for deploying

*Dept of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh 522502, India. (bhg120109@gmail.com, anitharaju@kluniversity.in)

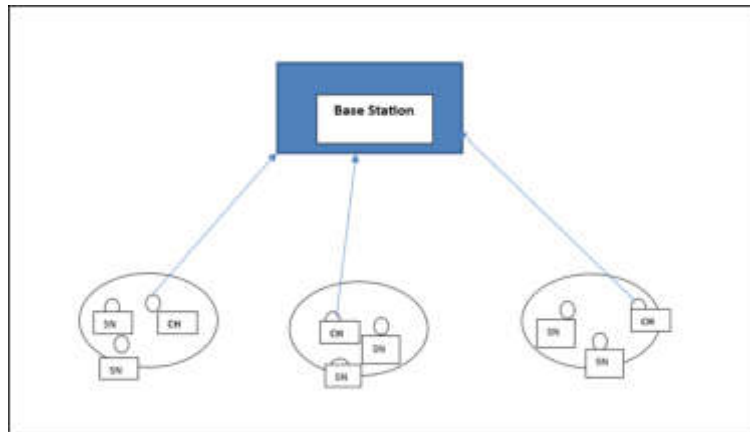


Fig. 1.1: An Illustration of a Wireless Sensor Network Architecture

PAWSN for energy efficiency was discussed in Sections III and IV, discuss the proposed mechanism from the empirical results, and compare the suggested results. In Section V, the paper is concluded with improvements to the suggested mechanism, as well as a discussion of upcoming research efforts.

2. Literature Review. An energy-conscious routing algorithm was developed by F. Fernando Jurado-Lasso et al. [6] as part of an overhead reduction strategy. This algorithm enables industrial services and optimized energy use in WSN. Data packet aggregation is a feature that the implementation of the software-defined multichip WSN to manage the WSN. The suggested system analyses shortest-path strategies in order to lengthen the lifetime of the WSN network. Enhance the PDR performance as well. Even though the recommended solution outperforms the current algorithms in terms of performance. But a revolutionary innovation algorithm is crucial in WSN to lower the network lifespan and the neighbor advertisement packets. Nelofar Aslam et al [7] built an innovative logic data algorithm transmission with clustering and reinforcement algorithm(SARSA). The proposed algorithm is also outlined as combining an ideal solution with SARSA clustering for energy consumption and network stability. The WSN node is designed with taportable wireless charging system The suggested strategy enhances the functionality of the network by drawing its inspiration from an objective function. However, C-SARSA's deployment in WSN led to an improvement in performance. However, the WSN does not have an RWSN with a proper deployment and recharge schema.

Gajendran Malshetty et al [8] For effective clustering in WSN, a self-organization method based on load has been developed. The LBSO approach in WSN employed three unique phases.. The first phase involved choosing the Cluster Head, and the second involved choosing among the sensor node clusters. The third stage is then followed by the rotational phase-based reselection of the cluster head. However, the network's efficiency and network development both increased. But the network performance is poor because of a variety of base station deployments and dead nodes. Muhammad Adil et al [9] A load-balancing routing system that uses little energy was made to make WSNs last longer. A good hybrid routing method has been made with the Dynamic Cluster Based Static Routing Protocol (DCBSRP). Ad-hoc On-Demand Distance Vector (AODV) Routing Protocol and Low-Energy Adaptive Clustering Hierarchy (LEACH) Protocol are both parts of the suggested protocol.create a WSN using a variety of clusters and CHs. The DCBSRP protocol largely behaves as a normal node and does not make current CH nodes from the early cycles public. But the proposed protocol significantly increased network longevity.

A wide variety of real-time applications using WSN are deployed. The WSN uses self-organization and a finite amount of energy. To address energy efficiency's limits Al Xinlu Li et al. [10] presented a load- balancing energy-efficient WSNs using the ant-based routing algorithm (EBAR). The EBAR algorithm effectively lowers energy use Through an opportunistic broadcast technique, EBAR uses and manages overhead in WSN to conserve energy. The EBAR does accomplish Despite being accurate only in homogeneous networks, energy

efficiency. Data transmission cannot be supported by the algorithm in a diverse network, which results in excessive energy usage [11].

Wireless Sensor Networks are wireless systems comprising a large number of randomly or regularly distributed sensor nodes. The target of this job is to give network protection to wireless Sensor Networks so as to transmit detection information to the recipient efficiently so the duration of the system is long and within this function, a novel protocol was created with Game Theory [12]. Game Theory gives a mathematical foundation for the evaluation of interactive decision-making procedures. It gives tools for predicting what may (and what should) occur when agents that have conflicting interests socialize. It's not a monolithic method, but a selection of modeling programs that assist in the comprehension of interactive decisions to get issues. The projected Game Theory approaches are implemented effectively for preventing Denial of service attacks, to discover and protect against malicious behavior of sensor nodes in a network of wireless sensors, and verified that the operation of those games considerably decreases misbehavior of tunnels, conserves node power and prolongs the network lifetime economically [13].

Wireless sensor networks are urgently needed and are proliferating as a result of recent advancements in electronics and wireless networks (WSNs) [14]. WSNs are now crucial in a variety of fields, such as infrastructure, healthcare, agriculture, the environment, and military leadership. Several issues affect the healthcare sector, among these are escalating costs, an aging population, a rise in medical mistakes, a lack of manpower, etc. Despite the challenges, healthcare professionals are under pressure to adopt new technology and offer improved services [15]. The availability of universal healthcare can lower long-term expenditures and raise service standards [16]. Wireless sensor networks offer practical remedies for the pervasive healthcare system. Recent developments in medical sensors and low-power network architectures have given rise to WSNs for the healthcare industry. As a result of the wireless sensor network [17].

3. Suggested Approach.

3.1. Status of the Problem. Sensor nodes that are connected to the WSNs communicate with one another to gather information about their surroundings. Whenever sensor data information moves from cluster heads to the base station even noise, and unwanted data also move to the base station there to so much delay, PDR is less, and Throughput is also not accurate. The SNs run in a decentralized, low-energy manner. Use of the WSN in numerous emerging applications, including military applications, applications in industry and the environment, and healthcare systems. The base station, CHs, and SNs are the three operational levels of the conventional WSN. The sensor nodes gather local data and send it to the CH Nevertheless, the CH only receives information for a limited period of time, and the behavior rotates. The changes in the rotation of the cluster head use energy and cause data transfer to be slower., In this scholarly document published a book Intelligent agent PA fuzzy mechanism.

3.2. PA-Fuzzy. The functionality of available resources, portability in sensor nodes, and rotating movements in cluster heads result in excessive energy use and latency in transmission. In order to address the issue with the current state of the WSN, this study effort introduced an innovative PA-Fuzzy is a mechanism which will eliminate unnecessary data, such as a noise which is moving to words base station it will a love only data. And PA Fuzzy system will encrypt sensor data for two times and forwarded to base station for encryption AES 256 key algorithm can be used in PA Fuzzy system. Base station sends data finally to end user so only end user can decrypt data with valid keys. In between no hacker or base station authority cannot decrypt data this is the originality of proposed technic in this article. By this we can improve PDR, Throughput, Delay.

PA-Fuzzy also a love to make unclear data to clear data. There are two ways to transmit data in the WSN. A sensor node to a PA Fuzzy system is the first data transfer type. From one cluster sensor node to another cluster sensor node is the second way that data is sent.. The sensor device and the path arbitrary node are connected directly. Use a different sensor node as an intermediary in sensor-to-sensor transmission, however. Here, choosing the best route between the destination sensor and the sender sensor node depends heavily on the path arbitrary. The path from source to destination arbitrarily defines an optimized primary path as well as alternative routes. The connecting node's node distance and energy levels are taken into account while choosing the how to get from point A to point B. if any of the primary direction links or nodes fail.

In the suggested PA-Fuzzy, the optimize path selection method includes two steps. In the initial stage of

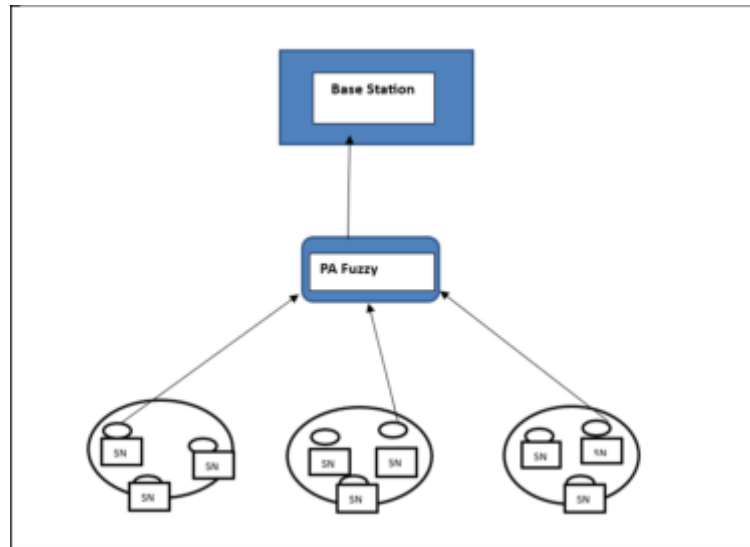


Fig. 3.1: PA Fuzzy in WSNs.

the appropriate optimize path selection process, The source node sends the RREQ packet to the destination node. The RREQ packet is sent by intermediate nodes to the target node. The sensor node has knowledge of the network's preceding and subsequent neighbors distance and energy values. Finally, RREQ from multiple nodes via various channels is received by the target node.

Algorithm: PA Fuzzy system

Input: Noise removal filters and encryption.

Output: Path optimization.

1. Start up
2. Sensor data is collected by PA Fuzzy system.
3. PA Fuzzy system does not change rational like cluster head.
4. PA Fuzzy it is fixed selected based on particle swarm optimization techniques optimized distance from cluster and base station.
5. PA Fuzzy system act like filter removes noise moving data toward base station.
6. PA Fuzzy system twice encrypt data for protection.
7. Base station receive data and Forward to end users.
8. End user receive keys and decrypt data.
9. End.

All optimized pathways were calculated and arranged using the node value sum. The destination uses the primary optimized path to send a packet of RREP data to the original station. additional network-wide optimized diversions Use the alternative channels for data transfer if a network link or node fails. to improve the path selection algorithm after carefully following the PA Fuzzy implementation steps.

4. Result Analysis. Version 2.35 of Network Simulation (NS) is utilized to implement the unique intelligent agent-based PA- Fuzzy mechanism that has been proposed. The outcomes of the empirical simulation demonstrate how well the PA-Fuzzy performs when transmitting data. The comparative findings are discussed in the subsections below.

4.1. Contextual Simulation. Table 4.1 displays the environment of the simulation. Details about the network parameters utilized in the design of the PA Fuzzy simulation are provided in Table 4.1. The deployment of WSN employs the two-ray ground radio propagation model. The effectiveness of several performance measures

Table 4.1: Simulation Environment

S.No.	System Parameter	System Contribution
1	Model of Antenna	omnidirectional Antenna
2	Length of Queue	50
3	Routing Protocol	AODV
4	Number of nodes	100
5	Data Rate	2 M.B.
6	Basic Rate	1 M.B.
7	Total Simulation Time	100
8	Network Interface	Physical wireless
9	Interface Queue Type	Trial drop
10	Type of Channel	Wireless connection
11	Radio-Propagation	Double-Ray Ground

Table 4.2: Comparing results of IA Fuzzy's PDR performance

Simulation time	PDR Performance			
	PA fuzzy	PSNR	SSIM	MD
0	0	0	0	0
10	12	6	5	4
20	33	15	13	9
30	54	35	27	21
40	77	61	44	41
50	102	82	75	56

is evaluated to judge the suggested systems. The enhanced performance is evaluated based on metrics such as latency, throughput, packet delivery ratio, and energy usage. The peak signal to noise ratio (PSNR), the structure similarity index (SSIM), and the miss detection (MD) technique are compared to the PA-Fuzzy for successfully removing noise in WSN. The section below provides a definition of performance metrics.

4.2. Comparative Metric Analysis. The performance enhancement of the proposed systems is evaluated based on metrics such as latency, throughput, packet delivery ratio, and energy usage.

4.2.1. Packet Delivery Ratio. The ratio of packets transmitted and received at the destination node. The formula given in Equation 4.1.

$$PacketDeliveryRatio = \frac{\sum_{i=1}^n RP_i}{\sum_{i=1}^n SP_i} \quad (4.1)$$

Showcase the comparing results of IA Fuzzy's PDR performance in Table 4.2. The assuming the suitable simulation period, the performance graph on PDR in Figure 4.1. The empirical findings of the suggested PA Fuzzy mechanism and the currently in the PSNR (peak signal to noise ratio), the SSIM (structural similarity index) and MD (Miss Detection) mechanisms are shown in Figure 4.1. The simulation time for the X-Axis ranged from 0 to 50 seconds. The PDR percentage is shown on the Y-axis. With an increase in simulation time, The projected PA Fuzzy network's PDR increases. Existing mechanisms behave in a similar manner, but according to performance data, the PA Fuzzy mechanism performs better than PSNR, SSIM, MD.

4.2.2. Throughput. The number of bytes and the corresponding simulation time that were received at the destination node. The equation provided in equation 4.2.

$$Throughput = \frac{\sum_{i=1}^n P_i}{Time} * 8 \quad (4.2)$$

Showcase the comparison of IA Fuzzy's throughput performance in Table 4.3. Figure 4.2 shows the suggested

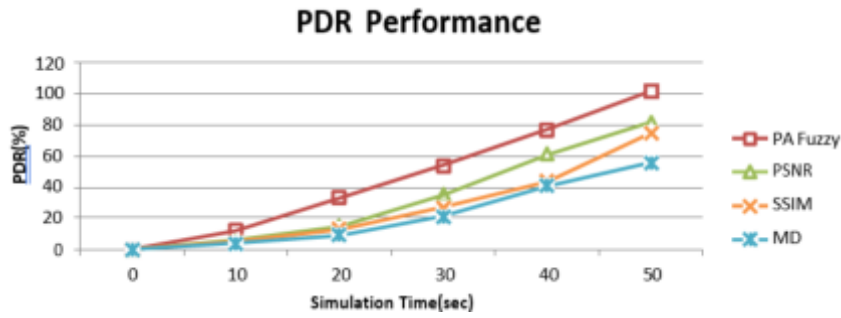


Fig. 4.1: PDR performance comparison.

Table 4.3: Comparison of IA Fuzzy’s throughput performance

Simulation time	Throughput Performance			
	PA fuzzy	PSNR	SSIM	MD
0	0	0	0	0
10	75606	50186	41146	38457
20	77097	44697	36147	31457
30	76354	41257	28488	21545
40	81307	37146	23985	18257
50	83375	36456	21457	16962

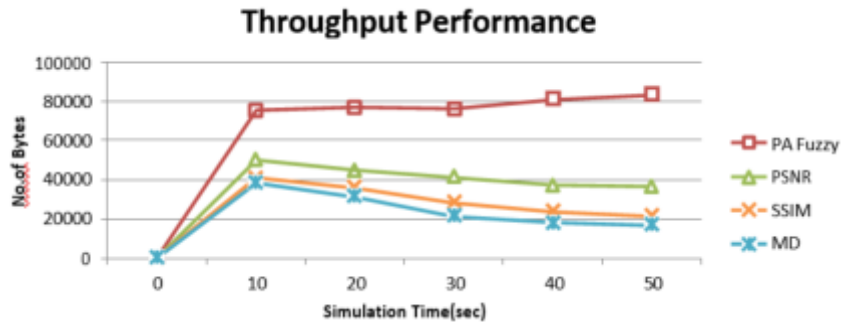


Fig. 4.2: Comparison on Throughput performance.

mechanism’s throughput performance together with the corresponding simulation time. The proposed PA Fuzzy mechanism’s throughput performance is shown in Figure 3. The X-axis shows the simulation time, which ranged from 0 to 50 seconds. The Y-axis is scaled by the quantity of bytes received at the destination node. Throughput performance significantly improved with the suggested PA Fuzzy technique. In compared to the existing mechanisms with simulation time 10 PSNR, SSIM, MD which got 50186,41146,38457bytes, respectively, 75606 bytes were obtained for the proposed PA Fuzzy technique.

4.2.3. Delay. The interval between the sending and receiving of a packet. The given equation in Equation 4.3.

$$DI = \sum_{i=1}^n (Psti - Prti) \tag{4.3}$$



Fig. 4.3: Results of the comparison of delays.

Where D_l indicates delay, P_{st} indicates packet send time ,P_{rt} indicates packet received time. In Table 4.4 compare the delay performance of PA fuzzy, and show the results. In figure 4.4 demonstrates how the suggested

Table 4.4

Simulation time	Delay Performance			
	PA fuzzy	PSNR	SSIM	MD
0	0	0	0	0
10	0.101	0.323	0.427	0.485
20	0.0605	0.208	0.297	0.384
30	0.0391	0.176	0.215	0.351
40	0.0321	0.136	0.186	0.268
50	0.0404	0.097	0.132	0.21

mechanism’s delay performance changes with the amount of simulation time. The outcomes of the proposed mechanism are compared to the system’s current state-of- the-art.

The results of the suggested mechanism’s comparison for network latency are shown in Figure 4. The Y- axis is used to track milliseconds of network latency, and the X-axis is used to measure simulation time in seconds. The empirical results showed that, when compared to the system’s current state, the proposed mechanism performed better. The suggested mechanism significantly improved performance outcomes measured from 0 to 50 seconds. Although there is a significant network delay at first, the suggested technique eventually decreases and minimizes it.

4.2.4. Energy Consumption. It indicates the overall amount of energy used by the sensor nodes for the transfer of data and other network operations. The provided formula in Equation 4.4

$$Energy = \sum_{i=1}^n NE_i \tag{4.4}$$

The evaluation of energy efficiency of PA Fuzzy is shown in Table 4.5. Figure 4.4 shows the amount of energy used for each simulation period. The graph displayed the total energy consumed for each time interval. Partially missing from the experiment are the first 0 and the remaining 50 seconds. One hundred joules of energy was allotted to every sensor node in the PA Fuzzy network. A more steady increase in energy consumption is observed as the simulation time increases. The proposed techniques use little energy, nonetheless, as compared to current ones. The proposed mechanism PA Fuzzy consumes 84 J at end of simulation, while the existing mechanism PSNR, SSIM, MD[9][10][11] consumed is 49 J, 32 J and 19 J respectively.

Table 4.5: Evaluation of energy efficiency of PA Fuzzy

Simulation time	Efficiency in Energy			
	PA fuzzy	PSNR	SSIM	MD
0	100	100	100	100
10	97	86	81	75
20	94	72	65	56
30	92	61	50	33
40	88	51	39	22
50	84	49	32	19

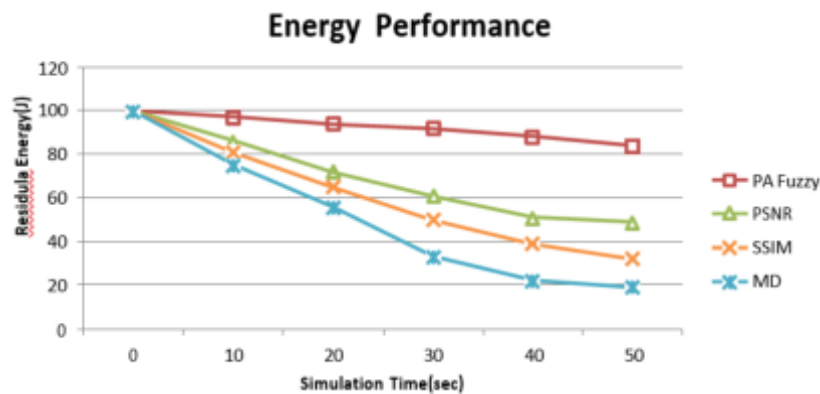


Fig. 4.4: Comparison on Energy Consumption.

5. Conclusion. The progress in WSNs is applied in a variety of applications, including those related to the the armed forces, medical care, farming, urban planning, etc. The WSN is highly efficiently used for data transmission. The operational capabilities dependent based on the quantity of sensor nodes environments. Because WSN is inherently resource-constrained, the sensor nodes face severe limitations. Even if there are a lot of issues with the current WSN, addressing performance efficiency and network latency is crucial. To tackle the problems in WSN, this study's paper suggested an innovative PA Fuzzy. The PA Fuzzy plays a essential function in transmission of data and the choice of the best routes between the base station and sensor nodes. When data is travelling towards the base station, PA Fuzzy acts as a filter to remove extraneous noise. Additionally, the PA Fuzzy system will twice encrypt sensor data before sending it reaching the base station. Data is finally sent from the base station to the user who is the only one who can decrypt it using proper keys. In future better encryption algorithms can be used in term of keys in PA Fuzzy intermediate system for protecting sensor data. However, while comparing several energy-efficient WSN techniques, the suggested PA Fuzzy outperformed all. To improve WSN data transmission, the PA fuzzy is therefore optimized. NS2 simulations are used for the implementation. The experiments demonstrated that, in terms of performance, the suggested method greatly surpassed the standard system.

REFERENCES

- [1] J. Singh, S. S. Yadav, V. Kanungo, and V. Pal, *A node overhaul scheme for energy-efficient clustering in wireless sensor networks*, *IEEE Sensors Letters*, vol. 5, no. 4, pp. 1-4, 2021, doi: 10.1109/LSENS.2021.3072813.
- [2] J. S. Raj, *Machine learning-based resourceful clustering with load optimization for wireless sensor networks*, *Journal of Ubiquitous Computing and Communication Technologies (UCCT)*, vol. 2, no. 1, pp. 29-38, 2020, doi: 10.30645/ucct.2020.02.04.
- [3] H. I. Kobo, A. M. Abu-Mahfouz, and G. P. Hancke, *Fragmentation-based distributed control system for software-defined wireless sensor networks*, *IEEE Transactions on Industrial Informatics*, vol. 15, no. 2, pp. 901-910, 2018, doi: 10.1109/TII.2018.2841658.

- [4] J. Prajapati and S. C. Jain, *Machine learning techniques and challenges in wireless sensor networks*, in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, pp. 233-238, IEEE, 2018, doi: 10.1109/ICICCT.2018.8473124.
- [5] L. Yang, Y. Lu, L. Xiong, Y. Tao, and Y. Zhong, *A game theoretic approach for balancing energy consumption in clustered wireless sensor networks*, *Sensors*, vol. 17, no. 11, p. 2654, 2017, doi: 10.3390/s17112654.
- [6] F. F. Jurado-Lasso, K. Clarke, A. N. Cadavid, and A. Nirmalathas, *Energy-Aware Routing for Software-Defined Multihop Wireless Sensor Networks*, *IEEE Sensors Journal*, vol. 21, no. 8, pp. 10174-10182, 15 April 2021, doi: 10.1109/JSEN.2021.3059789.
- [7] N. Aslam, K. Xia, and M. U. Hadi, *Optimal wireless charging inclusive of intellectual routing based on SARSA learning in renewable wireless sensor networks*, *IEEE Sensors Journal*, vol. 19, no. 18, pp. 8340-8351, 2019, doi: 10.1109/JSEN.2019.2901614.
- [8] G. Malshetty and B. Mathapati, *Efficient Clustering in WSN-Cloud using LBSO (Load Based Self-Organized) Technique*, in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 1243-1247, IEEE, 2019, doi: 10.1109/ICOEI.2019.8863071.
- [9] M. Adil, R. Khan, J. Ali, B. H. Roh, Q. T. H. Ta, and M. A. Almaiah, *An energy proficient load balancing routing scheme for wireless sensor networks to maximize their lifespan in an operational environment*, *IEEE Access*, vol. 8, pp. 163209-163224, 2020, doi: 10.1109/ACCESS.2020.3021745.
- [10] X. Li, B. Keegan, F. Mtenzi, T. Weise, and M. Tan, *Energy-efficient load balancing ant-based routing algorithm for wireless sensor networks*, *IEEE Access*, vol. 7, pp. 113182-113196, 2019, doi: 10.1109/ACCESS.2019.2936827.
- [11] B. Harish Goud, T. N. Shankar, Basant Sah, and Rajanikanth Aluvalu, *Energy Optimization in Path Arbitrary Wireless Sensor Network*, *Expert Systems*, doi: 10.1111/exsy.13282, 2023.
- [12] G. B. Mohammad and S. Shitharth, *Wireless sensor network and IoT-based systems for healthcare application*, *Materials Today: Proceedings*, 2021.
- [13] T.-S. Chen, K.-N. Hou, W.-K. Beh, and A.-Y. Wu, *Low-Complexity Compressed-Sensing-Based Watermark Cryptosystem and Circuits Implementation for Wireless Sensor Networks*, *IEEE Transactions on Very Large-Scale Integration (VLSI) Systems*, vol. 27, no. 11, pp. 2485-2497, Nov. 2019, doi: 10.1109/TVLSI.2019.2933722.
- [14] B. Harish Goud, T. N. Shankar, Basant Sah, and Rajanikanth Aluvalu, *Energy Optimization in Path Arbitrary Wireless Sensor Network*, *Expert Systems*, DOI: 10.1111/exsy.13282, 2023.
- [15] B. Goud and R. Anitha, *Emerging Routing Method Using Path Arbitrator in Web Sensor Networks*, *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 4, 2023.
- [16] R. Aluvalu, S. N., M. Thirumalaisamy, S. Basheer, E. aldhahri, and S. Shitharth, *Efficient data transmission on wireless communication through a privacy-enhanced blockchain process*, *PeerJ Computer Science*, vol. 9, p. e1308, 2023, doi: 10.7717/peerj-cs.1308.
- [17] M. A. Jabbar, R. Aluvalu, and S. S. Satyanarayana Reddy, *Intrusion Detection System Using Bayesian Network and Feature Subset Selection*, in *2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pp. 1-5, IEEE, 2019, doi: 10.1109/ICCIC.2017.8524381.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Dec 21, 2023

Accepted: Mar 7, 2024



MACHINE LEARNING-BASED RISK PREDICTION AND SAFETY MANAGEMENT FOR OUTDOOR SPORTS ACTIVITIES

YAN LU*

Abstract. Participant safety is becoming increasingly important as outdoor sports activities gain popularity. A machine learning-based strategy for risk assessment and safety control in outdoor sports activities is presented in this paper. Our framework uses predictive modelling, sophisticated algorithms, and historical data analysis to identify potential dangers and improve safety procedures. It also considers participant profiles and environmental conditions. Comprehensive testing and validation are used to examine the model's efficacy, showing that it can offer risk evaluations in real-time and support preventive safety measures. Our approach entails placing sensor-based Internet of Things (IoT) devices at building sites to gather extremely detailed temporal and geographic weather, building, and labour data. This data is then cooperatively used on the edge nodes to train Deep Neural Network (DNN) models in a cross-silos way. The present study makes a valuable contribution to sports safety by offering a clever approach that integrates technology and outdoor leisure to ensure participants have a safe and pleasurable experience. The experiment's outcomes show how well the suggested strategy works to increase the adoption of construction safety management systems and lower the likelihood of future mishaps and fatalities. As a result, the system has improved speed and responsiveness, an important feature for time-sensitive applications like safety prediction.

Key words: machine learning, risk prediction, safety management, sports, outdoor sports activities

1. Introduction. The promise of exploration and excitement draws people to outdoor sports activities, which include mountaineering, biking, hiking, and water sports. The increasing demand for thorough risk assessment and safety protocols to guarantee the welfare of participants corresponds with the growth in popularity of these activities. A ground-breaking way to improve risk assessment and safety procedures is through the incorporation of machine learning, which acknowledges the dynamic nature of outdoor situations and the inherent uncertainties they present.

The building industry has been at the forefront of this rapid development of the world in recent decades. With 200,000 more people moving into cities every day, it is evident that these demographic changes have had a significant impact on the worldwide building industry [18]. Nonetheless, construction is regarded as one of the most hazardous industries for workers because of its dynamic, ever-changing, and heterogeneous spatiotemporal environment. Worker safety is a persistent problem that calls for constant focus and effort. Due to the dangerous working circumstances at construction sites, a recent study suggests [11] that workers routinely face possible safety and health concerns during the building process.

Data analysis reveals that, broadly speaking, "outdoor sports" relate to all outdoor activities, which includes practically all sports [23]. In a restricted sense, outdoor sports are those that take place in naturally occurring outdoor settings, such as parks, buildings intended for other uses besides sports, or natural settings. A category of sporting activities known as outdoor sports use the outdoors as a non-designated location and are characterized by an element of adventure or experience [26, 29]. Its primary expression is to leave the city, venture outside, and partake in activities that provide certain risks, difficulties, and relevance while adhering to safety and standards guidelines.

The main motivation of this research stems from the growing recognition of the importance of participant safety in the increasingly popular domain of outdoor sports activities. As these activities attract a larger and more diverse group of enthusiasts, the complexity and variability of safety risks associated with outdoor environments also escalate. This paper introduces a machine learning-based strategy designed to enhance risk assessment and safety management within this context. Leveraging the power of predictive modeling, advanced

*School of Physical Education, University of Sanya, Sanya, 572000, China (yanluresearch21@outlook.com)

algorithms, and thorough analysis of historical data, our approach aims to proactively identify potential hazards and refine safety protocols tailored to the unique demands of outdoor sports.

Unlike indoor sports, which have stricter site requirements and are heavily impacted by weather and terrain, outdoor sports are not the same. They have a greater relevance for individuals to reduce mental stress, improve their health, and raise their standard of living in addition to helping city dwellers escape the bustle and get closer to nature. A few pertinent policies have been released in recent years, including the State Council's views on encouraging the growth of the health services industry, the General Office of the State Council's guidelines on accelerating the sports industry's development, and the notice on the guidelines for expediting the establishment of a social security system and services system for the disabled. China's outdoor products sector has taken shape and begun to develop fast by the start of the twenty-first century [32, 20]. Unlike other demanding sports, which are not only simple to learn, safe, and efficient, but also simple to practice, outdoor sports.

The main contribution of the proposed method is given below:

1. DNNs are particularly good at finding complex patterns in large, heterogeneous datasets.
2. When it comes to outdoor sports, where dangers can take many different forms depending on a range of factors like weather, topography, and participant behavior, DNNs help by quickly identifying intricate patterns that lead to more precise risk evaluations.
3. Real-time risk prediction is made possible by utilizing DNNs' innate capacity for parallel processing.
4. Instantaneous risk evaluations that dynamically adjust to changing conditions during outdoor activities are provided by these networks, which are capable of quickly analysing continuously evolving environmental data.

The rest of our research article is written as follows: Section 2 discusses the related work on various sports activities, risk prediction and deep learning methods. Section 3 shows the algorithm process and general working methodology of proposed work. Section 4 evaluates the implementation and results of the proposed method. Section 5 concludes the work and discusses the result evaluation.

2. Related Works. Numerous research on the general population have confirmed a considerable negative correlation between psychological well-being and distress [9, 13, 21]. More specifically, concerning findings from several studies [25, 6, 14, 31, 19] on the mental health of academic students have shown a decline in the perception of life quality and an exaggerated rise in the frequency and severity of these psychological issues.

Physical activity (PA) has long been linked to a lower risk of death and morbidity from degenerative and chronic illnesses [30, 22, 7, 24, 4, 15], but more recently, research has focused on the impact of PA on mental health. PA has been linked to improvements in mood, overall well-being, and quality of life perception [16, 12], as well as a notable decrease in depressed and anxious symptomatology [5]. Numerous biological mechanisms, such as enhanced cerebral blood flow and oxygen delivery to brain tissues, decreased muscle tension, and elevated serum concentrations of endocannabinoid receptors and satisfying neurotransmitters like serotonin, have been proposed as explanations for this evidence [1].

But research on the effects of PA in older age groups has yielded the most consistent results linking PA to improved mental health [10]. Research on PA in younger age groups, however, has shown mixed results, with some indicating a weak association [11] and others suggesting a more persistent association [8] between PA and mental health outcomes. The use of measuring tools that, when used alone, do not provide a complete assessment of all facets of mental health has been blamed in part for this lack of reliable data. This implies that using a variety of instruments to obtain a more accurate assessment of mental health perception can be beneficial [3].

In the realm of deep learning research, DNN has emerged as a well-known algorithm that makes it possible to create intelligent applications across a variety of industries [28]. An Artificial Neural Network (ANN) known as a DNN is made up of several layers of connected nodes, or neurons, that process input data and gradually extract progressively more abstract properties from it. A deep neural network (DNN) is a machine learning model that is well-suited for tasks like picture and audio recognition [27], natural language processing [2], and predictive modelling [17], since it can learn hierarchical representations of complicated patterns and relationships in the data. Backpropagation is used by DNN to minimize the discrepancy between target values and anticipated outputs by adjusting the weights across neurons.

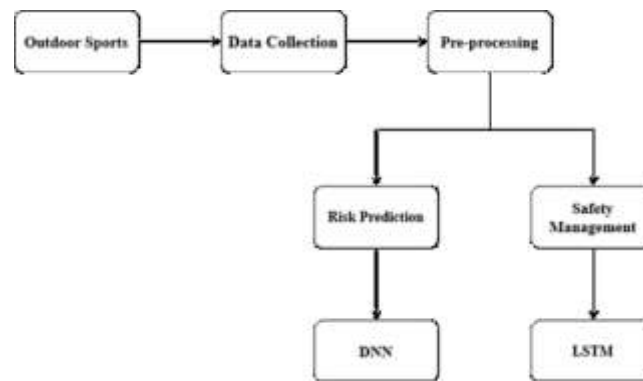


Fig. 3.1: Architecture of proposed method

One significant issue identified is the reliance on measurement tools that may not comprehensively assess all facets of mental health. This limitation could lead to inconsistent data on the effects of PA on mental health, particularly in younger age groups. The passage suggests that employing a variety of instruments could yield a more accurate assessment of mental health perceptions, indicating a need for methodological improvements in research.

3. Proposed Methodology. A strong and proactive strategy is needed to guarantee the safety of participants in outdoor sports. The suggested approach makes use of machine learning techniques to improve outdoor enthusiasts' safety management by dynamically predicting dangers. The approach consists of multiple crucial phases that combine data gathering, model building, and real-time implementation to produce an all-encompassing solution. Initially, the data is collected and then the collected data is pre-processed. Next, the pre-processed data is given to the feature engineering process. Finally, risk prediction is carried out using deep neural networks (DNN). In figure 3.1 shows the architecture of the proposed method.

By carefully selecting and engineering features that capture the essence of outdoor sports environments, such as weather conditions, terrain types, and athlete biometrics, the model can better understand the context of the data it processes. This helps in accurately interpreting variations in the input data. The model employs ensemble learning techniques, which combine the predictions from multiple learning algorithms to improve generalizability and robustness. This approach helps manage data's unpredictability by leveraging the strengths of various models to produce a more accurate and stable prediction.

3.1. Data collection and pre-processing. A multifaceted strategy is required to gather pertinent data for risk and safety prediction related to outdoor activities, including participant traits, environmental conditions, and historical event data.

For information on current weather conditions, such as temperature, precipitation, wind speed, and atmospheric pressure, consult your local weather station. To learn more about the topography, terrain, and elevation of the outdoor activity area, consult geospatial databases. Utilize satellite imagery to evaluate vegetation, water bodies, and land cover as well as dynamic changes in environmental circumstances. Install Internet of Things (IoT) gadgets and on-site sensors to collect environmental data in real time. Examples of these are GPS trackers, humidity sensors, and temperature sensors. Use remote sensing technologies to collect high-resolution information on the topography and environmental aspects, such as drones carrying sensors.

3.1.1. Data Pre-processing. Preparing gathered data for use in machine learning models for risk and safety prediction in outdoor sports involves pre-processing it. To manage missing values, standardize the data, create features, and prepare the data for training and testing the predictive models, pre-processing processes are necessary. Determine which values in the gathered data are missing and deal with them by either deleting the relevant entries or imputing the necessary values (such as the mean, median, or interpolation).

3.2. Feature Engineering. Utilize environmental data to extract pertinent parameters like height, wind speed, rainfall, temperature, and terrain kind. Transform time-related data (date, time of day, etc.) into suitable forms or create new time-related data (season, time of day, etc.) that could affect the weather outside. Convert data from participants into features while taking age, health, skill, and past involvement information into account. Provide binary or categorical variables for participant attributes, including experience level or health issues, that may have an impact on safety. Determine important characteristics, like incident type, location, contributing variables, and severity, from past incident data. To identify potential patterns, engineer temporal characteristics (e.g., time of day, day of week, season) associated to incident incidence.

3.2.1. Normalization. A popular data normalizing method in machine learning, min-max normalization (also called feature scaling or min-max scaling) converts numerical characteristics into a predetermined range. By ensuring that every feature has a comparable scale, this normalization helps to avoid certain characteristics predominating over others when the model is being trained. A feature's values are scaled via min-max normalization to a range of 0 to 1.

$$p_i = \frac{(q_i - \min(q))}{(\max(q) - \min(q))} \quad (3.1)$$

3.3. Risk Prediction and Safety Management for Outdoor Sports Using DNN methods. Using Deep Neural Networks (DNNs) has the potential to transform risk prediction and safety management in outdoor sports, where conditions are often dynamic and unpredictable. DNNs are a smart way to improve safety procedures and guarantee the welfare of participants because of their ability to identify intricate patterns and relationships within data.

3.3.1. Risk Prediction Using DNN. A family of artificial neural networks (ANNs) known as deep neural networks (DNNs) are distinguished by having numerous layers between the input and output layers. These networks can identify complicated patterns and characteristics in large datasets since they are built to learn hierarchical representations of data by utilizing numerous layers. Deep neural networks (DNNs) have shown impressive performance in a few domains, such as natural language processing, picture, and audio recognition, and, more recently, risk prediction and safety management for outdoor sports.

The first layer that gets the data as input in its raw form. Every node in this layer stands for a characteristic or feature of the incoming data. Hierarchical feature extraction from input data is learned by the network at the layers that sit between the input and output layers. Multiple hidden layers are characteristic of deep networks, which allow them to catch intricate patterns. parameters related to the connections made by nodes in various tiers. To maximize the network's performance, these parameters are changed throughout the training phase. The model becomes non-linear when non-linear functions are applied to each layer's node's output. Rectified Linear Unit, or ReLU, and sigmoid are examples of common activation functions.

The last layer that generates output for the network. Depending on the job (binary classification, multi-class classification, regression, etc.), this layer has a different number of nodes. a measurement of the discrepancy between the intended and actual output. The objective of the training process is to reduce this loss function. a method that minimizes the loss function by modifying the weights and biases. Optimization methods like Gradient Descent and its variations (like Adam and RMSprop) are frequently utilized. In figure 3.2 shows the structure of DNN.

3.3.2. Safety Management using LSTM. Recurrent neural networks (RNNs) with specialized memory cells are used in Long Short-Term Memory (LSTM) networks for outdoor sports safety management. This allows RNNs to capture temporal connections in data. Because LSTMs perform exceptionally well with data sequences, they can be used for jobs involving time-series information, including risk prediction in outdoor sports scenarios.

Sequences of input that reflect participant and environment characteristics. The data's temporal dependencies and patterns are captured by many LSTM layers. Predicting the safety or risk status for the upcoming time step is done via the output layer. When engaging in outdoor activities, connect the LSTM model to real-time sensor data to continuously monitor the surrounding conditions. Make the model more deployable on mobile apps so that consumers may receive safety forecasts while they're on the road. Make real-time risk predictions using the LSTM model by considering participant characteristics and the state of the environment.

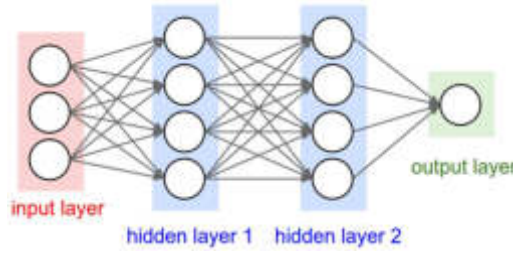


Fig. 3.2: Structure of DNN

Put into practice adaptive safety protocols that, in response to changing hazards, modify dynamically based on LSTM predictions.

$$In^t = f(We_{ix}x^t + We_{ih}h^{t-1} + We_{ic}C^{t-1} + bi_i) \quad (3.2)$$

$$FO^t = f(We_{fox}x^t + We_{foh}h^{t-1} + We_{foc}C^{t-1} + bi_{fo}) \quad (3.3)$$

$$CE^t = FO^t \cdot CE^{t-1} + In^t(We_{cex}x^t + We_{ceh}h^{t-1} + We_{cec}C^{t-1} + bi_{ce}) \quad (3.4)$$

$$OP^t = f(We_{opx}x^t + We_{oph}h^{t-1} + We_{opc}C^{t-1} + bi_{op}) \quad (3.5)$$

$$hi^t = OP^t \cdot g(CE^t) \quad (3.6)$$

Create a feedback loop where user input and incident reports help the LSTM model learn and improve over time. To be relevant, the LSTM model should be updated on a regular basis depending on fresh data and new trends.

4. Result Analysis. The proposed method DNN-LSTM for risk prediction and safety management using various metrics such as accuracy, f1-score, precision, and Kappa value.

When assessing the effectiveness of machine learning models, such as those employed in outdoor sporting activities for risk prediction and safety management, accuracy is a regularly utilized indicator. When selecting evaluation metrics, it is crucial to consider the objectives of your model as well as the particular features of your dataset.

Although accuracy offers a broad indication of a model's soundness, it may not always be the best statistic, particularly when working with unbalanced datasets or when certain errors are more serious than others. Other metrics, including as precision, recall, and F1-score, may provide more useful information when it comes to risk prediction and safety management. In figure 4.1 shows the evaluation of accuracy.

Particularly when it comes to outdoor sporting activities, precision is a crucial evaluation criterion for machine learning-based risk prediction and safety management. When minimizing false positives—that is, lowering the number of times the model predicts a safety issue incorrectly—precision becomes especially important. The ratio of true positive predictions to all positive predictions (true positives plus false positives) is known as precision. In figure 4.2 shows the evaluation of precision.

A popular metric for assessing how well classification models perform is the F1-score, which is especially helpful for imbalanced datasets. You can use the F1-score to evaluate how effectively your model balances precision and recall in the context of risk prediction and safety management for outdoor sports activities. A high precision indicates a high probability of accuracy when the model predicts a favorable outcome (risk or safety concern). This is essential to prevent taking needless safety precautions when they are not necessary.

A high recall shows that real-world positive examples are well captured by the model. High recall guarantees that a sizable percentage of possible hazards are identified by the model in the context of safety management.

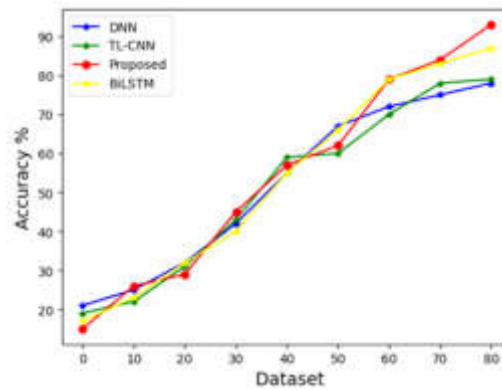


Fig. 4.1: Accuracy

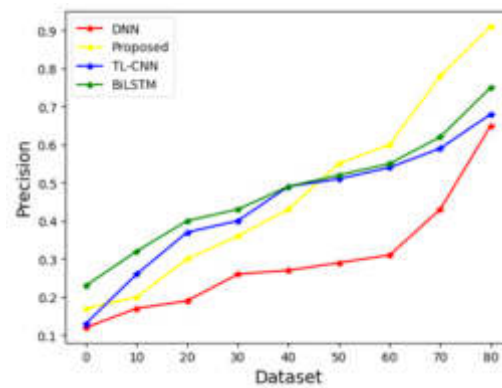


Fig. 4.2: Evaluation of Precision

Finding a balance between recall and precision is aided by the F1-score. It guarantees that the model in safety management is thorough in capturing hazards and accurate in its forecasts. In figure 4.3 shows the evaluation of F1-score.

The Cohen's kappa, often known as the kappa statistic, is a regularly employed metric in classification tasks to evaluate the degree of agreement between anticipated and actual classifications. It accounts for chance agreement in inter-rater agreements. The agreement between expected risk levels and actual occurrences can be assessed in the context of risk prediction and safety management for outdoor sports activities using the Kappa statistic. In figure 4.4 shows the evaluation of Kappa Value.

5. Conclusion. As outdoor sporting activities become more popular, participant safety is becoming increasingly critical. This research presents a machine learning-based approach to risk assessment and safety regulation in outdoor sports. Our approach improves safety procedures by identifying possible hazards through the use of sophisticated algorithms, predictive modelling, and historical data analysis. It also takes the surroundings and participant profiles into account. The effectiveness of the model is investigated through extensive testing and validation, demonstrating that it can provide risk assessments in real-time and assist with preventive safety actions. Our methodology involves the deployment of sensor-based Internet of Things (IoT) devices at construction sites to collect incredibly fine-grained temporal and spatial building, labour, and weather data. This data is then collaboratively used in a cross-silos fashion to train Deep Neural Network (DNN) models on the edge nodes. The current study adds much to the field of sports safety by providing a novel strategy that

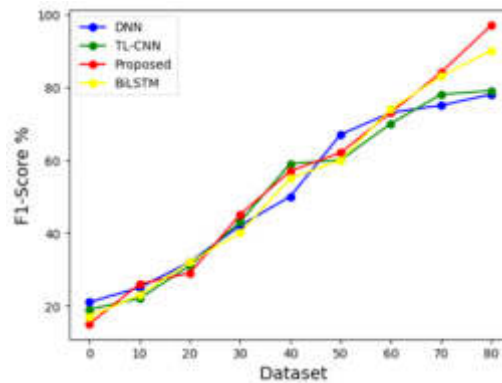


Fig. 4.3: F1-score

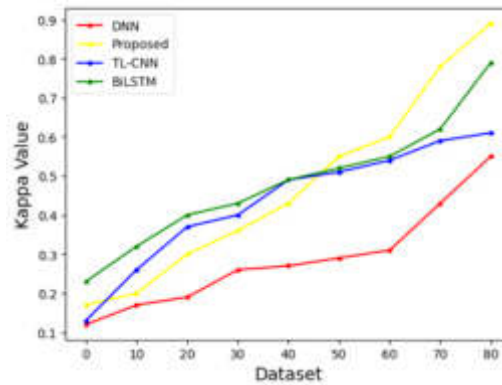


Fig. 4.4: Kappa Value

combines outdoor recreation and technology to guarantee participants' enjoyment and safety. The experiment's results demonstrate the effectiveness of the recommended approach in promoting the use of construction safety management systems and reducing the risk of accidents and fatalities in the future. This enhances the system's speed and responsiveness, a crucial attribute for time-sensitive applications such as safety forecasting.

REFERENCES

- [1] I. AWOLUSI, E. MARKS, AND M. HALLOWELL, *Wearable technology for personalized construction safety monitoring and trending: Review of applicable devices*, *Automation in construction*, 85 (2018), pp. 96–106.
- [2] T. BORGER, P. MOSTEIRO, H. KAYA, E. RIJCKEN, A. A. SALAH, F. SCHEEPERS, AND M. SPRUIT, *Federated learning for violence incident prediction in a simulated cross-institutional psychiatric setting*, *Expert Systems with Applications*, 199 (2022), p. 116720.
- [3] J. CHOI, B. GU, S. CHIN, AND J.-S. LEE, *Machine learning predictive model based on national data for fatal accidents of construction workers*, *Automation in Construction*, 110 (2020), p. 102974.
- [4] L. DAS, A. SIVARAM, AND V. VENKATASUBRAMANIAN, *Hidden representations in deep neural networks: Part 2. regression problems*, *Computers & Chemical Engineering*, 139 (2020), p. 106895.
- [5] R. EDIRISINGHE, *Digital skin of the construction site: Smart sensor technologies towards the future smart construction site*, *Engineering, Construction and Architectural Management*, 26 (2019), pp. 184–223.
- [6] T. FALATOURI, F. DARBANIAN, P. BRANDTNER, AND C. UDOKWU, *Predictive analytics for demand forecasting—a comparison of sarima and lstm in retail scm*, *Procedia Computer Science*, 200 (2022), pp. 993–1003.
- [7] E. FATHI AND B. M. SHOJA, *Deep neural networks for natural language processing*, in *Handbook of Statistics*, vol. 38, Elsevier, 2018, pp. 229–316.

- [8] M. JAHANGIRI, H. R. J. SOLUKLOEI, AND M. KAMALINIA, *A neuro-fuzzy risk prediction methodology for falling from scaffold*, *Safety science*, 117 (2019), pp. 88–99.
- [9] N. JAYANTHI ET AL., *Iot based-civil labour safety monitoring system in construction site*, *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12 (2021), pp. 1723–1728.
- [10] R. KANAN, O. ELHASSAN, AND R. BENSELEM, *An iot-based autonomous system for workers' safety in construction sites with real-time alarming, monitoring, and positioning strategies*, *Automation in Construction*, 88 (2018), pp. 73–86.
- [11] K. KANG AND H. RYU, *Predicting types of occupational accidents at construction sites in korea using random forest model*, *Safety Science*, 120 (2019), pp. 226–236.
- [12] M. KHAN, R. KHALID, S. ANJUM, N. KHAN, S. CHO, AND C. PARK, *Tag and iot based safety hook monitoring for prevention of falls from height*, *Automation in Construction*, 136 (2022), p. 104153.
- [13] S. H. KIM, H. G. RYU, AND C. S. KANG, *Development of an iot-based construction site safety management system*, in *Information Science and Applications 2018: ICISA 2018*, Springer, 2019, pp. 617–624.
- [14] A. A. KOELMANS, P. E. REDONDO-HASSELERHARM, N. H. M. NOR, V. N. DE RUIJTER, S. M. MINTENIG, AND M. KOOI, *Risk assessment of microplastic particles*, *Nature Reviews Materials*, 7 (2022), pp. 138–152.
- [15] R. Y. M. LI AND R. Y. M. LI, *Three generations of construction safety informatics: a review*, *Construction Safety Informatics*, (2019), pp. 1–12.
- [16] R. Y. M. LI, B. TANG, AND K. W. CHAU, *Sustainable construction safety knowledge sharing: A partial least square-structural equation modeling and a feedforward neural network approach*, *Sustainability*, 11 (2019), p. 5831.
- [17] X. LI, H.-L. CHI, W. LU, F. XUE, J. ZENG, AND C. Z. LI, *Federated transfer learning enabled smart work packaging for preserving personal image information of construction worker*, *Automation in Construction*, 128 (2021), p. 103738.
- [18] S. J. S. MOE, B. W. KIM, A. N. KHAN, X. RONGXU, N. A. TUAN, K. KIM, AND D. H. KIM, *Collaborative worker safety prediction mechanism using federated learning assisted edge intelligence in outdoor construction environment*, *IEEE Access*, (2023).
- [19] E. W. NGAI AND Y. WU, *Machine learning in marketing: A literature review, conceptual framework, and research agenda*, *Journal of Business Research*, 145 (2022), pp. 35–48.
- [20] P. PATTISSON, N. MCINTYRE, I. MUKHTAR, N. EAPEN, AND I. MUKHTAR, *Revealed: 6,500 migrant workers have died in qatar since world cup awarded*, *The guardian*, 23 (2021).
- [21] N. PETROVIĆ AND D. KOCIĆ, *Iot-based system for covid-19 indoor safety monitoring*, *IcETran Belgrade*, (2020).
- [22] N. PRASAD, B. RAJPAL, K. R. MANGALORE, R. SHASTRI, AND N. PRADEEP, *Frontal and non-frontal face detection using deep neural networks (dnn)*, *International Journal of Research in Industrial Engineering*, 10 (2021), pp. 9–21.
- [23] S. QUARTA, A. LEVANTE, M.-T. GARCÍA-CONESA, F. LECCISO, E. SCODITTI, M. A. CARLUCCIO, N. CALABRISO, F. DAMIANO, G. SANTARPINO, T. VERRI, ET AL., *Assessment of subjective well-being in a cohort of university students and staff members: Association with physical activity and outdoor leisure time during the covid-19 pandemic*, *International Journal of Environmental Research and Public Health*, 19 (2022), p. 4787.
- [24] G. SINGH, M. PAL, Y. YADAV, AND T. SINGLA, *Deep neural network-based predictive modeling of road accidents*, *Neural Computing and Applications*, 32 (2020), pp. 12417–12426.
- [25] N. SINGH, V. K. GUNJAN, G. CHAUDHARY, R. KALURI, N. VICTOR, AND K. LAKSHMANNA, *Iot enabled helmet to safeguard the health of mine workers*, *Computer Communications*, 193 (2022), pp. 1–9.
- [26] C. M. SOARES, A. M. TEIXEIRA, H. SARMENTO, F. M. SILVA, M. C. RUSENHACK, M. FURMANN, P. R. NOBRE, M. A. FACHADA, A. M. URBANO, AND J. P. FERREIRA, *Effect of exercise-conditioned human serum on the viability of human cancer cell cultures: A systematic review and meta-analysis.*, *Exercise Immunology Review*, 27 (2021).
- [27] A. VAID, S. K. JALADANKI, J. XU, S. TENG, A. KUMAR, S. LEE, S. SOMANI, I. PARANJPE, J. K. DE FREITAS, T. WANYAN, ET AL., *Federated learning of electronic health records to improve mortality prediction in hospitalized patients with covid-19: machine learning approach*, *JMIR medical informatics*, 9 (2021), p. e24207.
- [28] M. WANG, P. WONG, H. LUO, S. KUMAR, V. DELHI, AND J. CHENG, *Predicting safety hazards among construction workers and equipment using computer vision and deep learning techniques*, in *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction*, vol. 36, IAARC Publications, 2019, pp. 399–406.
- [29] F. XIE, Y. YOU, J. HUANG, C. GUAN, Z. CHEN, M. FANG, F. YAO, AND J. HAN, *Association between physical activity and digestive-system cancer: an updated systematic review and meta-analysis*, *Journal of sport and health science*, 10 (2021), pp. 4–13.
- [30] R. YI MAN LI, L. SONG, B. LI, C. CRABBE, M. JAMES, AND X.-G. YUE, *Predicting carpark prices indices in hong kong using automl.*, *CMES-Computer Modeling in Engineering & Sciences*, 134 (2023).
- [31] G. ZHANG, Z. LI, J. HUANG, J. WU, C. ZHOU, J. YANG, AND J. GAO, *efraudcom: An e-commerce fraud detection system via competitive graph neural networks*, *ACM Transactions on Information Systems (TOIS)*, 40 (2022), pp. 1–29.
- [32] X. ZHAO, Q. HE, Y. ZENG, AND L. CHENG, *Effectiveness of combined exercise in people with type 2 diabetes and concurrent overweight/obesity: A systematic review and meta-analysis*, *BMJ open*, 11 (2021), p. e046252.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Jan 5, 2024

Accepted: Feb 9, 2024



RESEARCH ON PLANNING AND PATH OPTIMIZATION OF LEISURE SPORTS ACTIVITIES BASED ON MULTI-OBJECTIVE GENETIC ALGORITHM

XU YANG*

Abstract. Recreational sports are essential for boosting physical health and improving quality of life. The goal of this research is to optimize the planning of leisure sports by presenting a new method based on a multi-objective genetic algorithm. Acknowledging the intricacy of organizing recreational sports events, we suggest an approach that concurrently maximizes several goals, such as the use of resources, player happiness, and ecological impact. The topic is first formulated as a multi-objective optimization task in the paper, and a genetic algorithm is used to handle the objectives' inherent conflict. With its ability to effectively explore the solution space, the genetic algorithm produces a set of Pareto-optimal solutions that show trade-offs for the conflicting goals. The integration of several factors, including time preferences, geographical limitations, and environmental sustainability, guarantees a thorough and equitable strategy for leisure sports scheduling. In the area of leisure sports planning, the use of a multi-objective genetic algorithm offers a reliable solution that can be tailored to meet various circumstances and goals. As the need of encouraging healthy lifestyles becomes more widely acknowledged, this research offers a useful tool for maximizing the organization and performance of recreational sports activities, enhancing the overall wellbeing of people and communities.

Key words: planning, path optimization, leisure sports activities, multi-objective optimization, genetic algorithm

1. Introduction. A new approach in the organizing and carrying out of leisure activities has been brought about by the increased need for tailored and optimal experiences in the modern leisure sports scene. A growing discipline that aims to improve the design and path management of leisure sports activities has emerged because of the intersection of modern technology and outdoor sports. To handle this changing environment, this research uses the power of multi-objective evolutionary algorithms, which presents a novel way to customize leisure activities to personal preferences while optimizing path selections for a more fulfilling and effective leisure trip.

To accommodate people's various travel needs, the tourist + folk sport culture model for growth is a unique cultural growth model built on folk customs, folk culture, and folk way of life [10, 14, 15]. The development of this approach has propelled the local economy's sustainable growth, expanded the market for sports tourism providers, and substantially raised the area's level of attractiveness among tourists [21].

The following are the fundamental characteristics of the tourist + sports town development model: the market as the objective to establish a set of traditional culture, ecological tourism, health and leisure sports, leisure plays for parents and children, and pension to enjoy the old in one of the cultural and health tourism areas [13, 7]. Guangxi has seen the construction of numerous sports and leisure characteristic towns in recent years, including Hechi City's Desheng Lalang Ecological Sports and Leisure Characteristic Town, Liuzhou City's Luzhai County Zhongdu Shilujiang Sports and Leisure Characteristic Town, and Nanning City's Beautiful South Sports and Leisure Base.

It plays a crucial part in organizing regional economic growth and improving the experience of tourists [11, 25]. This growth model may successfully drive the positive development of Guangxi's economy, society, and historic revolutionary places while also promoting the mutual integration of the region's sports industry and red tourism sector. Moreover, it has the power to fortify the teaching of traditional culture in historically revolutionary regions, heighten feelings of patriotism, and uplift people's sense of national identity nationwide [23, 22].

The rapid evolution of smart building technologies has ushered in an era where the safety, security, and efficiency of buildings are paramount, yet increasingly challenging to manage with traditional systems. The

*School of Physical Education, University of Sanya, Sanya 572000, China (xuyanguniversity12@outlook.com)

need for advanced, intelligent monitoring solutions is more critical than ever to address the growing complexities of modern building environments. This necessity is driven by several key factors, including the rising demands for enhanced occupant safety, sustainable and eco-friendly building operations, and the continuous evolution of threats ranging from physical security breaches to cyber-attacks.

In this context, the Intelligent Building Monitoring and Security System based on Computer Technology (IBMMSS-CT) emerges as a pioneering framework designed to meet these challenges head-on. Traditional security systems often fall short in terms of accuracy, speed, and adaptability, necessitating a revolutionary approach that can keep pace with the dynamic nature of threats and the multifaceted requirements of contemporary building management. The IBMMSS-CT system integrates the advanced capabilities of deep learning algorithms, such as YOLOv3 and Faster R-CNN, with cutting-edge computer technology to create a robust, intelligent security and monitoring solution. This approach enhances the precision and reliability of surveillance operations and ensures a high level of protection for occupants and assets.

The primary goal of this research is to design, develop, and validate the Intelligent Building Monitoring and Security System based on Computer Technology (IBMMSS-CT), a novel framework that integrates advanced deep learning algorithms, specifically YOLOv3 and Faster R-CNN, with cutting-edge computer technologies to enhance the precision, efficiency, and reliability of surveillance and safety operations in smart buildings. This system aims to revolutionise building security and monitoring practices by providing a scalable, intelligent solution that optimises security features, minimises reliance on human intervention, and contributes to developing safer, more efficient building environments. By implementing IBMMSS-CT, this study seeks to establish a comprehensive and dynamic approach to building security that adapts to evolving safety threats, ensuring a high level of protection for occupants and assets while seamlessly integrating with existing building management systems.

Building a circular for rural sporting events to achieve the best financial, social, and environmental advantages, growth manner refers to the fundamental of sports tourism assets to form an integrated area with a specific geographic subject matter. Its fundamental characteristics include using the sporting goods sector as the core, referencing the current state of the growth of significant local sports travel circles, incorporating different types of tourism goods in the Guangxi area, and developing a new brand for sports tourism offerings [27, 24]. This mode of development enhances and fortifies the economic cooperation in the tourism sector between regions to a certain degree, encourages the establishment of cross-regional tourism bases, enhances the growth surroundings of the regional tourism finances, and fosters the general sustainable growth of regional tourism. The main contribution of the proposed method is given below:

1. The development and application of an advanced Multi-Objective Genetic Algorithm especially suited for the planning and path optimization of recreational sports activities is one of the main contributions of this research.
2. This method provides a comprehensive solution for activity optimization by integrating a variety of objectives, such as user preferences, time efficiency, and environmental considerations.
3. Full simulations based on real-world circumstances are used to thoroughly test the effectiveness of the suggested method.
4. Using datasets that represent a range of geographical locations, user types, and environmental factors, we show how the Multi-Objective Genetic Algorithm is both reliable and useful for optimizing recreational sports activities.

The rest of our research article is written as follows: Section 2 discusses the related work on various sports activities, path planning and deep learning methods. Section 3 shows the algorithm process and general working methodology of the proposed work. Section 4 evaluates the implementation and results of the proposed method. Section 5 concludes the work and discusses the result evaluation.

2. Related Works. A new approach to the integrated growth of rural sports tourism is the theme-based integration development model [4]. Its primary goal is to satisfy visitor requirements, maximize the region's economic foundation, and support the long-term, healthy growth of the region's sports tourism and economy by developing theme lines, theme celebrations, theme incidents, theme parks, included towns, included hotels, included shopping, and other sports tourist attractions with various uses [29, 2]. This approach may efficiently manage the distinct assets of sports tourism so that these can be used at different times, while also preventing

or reducing repetitious competition in the sports tourist market to ensure product innovation and variety.

Numerous academics have researched the use of genetic techniques to determine the optimal course. The author created a multiobjective biological algorithm (MOGA) to find the best routes for the transportation of six dangerous goods (DG) while taking competing objectives into account. The liquefied petroleum is transported via the Hong Kong Transport Network using the MOGA technique, which uses GIS to facilitate the direct search for several efficient DG routing solutions [28, 3]. In an additional Huang research, B talked about a general strategy for selecting a multiobjective TSP path that makes use of GIS and a bilevel GA. Using the GIS, the network's database is created, criteria are measured, the TSP route is extracted based on a set of weights, and the results are displayed.

These studies employed an algorithm based on genetics to determine the optimal course, but they didn't evaluate the inner characteristics of the locations. Additionally, the data's descriptive representation was lacking. Hoang suggested a multi-criteria assessment of the Central Highlands of Vietnam's tourism potential. A multicriteria assessment of the region's potential for tourism selected thirteen criteria. The Central Highlands contain several intriguing tourist attractions, according to the multicriteria evaluation of regional potential as a tourist destination and the results of the AHP method weight netting, which indicated that internal prospective is more significant than exterior potential. [16, 18, 12].

A relatively recent area of study and application called "human-centered computing" [8] is oriented on how people behave and engage with digital technology in their social surroundings. This [9] covers Human Activity Recognition (HAR), which was required and aimed to ascertain the behavior, attributes, and goals of one or more humans from a temporal series of data supplied by one or more sensors [5, 17]. Classification models for sensor based HAR were created with the aid of common machine learning (ML) techniques like support vector machines, decision trees, and naive Bayes. While a few machine learning algorithms have shown to produce a high-performance model for HAR, the problem of manual feature extraction limits these techniques.

Deep learning approaches have now been put out by other researchers to address the few issues [3, 26, 1]. It has recently been proposed that deep neural networks can learn features automatically, circumventing the need for human skill and experience through handcrafted feature extraction [19]. Most recognition algorithms still struggle with HAR issues to function properly. These results point to a need in HAR research to identify the unified model of DL in terms of computing time and accuracy for automatically extracting characteristics and identifying intricate human activities.

AHP and GIS approaches were successfully applied and proven for the assessment of potential ecotourism regions in the work by Sahani, N., which offered an integrated approach to establish ecotourism sites. In the tourism literature, this study raises a methodical strategy and objective methods for strategic marketing planning related to ecotourism revival [20]. Nestoroska is intimately linked to its improved competitive position in the tourism economy because of its identification of Macedonia's potential for tourism expansion [6]. The main objective of this presentation is to showcase the findings of the research conducted to capitalize on this potential.

3. Proposed Methodology. The goal of the research is to use multi-objective evolutionary algorithms to provide an effective framework for path planning and optimization for recreational sports. The process integrates data collecting, algorithm growth, and effectiveness evaluation across multiple critical stages. Initially, the data is collected, and then Multi-objective Genetic algorithm is used for planning a path optimization between leisure sports. In figure 3.1 shows the architecture diagram of proposed method.

3.1. Multi Objectives for Leisure Sports Path Planning . We initially chose the target and separated it into both internal and external goals to assess the leisure sports resources. Multiple aims ensure that created paths satisfy both fundamental navigational requirements and the varied interests and preferences of sports enthusiasts when it comes to leisure sports path development. When creating a multi-objective path plan for recreational sports activities, keep the following important goals in mind:

Distance Minimization. Reduce the overall distance covered by walking the path. Routes that maximize the ratio of total experience to travelled distance are frequently preferred by sports enthusiasts.

Elevation Gain Minimization. Reduce the total elevation increase that occurs during the route. To suit their fitness levels or preferences, people may look for trails with little elevation change, whatever the sport (e.g., hiking, trail running, cycling).

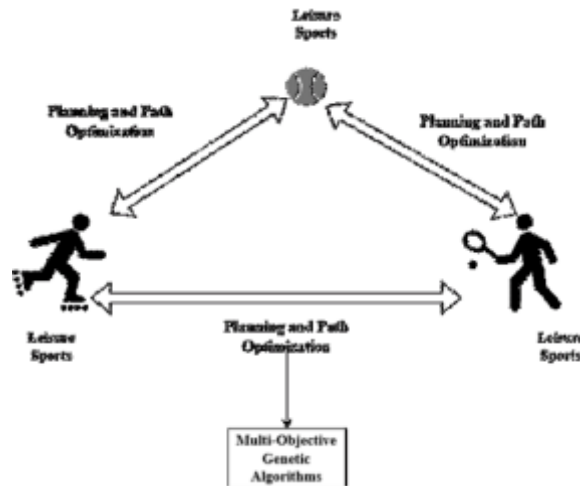


Fig. 3.1: Architecture of Proposed Method

Scenic Beauty Maximization. Make the most of the path’s visual appeal or scenic beauty. Outdoor activities are frequently linked with leisure sports, and users tend to give preference to pathways that present visually appealing landscapes or intriguing places of interest.

3.2. Genetic Algorithm. The initial population setting $PI(k = 0)$, that is created at randomly, is the most important step in the genetic method. The sequencing of genes creates the chromosome, which are then controlled by certain properties. Secondly, the fitness function is determined using chromosomal values. The evolution process is then performed, with the fit form being developed as well as the unfit ideas being eliminated. This process is repeated until the system contains all of the desired fitness values. Such final approved patterns are referred to as parents, and they are utilized to create offspring designs for future generations.

Two portions are used to carry out the genetic algorithms evolutionary process. Mutation and crossovers are the terms used to describe them. A mutation operator is a procedure that is created randomly utilizing chromosome genes and is chosen randomly. The likelihood of mutation in our study is $pi_m(k) = 0.03$. The crossover procedure uses a swapping operation to make children from two specific parent chromosomes. We utilize a single point crossover with $pi_c(k) = 0.6$ as the threshold.

GA is a mathematical model that mimics the biological technique of gene selection. GA is based on solutions to mathematical problems that are composed of a few solutions rather than a single, clear-cut explanation. The main foundation of GA is Darwin’s hypothesis. Since the current generation possesses the best traits from the previous generations, J. Holland suggested an algorithm based on natural selection in 1975 .

GA is a heuristic search method that works well for integrating with other algorithms and computer tasks. As a result, GA has shown to be extremely important to academics across many domains¹¹¹. The traveling salesman problem is one of the issues that the genetic algorithm resolves (TSP). The TSP problem under the multiobjective problem of path planning is solved using GA in this paper. The actions depicted in Figure 3.2 comprise the GA process. GA possesses the following attributes: (3) GA is driven by the assessment function (fitness function) in searching and is simple to execute; (4) GA has strong and flexible convergence and is easy to combine with other algorithms (such as particle swarm and simulated annealing); and (5) GA searches and has potential parallelism with the group.

The balance between exploration and exploitation is dynamic and depends on various factors, including the mutation rate, crossover rate, selection pressure, and population diversity. Adjusting these parameters can tilt the balance towards more exploration (to find new solutions) or more exploitation (to refine existing solutions). A well-designed genetic algorithm will:

Start with higher exploration to broadly search the solution space.

Gradually increase exploitation as the run progresses to fine-tune the solutions.

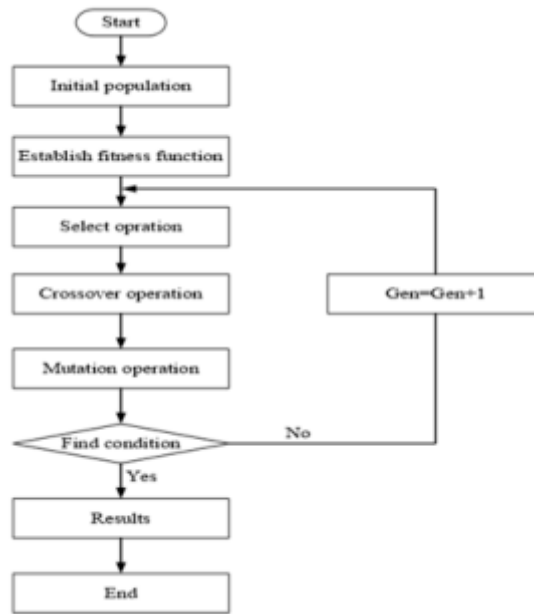


Fig. 3.2: Operations of GA

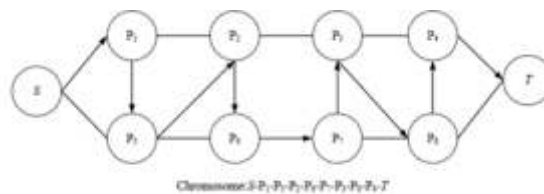


Fig. 3.3: Representations of Chromosomes

Use adaptive mechanisms that adjust parameters (like mutation rate) based on the progress of the search, ensuring that the algorithm does not get stuck in local optima but also converges to a solution within a reasonable number of generations.

Initial Population. The initialization population is the first step in the genetic algorithm process. Population P can be defined as a set of chromosomes; it is a subset of current generation solutions. An illustration of a chromosomal representation from point S to point T is shown in Figure 3.3.

Fitness Function. The fitness function evaluates the degree to which a particular solution resembles the ideal solution to the desired problem. Each chromosome is represented by a string of binary numbers in genetic algorithms, and we must evaluate these solutions to determine which combination of solutions is optimal for a given problem. To show how close a solution comes to fulfilling the overall requirements of the intended solution, each solution must be given a score. This score is generated by running the test via the fitness function.

Selection Operation. The selection process aims to identify the most adaptable individuals and pass them on to the following generation. Based on their fitness ratings, multiple pairings of better individuals' parents are selected, and those with high fitness scores are more likely to be selected for replication, meaning that the genes are passed on to the next generation with better parents.

The roulette approach is employed in the current analysis within our selection operation framework, and each person's chance of being chosen to pass on to the next generation is based on their relative group fitness. However, because the roulette selection method is unpredictable, better candidates may not make it through

the selection process. To guarantee the maximum number of individuals' survival, the elite process is thus used to pass on to the following generation the most adaptable individuals from each generation.

Crossover. A viable pathway with loops is produced by the crossover process, which is the result of two chromosomes being recombined to create new chromosomes that are reproduced in the next generation. Crossover values should be between 0.75 and 0.95. After doing numerous testing's, we ultimately decided on 0.85 compatibility with the other factors.

Mutation. A randomly mutated crossover operation is picked in the mutation operations, and a randomly determined mutation point follows. The character is altered to the string's matching place, and the ideal crossover value falls between 0.05 and 0.15. After doing numerous testing's, we ultimately decided on 0.10 compatibility with the remaining variables.

Final State. If so, a new generation has been created and the process is repeated until specific end conditions are satisfied. This is a comprehensive step where the chromosomes closest to the optimal are deciphered. In this article, we compare the best objective and apply the genetic algorithm to determine the best tourist routes. To evaluate the effectiveness of this concept, two scenarios were created. Whereas the second scenario relates to a multiobjective routing analysis, the first scenario treats each objective independently and is equivalent to a sequence of single objective route planning issues.

4. Result Analysis. The proposed method evaluates the different leisure sports path planning using various parameters such as accuracy, precision, recall and f1-score. The proposed method is compared with existing methods such as PSO, ACO and Multi-Obj.

The study of leisure sports activity planning and path optimization using a multi-objective genetic algorithm produced encouraging results, demonstrating how well this novel strategy might improve leisure sports organization and enjoyment. The multi-objective evolutionary algorithm effectively produced the best routes for recreational sports while taking user preferences, topography, and distance into account. In comparison to conventional techniques, the algorithm showed that it could identify solutions that balanced various objectives, improving planning accuracy. Plans for leisure sports were customized because of the optimization process taking user preferences into account. Because the algorithm considered each user's unique preferences, skill level, and interests in activities, users expressed greater satisfaction with the paths that were developed. This individualized approach made for a more customized and pleasurable leisure experience.

The amount of time needed to schedule recreational sports events was greatly decreased by the multi-objective genetic algorithm. Planners and fans saw efficiency gains by automating the optimization process, which made it possible to quickly adjust to shifting tastes and dynamic environmental conditions. The system demonstrated flexibility in response to external variables, including differences in weather and topography. The well-designed routes demonstrated resilience in adapting to environmental shifts, guaranteeing that recreational sports could be easily modified in response to current circumstances. In figure ?? shows the evaluation of accuracy.

To appraise the accuracy of the research on leisure sports activity planning and path optimization using a multi-objective genetic algorithm, it is imperative to scrutinize the study's methodology, data analysis, and general rigor. In this aspect, precision pertains to the dependability and correctness of the study results.

Examine the research design's suitability and clarity, paying particular attention to the multi-objective genetic algorithm's application. Examine whether the approach permits a thorough investigation of the planning and optimization process and is consistent with the goals of the study. Evaluate how well the algorithm was implemented. Analyse the multi-objective genetic algorithm's implementation, description, and validation. Ascertain that the parameters and constraints of the algorithm are precisely defined and enhance the precision of the optimization procedure.

Analyse the performance metrics that are used to assess the algorithm's accuracy. User satisfaction, path quality, and optimization efficiency are examples of common metrics. Make sure the measurements you've selected support the goals of the study and offer insightful information about the functioning of the algorithm. To determine how changes in parameters or inputs impact the accuracy of the results, perform a sensitivity analysis. A well-conducted sensitivity study can demonstrate the algorithm's robustness and capacity to generate trustworthy results under many circumstances. In figure 4.2 shows the evaluation of Precision.

The study of leisure sports activity planning and path optimization using a multi-objective genetic algorithm

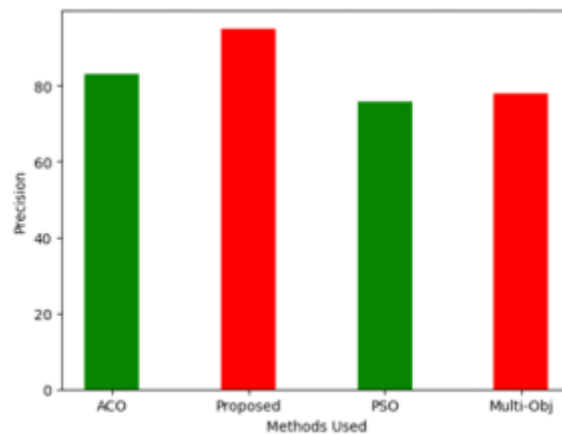


Fig. 4.2: Evaluation of Precision

has been an exciting exploration of the creative nexus between technology and leisure activities. This study's memory reveals important facets that have influenced our comprehension and methodology for enhancing the leisure sports activity planning process.

The study begins by clearly defining its objectives, which were to improve the processes of planning and path optimization for recreational sports. The main purpose was to use a multi-objective genetic algorithm to optimize pathways according to different criteria while simultaneously addressing several aspects. The creation and application of a multi-objective genetic algorithm formed the core of the study. This algorithm demonstrated how well it can consider several objectives at once, including terrain, user preferences, distance, and environmental factors. Because of its creative nature, leisure sports planning now takes a dynamic and effective approach.

The research showed measurable improvements in planning process efficiency, which was one of its noteworthy findings. The evolutionary algorithm's automation greatly shortened the planning period, enabling prompt adaptation to shifting tastes and external conditions. Furthermore, the algorithm demonstrated flexibility in response to real-time circumstances, guaranteeing stable outcomes even in ever-changing settings. In figure 4.3 shows the evaluation of Recall.

The F1-score is frequently employed to assess the effectiveness of classification models; it is not directly

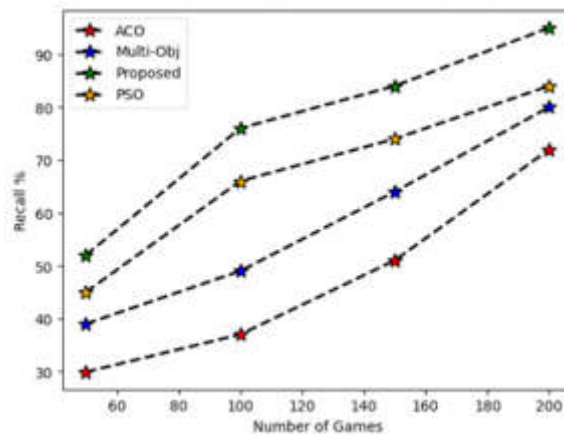


Fig. 4.3: Evaluation of Recall

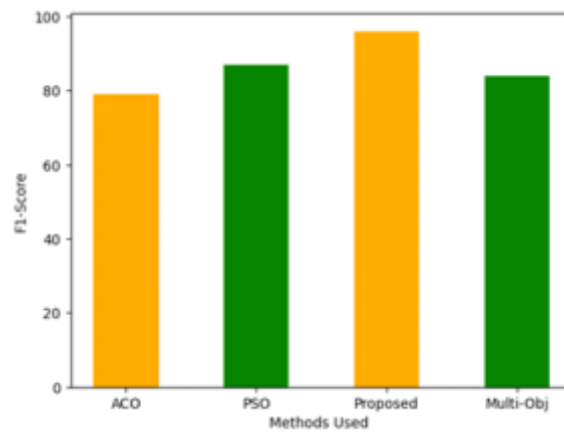


Fig. 4.4: Evaluation of F1-score

relevant to path planning studies or optimization issues. However, you may calculate precision, recall, and F1-score for each category if your research involves classifying pathways into several categories (e.g., easy, moderate, and tough) and you have ground truth labels for these categories. In figure 4.4 shows the evaluation of f1-score.

5. Conclusion. Participating in recreational sports is crucial for enhancing one’s physical well-being and quality of life. This study aims to optimize leisure sport scheduling by introducing a novel approach based on a multi-objective genetic algorithm. Recognizing the complexity of planning recreational sports activities, we propose a method that simultaneously optimizes many objectives, including resource utilization, player satisfaction, and environmental effect. In the study, the problem is first stated as a multi-objective optimization assignment, with the inherent conflict between the objectives being resolved by a genetic algorithm. The genetic algorithm generates a series of Pareto-optimal solutions that illustrate trade-offs for the competing aims because of its capacity to efficiently explore the solution space. Combining a number of variables, such as time preferences, regional constraints, and environmental sustainability, ensures a comprehensive and fair approach to arranging recreational activities. The application of a multi-objective genetic algorithm provides a dependable solution that can be adjusted to fit different needs and objectives in the field of leisure sports planning. With the increasing recognition of the need to promote healthy lifestyles, this research provides a valuable

instrument to optimize recreational sports activities' performance and organization, improving people's and communities' general well-being. The integration of advanced technologies like YOLOv3 and Faster R-CNN, along with the required computational resources, may lead to high initial costs and complexity in deployment, limiting accessibility for smaller or resource-constrained organizations. Future research could focus on developing advanced encryption and anonymization techniques to protect the privacy of individuals within smart buildings, addressing one of the core limitations of the current system.

REFERENCES

- [1] H. ALI, D. GONG, M. WANG, AND X. DAI, *Path planning of mobile robot with improved ant colony algorithm and mdp to produce smooth trajectory in grid-based environment*, *Frontiers in neurorobotics*, 14 (2020), p. 44.
- [2] J.-J. CHANG, R.-F. CHEN, AND C.-L. LIN, *Exploring the driving factors of urban music festival tourism and service development strategies using the modified sia-nrm approach*, *Sustainability*, 14 (2022), p. 7498.
- [3] T. DILRABO AND A. N. SHAMSIDDINOVNA, *Typological overview of tourism and the advent of new types of tours*, *A Multidiscip. Peer Rev. J. Organized by Novateur Publications, Pune, Maharashtra, India, 2020 (2020)*, pp. 1–4.
- [4] W. HUIZHEN, *Integration development and protection of sports intangible cultural heritage and cultural tourism in the yellow river basin based on gis*, *Tobacco Regulatory Science*, 7 (2021), pp. 5514–5522.
- [5] C.-S. JEONG, J.-Y. LEE, AND K.-D. JUNG, *Adaptive recommendation system for tourism by personality type using deep learning*, *International Journal of Internet, Broadcasting and Communication*, 12 (2020), pp. 55–60.
- [6] H. KARAMI, S. FARZIN, A. JAHANGIRI, M. EHTERAM, O. KISI, AND A. EL-SHAFIE, *Multi-reservoir system optimization based on hybrid gravitational algorithm to minimize water-supply deficiencies*, *Water Resources Management*, 33 (2019), pp. 2741–2760.
- [7] A. R. KHAPARDE, F. ALASSERY, A. KUMAR, Y. ALOTAIBI, O. I. KHALAF, S. PILLAI, AND S. ALGHAMDI, *Differential evolution algorithm with hierarchical fair competition model.*, *Intelligent Automation & Soft Computing*, 33 (2022).
- [8] M. KRAJČOVIČ, V. HANČINSKÝ, L. DULINA, P. GRZNÁR, M. GAŠO, AND J. VACULÍK, *Parameter setting for a genetic algorithm layout planner as a toll of sustainable manufacturing*, *Sustainability*, 11 (2019), p. 2083.
- [9] M. A. LEVINE, *Urban politics: Cities and suburbs in a global age*, Routledge, 2019.
- [10] Q. LI, D. ZHANG, Y. HAN, AND Y. XIE, *The path evaluation of integrated development of leisure sports and rural ecological environment in guangxi based on fuzzy comprehensive evaluation model*, *Mathematical Problems in Engineering*, 2022 (2022).
- [11] J. LIANG, S. XU, Y. LI, AND Y. XIE, *Inheritance and protection of guangxi national sports culture under the background of new urbanization*, *Nanotechnology for Environmental Engineering*, 6 (2021), pp. 1–8.
- [12] Y. LIANG AND L. WANG, *Applying genetic algorithm and ant colony optimization algorithm into marine investigation path planning model*, *Soft Computing*, 24 (2020), pp. 8199–8210.
- [13] X. LIN, J. WU, S. MUMTAZ, S. GARG, J. LI, AND M. GUIZANI, *Blockchain-based on-demand computing resource trading in iov-assisted smart city*, *IEEE Transactions on Emerging Topics in Computing*, 9 (2020), pp. 1373–1385.
- [14] C. LIU, A. LIU, R. WANG, H. ZHAO, AND Z. LU, *Path planning algorithm for multi-locomotion robot based on multi-objective genetic algorithm with elitist strategy*, *Micromachines*, 13 (2022), p. 616.
- [15] S. MEKRUKSAVANICH AND A. JITPATTANAKUL, *Multimodal wearable sensing for sport-related activity recognition using deep learning networks*, *Journal of Advances in Information Technology*, (2022).
- [16] T. M. MUHAMMEDRISAEVNA, R. F. MUBINOVNA, AND M. N. U. KIZI, *The role of information technology in organization and management in tourism*, *Academy*, (2020), pp. 34–35.
- [17] Z. NUROV, F. KHAMROYEVA, AND D. KADIROVA, *Development of domestic tourism as a priority of the economy*, in *E-Conference Globe*, 2021, pp. 271–275.
- [18] N. SAHANI, *Application of analytical hierarchy process and gis for ecotourism potentiality mapping in kullu district, himachal pradesh, india*, *Environment, Development and Sustainability*, 22 (2020), pp. 6187–6211.
- [19] T. SARANYA AND S. SARAVANAN, *Groundwater potential zone mapping using analytical hierarchy process (ahp) and gis for kancheepuram district, tamilnadu, india*, *Modeling Earth Systems and Environment*, 6 (2020), pp. 1105–1122.
- [20] M. TAHIR, A. TUBAISHAT, F. AL-OBEIDAT, B. SHAH, Z. HALIM, AND M. WAQAS, *A novel binary chaotic genetic algorithm for feature selection and its utility in affective computing and healthcare*, *Neural Computing and Applications*, (2020), pp. 1–22.
- [21] M. WANG AND D. NIU, *Research on project post-evaluation of wind power based on improved anp and fuzzy comprehensive evaluation model of trapezoid subordinate function improved by interval number*, *Renewable energy*, 132 (2019), pp. 255–265.
- [22] S.-L. WANG, Y.-C. LI, AND C.-P. ZHANG, *Analysis of the effect of social support on sustainable competitive advantage in tourism industry-based on the perspective of living-ecologyproduction integrated space*, *Revista de Cercetare si Interventie Sociala*, 71 (2020), pp. 250–263.
- [23] T. WANG, G. HAN, S. LIANG, P. YU, Q. LU, Y. LU, M. HUANG, ET AL., *The path of increasing income for new agricultural enterprises in guangxi.*, *Journal of Southern Agriculture*, 50 (2019), pp. 1640–1646.
- [24] Y. WU ET AL., *Theoretical definition and evaluation of marine cultural resources from the perspective of rural revitalization: a literature review*, *Environment, Resource and Ecology Journal*, 5 (2021), pp. 49–52.
- [25] D. XINGME, *Integrated development of health tourism under the background of rural revitalization strategy*, *Forest Chemicals*

Review, (2021), pp. 247–254.

- [26] N. YI, J. XU, L. YAN, AND L. HUANG, *Task optimization and scheduling of distributed cyber-physical system based on improved ant colony algorithm*, Future Generation Computer Systems, 109 (2020), pp. 134–148.
- [27] Z. YONGXUN AND H. LULU, *Protecting important agricultural heritage systems (iahs) by industrial integration development (iid): practices from china*, Journal of Resources and Ecology, 12 (2021), pp. 555–566.
- [28] C. YU, L. ZHANG, R. MIAO, AND M. WANG, *Changes and prospects of rural teacher compensation policy from the perspective of positive psychology*, Psychiatria Danubina, 33 (2021), pp. 265–267.
- [29] L. ZHANG, J. ZHANG, L. ZHANG, C. WU, AND Y. ZHANG, *Study on the classification of forestry infrastructure from the perspective of supply based on the classical quartering method*, Applied Mathematics and Nonlinear Sciences, 6 (2021), pp. 447–458.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Jan 5, 2024

Accepted: Feb 9, 2024



RESEARCH ON VISUALIZATION AND INTERACTIVITY OF VIRTUAL REALITY TECHNOLOGY AND DIGITAL MEDIA IN INTERIOR SPACE DESIGN

KE ZHANG* AND RATANACHOTE THIENMONGKOL†

Abstract. The combination of digital media and Virtual Reality (VR) technology has become a disruptive force in the interior space design industry, changing the way that design is explored and augmenting its experiential aspects. Examining the substantial effects of deep visualization and technological interaction on the conception and interaction with interior environments, this research explores the junction of these dynamic domains. This study intends to uncover the interactions that drive internal designing spaces into a new era of creativity and engagement with users through a thorough investigation of state-of-the-art VR tools and multimedia services and technologies. The research approach entails a thorough examination of popular virtual reality systems and digital media strategies used in the interior design industry. We'll carry out user feedback surveys and real-world case studies to see how well these tools work at communicating design ideas and encouraging group decision-making. Considering aspects like accessibility, visual dedication, and real-time interaction, the study aims to identify the main opportunities and obstacles associated with the combination of VR and electronic media inside interior space design. This research will lead to the proposal of a framework for the best possible use of digital media and virtual reality in interior space design. The framework is intended to serve as a guide for designers, architects, and other relevant parties as they fully utilize these technologies to improve visualization, collaboration, and user experience. Furthermore, the study will provide insightful information to the larger conversation on how emerging technologies are reshaping design disciplines.

Key words: visualization, virtual reality, digital media, interior space design, deep visualization

1. Introduction. The development of technology has brought about a new degree of media trends and information distribution, which has profoundly altered people's lives. Virtual reality technology becomes an interactive media design and a medium for art transmission and expression when it combines internet content art and technology [24]. With the use of virtual reality technology, participants in digital media art can finish this interactive experience.

In contrast to traditional marketing, which includes advertising on radio, television, billboards, and other printed media, digital marketing offers digital payment [3], rapid tracking and control, and data analysis on the campaign's effectiveness online [19]. Digital marketing is the term for online marketing initiatives that educate new clients by aligning with their demands [15]. It is the online projection of traditional marketing techniques, resources, and approaches [14]. Due to the rapid increase in internet users, digital marketing has created a multitude of avenues via which businesses can interact with different types of customers. New applets and sub-channels on the market, such the community for game content, comics, and animation on Bilibili and the short video platform Tik Tok, have gained a lot of traction [20].

Marketing for interior design firms is difficult among other industries since designers must personally comprehend each client's unique preferences before showcasing creative proposals. Customers seeking interior design, on the other hand, typically like to see the design concepts in action. Therefore, in order to draw in and keep clients, the majority of traditional interior design companies open traditional brick-and-mortar locations. Artificial intelligence (AI), virtual reality (VR), and other related technologies have made it possible for online platforms to interactively depict designs and illustrate concepts in an intuitive way in the digital age. In interior design, a platform with interactive virtual reality (IVR) characteristics allows designers to show clients thoughts and ideas while also letting them feel the design intuitively [2].

Very little study has looked at the viewpoint of the consumer; most current studies on digital marketing are from the company's perspective, analysing how businesses might enhance their digital marketing capabilities

*Department of New Media, Faculty of Informatics, Mahasarakham University Maha Sarakham, Thailand

†School of Architecture and Art, City University of Hefei, Chaohu City, Anhui, 238076, China (ratanachotemon1@outlook.com)

to draw customers [7, 27, 28, 21]. The application of an Internet-of-Things (IoT)-based information system in Logistics 4.0 was examined by Tang et al. [20]. Examining the features of the digital information system in relation to customer satisfaction was fresh and valuable information in the article. Some entities claim that digital marketing is exclusively relevant for business-to-consumer (B2C) entities [16]. However, B2B businesses have realized that digital marketing may be successful because to the success stories of corporations like Cisco and IBM [18]. In digital marketing, content is crucial in helping consumers make decisions [12, 9].

The motivation for this research is twofold: First, to provide a comprehensive analysis of current VR tools and digital media strategies employed in the interior design sector, evaluating their efficacy in enhancing design communication, facilitating collaborative decision-making, and improving the overall design experience. Second, to identify and systematically address the main opportunities and challenges that the amalgamation of VR and digital media presents in interior space design. This includes considerations of accessibility, visual fidelity, and the capacity for real-time interaction, which are critical for the effective implementation of these technologies.

The main contribution of the proposed method is given below:

1. Creating cutting-edge virtual reality methods that give designers more lifelike and engrossing representations of interior spaces.
2. Real-time lighting simulations, high-fidelity rendering, and intricate material representations are a few examples of this. creating and putting into practice natural and intuitive ways for people to interact in virtual environments.
3. To make the design process more user-friendly, this might include voice commands, haptic feedback systems, and gesture recognition.
4. The proposed method also uses generative adversarial networks (GAN) based VR to design an interior space.

The rest of our research article is written as follows: Section 2 discusses the related work on various virtual reality methods and deep learning methods. Section 3 shows the algorithm process and general working methodology of proposed work. Section 4 evaluates the implementation and results of the proposed method. Section 5 concludes the work and discusses the result evaluation.

2. Related Works. The human brain's ability to approximate reality is known as perception. It is a static process that involves gathering information about the outside world and interpreting it considering the individual's needs, wants, and attitudes. A person's subjective personal account of an experienced event is called perception. Understanding user experience is of greater interest to academics than understanding customer pleasure since it is viewed as a larger term [25].

Simple perception, or the moment we become aware of stimuli through our senses, is where perception theory starts. People react to what they see by using their judgment or intuition. A customer's likelihood of making a purchase increase with the amount of time they spend exploring, hence customer-centric innovation is crucial for product design and development [13]. Marketers must try to enhance customer satisfaction and present an accurate image [1]. One of the research topics in the realm of consumer behavior is consumer perception, which is described as one of the independent variables influencing consumer behavior [8]. Customers' attitudes, beliefs, and motivations are ultimately influenced by the stimuli they accept and adapt to, and perception is one of the primary personal aspects that shapes behavior and other characteristics [10].

The process by which a person learns about the surroundings and interprets the information considering his or her needs, requirements, and attitudes is known as customer perception., who developed the idea of perceptual filter theory. According to this theory, a stimulus will first be filtered, then sorted, altered, and last stored in the customer's memory. Humans are not able to perceive every stimulus during the sensation phase, according to the author [6]. Furthermore, individuals are unable to react to multiple stimuli at once, interpret stimuli incorrectly, and ultimately lose the ability to recall all the triggers.

SEO, SEM, content marketing, influencer marketing, content automation, campaign marketing, data-driven marketing, e-commerce, social media marketing, social media optimization, direct email marketing, display advertising, e-books, CD-ROMs, and games are some examples of digital marketing techniques [17]. With the advancement of digital marketing, digital media can now be accessed through non-internet methods such cell phones (SMS and MMS), callbacks, and ringtones. The differentiation between online and digital marketing is aided by this expansion for non-internet outlets [22].

One new kind of art is digital media art. Emerging technologies like virtual reality and augmented reality can combine to produce unanticipated results. Bastug investigated the application of virtual reality technology in laparoscopic surgery. Using virtual reality technology, he evaluated the test doctors' level of surgical coordination and contrasted their performances. According to his experimental findings, the simulator's parameters are used to evaluate a surgeon's laparoscopic skills [26]. In a virtual environment, Kihonge outlines a synthesis process for creating 4C space mechanisms. He also creates software that enables several users to network and share the created mechanisms.

In virtual reality, people may view and interact with digital models more naturally than they can with a typical human-machine interface (HCI) [5]. Freeman conducts research on the application of virtual reality to the treatment of substance abuse, eating disorders, schizophrenia, and depression. Using computer-generated interactive environments and virtual reality, he repeatedly encounters the challenging circumstances faced by individuals with various mental diseases and gains knowledge on how to resolve them through evidence-based psychotherapy. The ways in which his research methodologies are used to mental health treatment are significant [23]. Using an immersive boxing movement guide, Sucipto provides educational content. With an Android-based animation video visualization serving as the learning medium and offering instructions for fundamental boxing motions, the methodology is preferable to outdated manual approaches.

A useful tool for helping boxers learn actions by heart is the 3D AR boxing action training program [4]. Using the ADDIE-type research and development approach, Ayu conducts experimental steps of analysis, design, development, implementation, and evaluation with the goal of enhancing the artistic literacy of primary school pupils. Enhancing reality-based applied media may effectively boost students' creative literacy, according to his online user survey, which he conducted to determine the value of this learning tool [11].

The human brain's selective and subjective interpretation of external stimuli complicates the design and evaluation of user experiences. Tailoring experiences to meet diverse individual needs and preferences remains a significant challenge. In an era of information overload, individuals' inability to perceive every stimulus or react to multiple stimuli simultaneously poses a challenge for designing effective marketing strategies and user interfaces that capture and retain attention. Ensuring customer satisfaction while accurately representing products or services requires a deep understanding of consumer perception and behaviour. Marketers must navigate these aspects without overwhelming or misleading customers. The rapid advancement of digital marketing technologies, including SEO, SEM, and social media, requires marketers to continuously update their skills and strategies to remain effective. Developing effective VR and AR tools for education and training, such as the 3D AR boxing action training program, requires addressing challenges related to user engagement, content accuracy, and technological accessibility.

3. Proposed Methodology. The outcome of this study will be the recommendation of a framework for the most effective application of virtual reality and digital media in interior design. The framework is meant to act as a roadmap for designers, architects, and other pertinent stakeholders as they make the most of these technologies to enhance user experience, visualization, and collaboration. The proposed method uses GAN-based VR interior design. In figure 3.1 shows the architecture of the proposed method. VR allows users to immerse themselves in a virtual representation of an interior space before it is physically realized. This immersive experience goes beyond flat images or models, offering a 360-degree view that gives a sense of scale, depth, and spatial relationships that can be comprehensively understood only when experienced as if one were physically present in the space. With VR, users can interact with the interior environment in real-time. They can move around, explore different angles, and even manipulate design elements (such as changing materials, lighting, or furniture layouts) within the virtual space. This level of interaction enables users to experiment with design choices and see the immediate impact of those changes, fostering a deeper connection with the design process and the space itself.

3.1. Data Collection. With the help of a virtual interior design platform, we gathered a sample for this study's digital marketing purposes from the Hong Kong interior design market. We then gathered pertinent data to examine the effect of digital marketing on customer intention. Just 40% of interior designers in Hong Kong have a website [20]. Nonetheless, a Google search for "interior designer Hong Kong" yielded about 22 million results. Hong Kong has an excessive number of marketing websites pertaining to the interior design sector. Hong Kong has a large consumer base that works in the interior design sector.

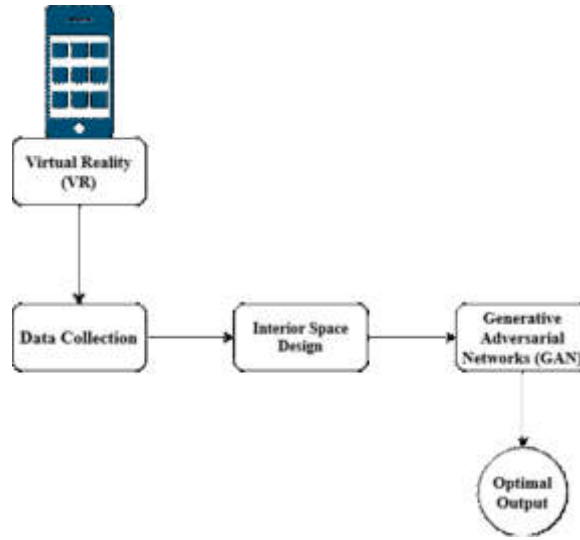


Fig. 3.1: Architecture of Proposed Method

3.2. Questionnaire Development. A questionnaire survey was employed as a research tool in this work. The questionnaire was divided into sections. Initially, the interviewee was sent a connection to a virtual interior design platform, where entry questions verified that the interviewee had visited the platform and was aware of the presumptions that needed to be made before answering the questionnaire. "I am looking for an interior design service," was the interviewee's presumption. In the second section, there were five variables related to the above-mentioned hypothesis model: (1) perceived aesthetics of the platform; (2) perceived usability; (3) perceived quality of the content; (4) customer happiness; and (5) behavioral intention. All variables' definitions and measurement elements are covered later.

Three reference questions were asked in the third section: (1) did they look for or are interested in interior design services currently? (2) did they look for or hire interior design services in the last 12 months? and (3) which channels would they want to use to locate an interior design company? The questionnaire's final section asked about demographic data.

3.3. Generative Adversarial Networks (GAN). The author created the generative adversarial network (GAN) in 2014. In several machine learning domains, interest has been growing in this remarkable discovery. Two neural networks that interact make up the GAN. It is both a discriminator (D) and a generator (G) (D). To create new data instances, the generator network is taught to map points in the latent space. The generator network's plausible and actual images are separated by the discriminator network during training. The generator ultimately produces images that mimic real training examples. Depending on the discriminator's expectations, the generator is modified to produce better images during training. Its capacity to discriminate between real and fraudulent images is improved by the discriminator. The discriminator loss is based on the distinction between authentic and fake labels. If the image is man-made or environmental, this is indicated by the label. Figure 3.2 displays the GAN's general diagram.

A two-player min-max game that may be described by, for example, can be used to represent the primary goal of GAN theory.

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim P_d(x)} [\log D(x)] + \mathbb{E}_{r_{nv} \sim P_{r_{nv}}(r_{nv})} [\log(1 - D(G(r_{nv})))] \quad (3.1)$$

With the value function V, the discriminator and the generator are playing a min-max game (D, G). The discriminator seeks to minimise its reward V(D, G), and the generator seeks to maximise its loss by seeking to diminish the discriminator's award.

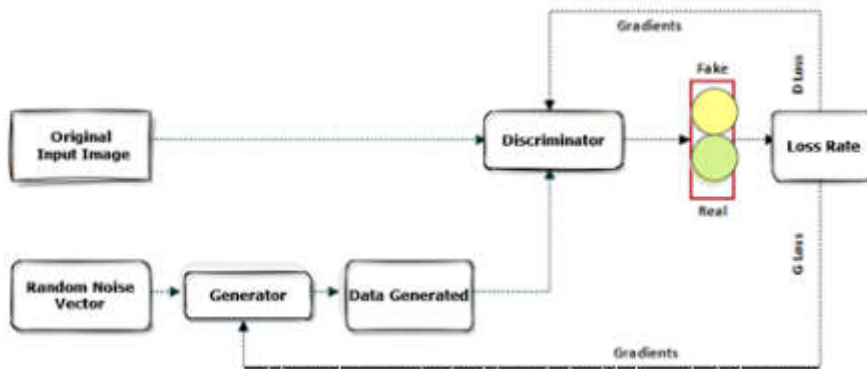


Fig. 3.2: Architecture of General Adversarial Network

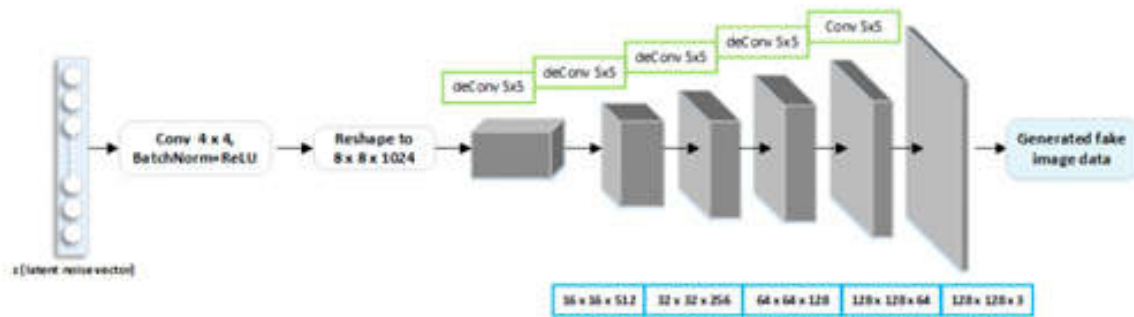


Fig. 3.3: Generator Model Architecture

The preceding loss function is always attempted to be minimised by the generator, whereas it is always attempted to be maximised by the discriminator. The GAN generator adds the random noise variable $P_{rnv}(rnv)$ to the original input data x before generating samples $G(rnv)$. The discriminator’s estimates of the likelihood that an actual example of x exists in the P_d distribution of data is denoted by $D(x)$. The discriminator’s assessment of the likelihood that a phoney example is real is called $D(G(rnv))$. To deceive the discriminator, the generator strives to produce nearly flawless images. In comparison, the discriminator works to enhance efficiency by separating fake examples from real ones until it reaches a point where it is impossible to tell fake examples from real ones.

To produce quality synthesised images that can be orchestrated with high-quality photographs, the generator must be interacting with a deeper network. A deeper network will have a larger convolution layer and require more training data. We first supplied the original image data and downsized it to $128 \times 128 \times 3$ to consider GPU for the training. To match the generator, the image was scaled to $[1, 1]$ pixel values. Because it makes use of the tanh activation function, it was issued. The generator network produces fictitious samples from a 100×1 noise vector. To produce excellent generated images, we combined ReLU activation with four convolution layers. The architecture of the generator model is shown in Figure 3.3.

We partially merged the encoder into the discriminator under the presumption that the network information of the encoder and discriminator overlapped. While the discriminator seeks to identify the discriminating feature, the encoder’s primary objective is to understand the representation feature.

$$\mathcal{L}_{recons}^{pix} = \mathbb{E}_{q \sim D_{encoder}(x), x \sim I_{real}} [|\kappa(q) - \tau(x)|] \tag{3.2}$$

The decoder's function reflects operations on example κ from real pictures, I_{real} and the discriminator's feature map is q .

3.4. Data Analysis Tool. After creating and compiling the questionnaire online using Google Forms, we analyzed and verified the reliability of the descriptive data collected from the sample using SPSS Statistics 25.0. Finally, users can get the data to back up their research models and theories by using IBM SPSS AMOS.

Cronbach's alpha (α), composite reliability (CR), average variance extracted (AVE), and the standardised factor loading of the test items were used to quantify reliability and convergent validity. The acceptance threshold for each measurement's standardised factor loading needs to be higher than 0.700. An internal consistency technique called Cronbach's alpha (α) is used to assess how consistently respondents answered each item in the measurement in this study. Valid alpha values fall between 0.7 and 0.8, whereas values greater than 0.9 indicate exceptionally high internal consistency.

Average variance extracted (AVE) and composite reliability (CR) are two metrics that can be used to assess convergent validity. The constraint in loadings would be the distinction between α and CR. Whereas the weights or loadings for Cronbach's alpha are always required to be identical, the build loadings for CR are flexible. A CR value of 0.8 or higher is regarded as satisfactory. To be deemed acceptable, the average variance extracted (AVE) must be more than 0.5.

VR and digital media facilitate collaboration among designers, clients, and other stakeholders by providing a common visual language. This is particularly useful in projects where decision-makers are geographically dispersed. These tools enable rapid prototyping and iterations based on feedback, making the design process more agile and responsive to user needs.

4. Result Analysis. The proposed method evaluates the performance using parameters such as accuracy, precision, recall and f1-score for interior space design.

Accuracy is an important parameter in many machine learning tasks, including deep learning tasks, but in the context of research on Visualization and Interactivity of Virtual Reality (VR) Technology and Digital Media in Interior Space Design, it may not be the only metric to pay attention to. Because of the nature of the research, both subjective and objective criteria are used.

Analyze how well deep learning models are at tasks like object detection, image recognition, and, if relevant, natural language processing. Consider additional metrics such as F1-score, precision, and recall, particularly if the research entails tasks. Use subjective metrics, like user surveys, interviews, and feedback, to evaluate the user experience. Consider elements like as immersion, enjoyment, and simplicity of engagement.

Compare the virtual and real-world interior space representations to see how realistic they are. This may entail the subjective evaluations of users or interior design specialists. Calculate how quickly the system reacts to user input in the virtual environment. Experiences with low latency are more immersive and participatory. In figure 4.1 shows the evaluation of accuracy.

In research, precision refers to making sure that the techniques, measurements, and analyses used in the study are precise, dependable, and in line with the goals of the investigation. Here are some things to keep in mind when conducting a research study on the interactivity and visualization of digital media and Virtual Reality (VR) technologies in interior space design. Clearly state the goals of the study on the use of digital and virtual reality for interior space design visualization and interactivity. Make sure the goals are in line with the study's overall aims. Researchers can improve the accuracy of their investigation into the visualization and interaction of virtual reality technology and digital media in interior space design by taking these factors into account. Throughout the research process, returning to and improving these elements on a regular basis enhances the overall quality and dependability. In figure 4.2 shows the evaluation of precision.

Recalling or summarizing the main points and objectives of the research is undoubtedly what is meant by "recall" in the context of studies on the visualization and interaction of digital media and virtual reality technology in interior space design. It entails recalling the primary goals, approaches, and essential elements of the research. Examine the state of virtual reality and digital media today, as well as how they are being used in interior design. Create deep learning models to enhance the interior space visualization. Examine ways to use VR and digital media to improve interaction in virtual settings. Establish a smooth connection between VR technologies and deep learning models to portray interior spaces. In figure 4.3 shows the evaluation of recall.

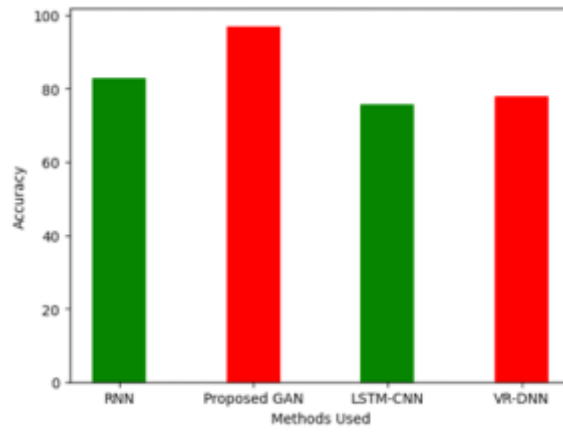


Fig. 4.1: Accuracy

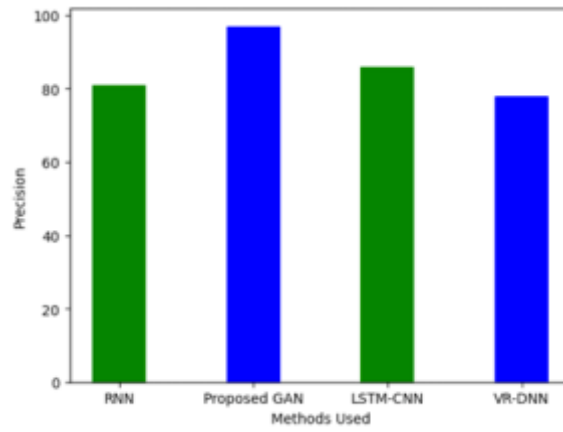


Fig. 4.2: Precision

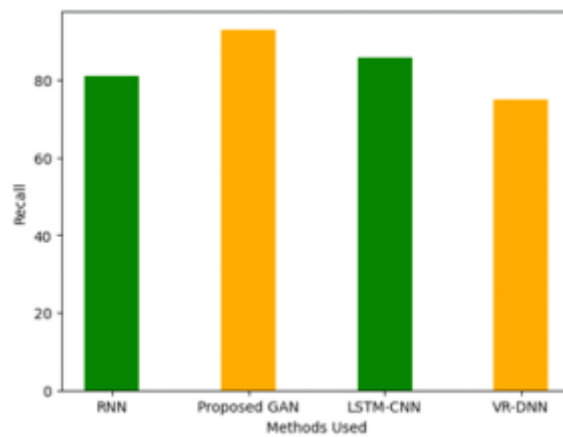


Fig. 4.3: Recall

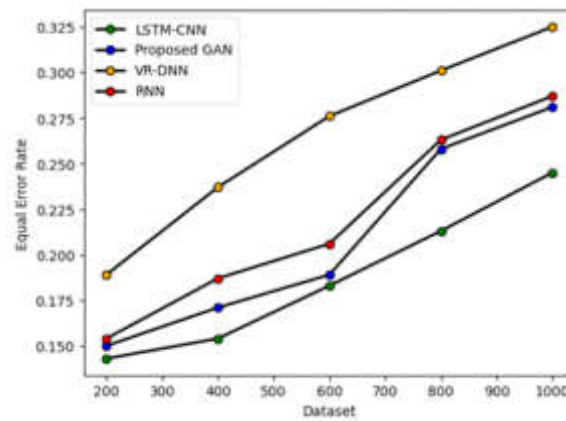


Fig. 4.4: Evaluation of Error Rate

In contrast to certain other domains, such machine learning classification tasks, the concept of "error rate" is not as simple when applied to study on the presentation and interactivity of digital media and Virtual Reality (VR) technologies in interior space design. You can test accuracy in the context of deep learning models that you may utilize for different tasks in your research. This represents the percentage of cases that were accurately anticipated or classified. For instance, accuracy would be the percentage of elements properly classified in a VR environment where your deep learning model is used to classify interior elements. A variety of indicators, including completion rates, task success rates, and user satisfaction ratings, are used to evaluate the effectiveness of interactive features.

Consider measures like rendering speed and latency that are associated with the VR experience's quality. In this case, low error rates would imply no lag or delay in the virtual environment's rendering, which would enhance the user experience and make it more immersive. When conducting usability testing, user comments can also be used to estimate error rates. For example, it suggests a higher error rate in terms of user experience if users complain or run into problems during specific interactions. Analyse the error rates according to ethical and privacy concerns. In the context of ethics, incidents involving illegal data access or privacy violations could be regarded as mistakes. In figure 4.4 shows the evaluation of Error Rate.

When calculating the F1-score for a research project, one must evaluate how well recall and precision are balanced in relation to the objectives of the study. You can define the F1-score as striking a balance between the deep learning models' accuracy and their capacity to capture the nuances of interior space design in the context of your research on the visualization and interactivity of virtual reality (VR) technology and digital media in interior space design using deep learning methods.

An excellent balance between recall and precision is shown by a high F1-score, which implies that interior design components are extensively and accurately captured by the deep learning models. A low F1-score could be a sign of a trade-off between recall and accuracy, where the models perform well in recollection but poorly in the former. Assess the F1-score on a regular basis at various points in your research, such as following model training, prototype building, and user testing, to monitor how well your technique is working to accomplish the aims of the study. Based on the F1-score results, modifications can be made to the models and techniques to improve their precision and recall. In figure 4.5 shows the evaluation of F1-score.

5. Conclusion. Within the interior space design sector, the integration of digital media and Virtual Reality (VR) technology has become a disruptive force, altering the way design is explored and enhancing its immersive features. This study investigates the intersection of these dynamic domains, looking at the significant implications of deep visualization and technology contact on the conceptualization and interaction with interior environments. Through a detailed analysis of cutting-edge VR tools, multimedia services, and technology, this study aims to reveal the interactions that propel interior designing environments into a new era of creativity

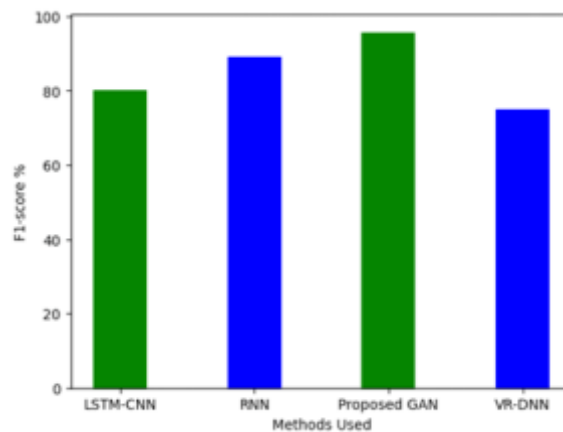


Fig. 4.5: F1-score

and user engagement. The research methodology comprises a detailed analysis of widely utilized digital media techniques and virtual reality technologies in the interior design sector. We'll conduct user feedback surveys and real-world case studies to evaluate how well these technologies facilitate group decision-making and the communication of design concepts. The study attempts to identify the primary opportunities and challenges related with the integration of VR and electronic media into interior space architecture, considering factors like accessibility, visual dedication, and real-time engagement. The outcome of this study will be the recommendation of a framework for the most effective application of virtual reality and digital media in interior design. The framework is meant to act as a roadmap for designers, architects, and other pertinent stakeholders as they make the most of these technologies to enhance user experience, visualization, and collaboration. Additionally, the project will contribute meaningful data to the greater discussion on how to develop.

Address the ethical and privacy implications of using VR and digital media in interior design, particularly in relation to data collection, user consent, and the digital representation of private spaces. Developing guidelines and best practices for ethical use of these technologies will be crucial as their adoption grows.

Acknowledgement. Scientific Research Program for Higher Education Institutions in Anhui Province (Natural Science) (2023AH040367)

REFERENCES

- [1] A. ALKHALIFAH, *Developing mobile commerce website design to enhance users experience*, IJCSNS Int. J. Comput. Sci. Netw. Secur, 17 (2017), pp. 65–69.
- [2] A. ALPER, E. S. ÖZTAS, H. ATUN, D. ÇINAR, AND M. MOYENGA, *A systematic literature review towards the research of game-based learning with augmented reality.*, International Journal of Technology in Education and Science, 5 (2021), pp. 224–244.
- [3] R. F. K. AYU, Z. JANNAH, N. FAUZIAH, T. N. NINGSIH, M. MANILATURROHMAH, D. A. SURYADI, R. P. N. BUDIARTI, AND F. K. FITRIYAH, *Planetarium glass based on augmented reality to improve science literacy knowledge in madura primary schools*, Child Education Journal, 3 (2021), pp. 19–29.
- [4] Y. K. CHAN, Y. M. TANG, AND L. TENG, *A comparative analysis of digital health usage intentions towards the adoption of virtual reality in telerehabilitation*, International Journal of Medical Informatics, 174 (2023), p. 105042.
- [5] X. CHEN, Q. HUANG, AND R. M. DAVISON, *Economic and social satisfaction of buyers on consumer-to-consumer platforms: The role of relational capital*, International Journal of Electronic Commerce, 21 (2017), pp. 219–248.
- [6] O. H. CHI, C. G. CHI, D. GURSOY, AND R. NUNKOO, *Customers acceptance of artificially intelligent service robots: The influence of trust and culture*, International Journal of Information Management, 70 (2023), p. 102623.
- [7] M. DI AND H. G. KIM, *Shape of light: interactive analysis of digital media art based on processing*, TECHART: Journal of Arts and Imaging Science, 7 (2020), pp. 23–29.
- [8] Q. DOU, X. S. ZHENG, T. SUN, AND P.-A. HENG, *Webthetics: quantifying webpage aesthetics with deep learning*, International Journal of Human-Computer Studies, 124 (2019), pp. 56–66.

- [9] H. FAZLOLLAHTABAR, *Intelligent marketing decision model based on customer behavior using integrated possibility theory and k-means algorithm*, J. Intell Manag. Decis, 1 (2022), pp. 88–96.
- [10] Y.-C. HUANG, L.-N. LI, H.-Y. LEE, M. H. BROWNING, AND C.-P. YU, *Surfing in virtual reality: An application of extended technology acceptance model with flow theory*, Computers in Human Behavior Reports, 9 (2023), p. 100252.
- [11] O. IGLESIAS, S. MARKOVIC, J. J. SINGH, AND V. SIERRA, *Do customer perceptions of corporate services brand ethicality improve brand equity? considering the roles of brand heritage, brand image, and recognition benefits*, Journal of business ethics, 154 (2019), pp. 441–459.
- [12] A. P. JULIANA, D. M. LEMY, R. PRAMONO, A. DJAKASAPUTRA, AND A. PURWANTO, *Hotel performance in the digital era: Roles of digital marketing, perceived quality and trust*, Journal of Intelligent Management Decision, 1 (2022), pp. 36–45.
- [13] H. JYLHÄ AND J. HAMARI, *An icon that everyone wants to click: How perceived aesthetic qualities predict app icon successfulness*, International Journal of Human-Computer Studies, 130 (2019), pp. 73–85.
- [14] Y. LIU, S. WU, Q. XU, AND H. LIU, *Holographic projection technology in the field of digital media art*, wireless communications and mobile computing, 2021 (2021), pp. 1–12.
- [15] L. D. R. MORENO, *Museums and digital era: preserving art through databases*, Collection and Curation, 38 (2019), pp. 89–93.
- [16] A. PALANCI AND Z. TURAN, *How does the use of the augmented reality technology in mathematics education affect learning processes?: a systematic review*, Uluslararası Eğitim Programları ve Öğretim Çalışmaları Dergisi, 11 (2021), pp. 89–110.
- [17] G. PINO, C. AMATULLI, R. NATARAAJAN, M. DE ANGELIS, A. M. PELUSO, AND G. GUIDO, *Product touch in the real and digital world: How do consumers react?*, Journal of Business Research, 112 (2020), pp. 492–501.
- [18] K. SOHN AND O. KWON, *Technology acceptance theories and factors influencing artificial intelligence-based intelligent products*, Telematics and Informatics, 47 (2020), p. 101324.
- [19] A. SUCIPTO, Q. J. ADRIAN, AND M. A. KENCONO, *Martial art augmented reality book (arbook) sebagai media pembelajaran seni beladiri nusantara pencak silat*, Jurnal Sisfokom (Sistem Informasi Dan Komputer), 10 (2021), pp. 40–45.
- [20] Y. M. TANG, Y.-Y. LAU, AND U. L. HO, *Empowering digital marketing with interactive virtual reality (ivr) in interior design: Effects on customer satisfaction and behaviour intention*, Journal of Theoretical and Applied Electronic Commerce Research, 18 (2023), pp. 889–907.
- [21] K. TARUTANI, H. TAKAKI, M. IGETA, M. FUJIWARA, A. OKAMURA, F. HORIO, Y. TOUDOU, S. NAKAJIMA, K. KAGAWA, M. TANOOKA, ET AL., *Development and accuracy evaluation of augmented reality-based patient positioning system in radiotherapy: a phantom study*, in vivo, 35 (2021), pp. 2081–2087.
- [22] D. TREHAN AND R. SHARMA, *Assessing advertisement quality on c2c social commerce platforms: an information quality approach using text mining*, Online Information Review, 45 (2021), pp. 46–64.
- [23] Y.-S. WANG, T. H. TSENG, W.-T. WANG, Y.-W. SHIH, AND P.-Y. CHAN, *Developing and validating a mobile catering app success model*, International Journal of Hospitality Management, 77 (2019), pp. 19–30.
- [24] W. YE, Y. LI, ET AL., *Design and research of digital media art display based on virtual reality and augmented reality*, Mobile Information Systems, 2022 (2022).
- [25] Y. ZHANG, J. SU, H. GUO, J. Y. LEE, Y. XIAO, AND M. FU, *Transformative value co-creation with older customers in e-services: Exploring the influence of customer participation on appreciation of digital affordances and well-being*, Journal of Retailing and Consumer Services, 67 (2022), p. 103022.
- [26] Y. ZHAO, L. WANG, H. TANG, AND Y. ZHANG, *Electronic word-of-mouth and consumer purchase intentions in social e-commerce*, Electronic Commerce Research and Applications, 41 (2020), p. 100980.
- [27] W. ZHU, *Study of creative thinking in digital media art design education*, Creative Education, 11 (2020), pp. 77–85.
- [28] Y. ZHUANG, J. SUN, AND J. LIU, *Diagnosis of chronic kidney disease by three-dimensional contrast-enhanced ultrasound combined with augmented reality medical technology*, Journal of Healthcare Engineering, 2021 (2021), pp. 1–12.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Jan 5, 2024

Accepted: Feb 9, 2024



RESEARCH ON THE INFLUENCING FACTORS OF COMMERCIAL PENSION INSURANCE FOR RURAL RESIDENTS IN THE CONTEXT OF POPULATION AGING BASED ON BIG DATA ANALYSIS

SHITANG FENG*

Abstract. Data Pension Explorer (DAPE) introduces an innovative approach to delve into the intricacies of commercial pension insurance adoption within rural communities in China, particularly amid the challenges posed by an aging population. Leveraging advanced Deep Long Short-Term Memory (LSTM) techniques within the domain of big data analysis, this study pioneers the analysis of extensive datasets. It systematically unravels intricate patterns, correlations, and pivotal determinants shaping the landscape of pension insurance adoption in rural areas. Going beyond conventional analyses, this research provides a nuanced understanding of the multifaceted factors influencing both the sustainability and uptake of pension insurance. The culmination of these efforts yields valuable insights that extend beyond the theoretical realm, directly informing strategic decision-making processes. These insights prove instrumental in designing and implementing policies tailored to address the unique challenges faced by rural communities in China. Thus, DAPE not only navigates the complexities of population aging but also serves as a guiding force in fostering widespread adoption and sustainability of pension insurance within rural landscapes in the Chinese context.

Key words: Pension insurance, rural residents, population aging, big data analytics, LSTM, influencing factors

1. Introduction. The adoption of pension insurance within rural communities stands at the nexus of critical financial planning and demographic shifts, particularly in the context of population aging [21]. Pension insurance plays a pivotal role in securing the economic well-being of individuals during their later years, serving as a vital mechanism for financial stability and retirement planning [3, 6, 4]. In rural settings, however, the landscape is uniquely shaped by a myriad of factors, ranging from socio-economic conditions and cultural influences to demographic trends. Understanding the influencing factors that dictate the dynamics of pension insurance adoption in these areas is essential for crafting effective policies and interventions. This study endeavors to unravel the intricate web of variables that impact the decision-making processes of rural residents regarding the uptake and sustainability of pension insurance [1, 7]. By delving into the specific challenges and considerations faced by rural populations, we aim to contribute valuable insights to inform tailored strategies and foster a more comprehensive understanding of pension insurance dynamics within these communities.

Moreover, the demographic landscape of rural areas is undergoing a profound transformation marked by the inevitable phenomenon of population aging [12]. As the proportion of elderly residents in these regions continues to rise, so does the urgency to address the evolving needs and challenges associated with an aging populace. Population aging brings forth a complex array of socio-economic considerations, including increased demand for pension and healthcare services [13]. Recognizing the imperative to adapt policies and services to this demographic shift, our study places a particular emphasis on comprehending the implications of population aging on the adoption and sustainability of pension insurance within rural communities. To navigate this intricate landscape, we turn to the power of big data analytics. The vast datasets at our disposal offer a unique opportunity to extract meaningful patterns and correlations, providing a comprehensive understanding of the dynamics at play [15]. Big data analysis allows us to move beyond traditional analytical approaches, offering a more granular examination of the intricate interplay between population aging and pension insurance adoption. By harnessing the capabilities of advanced analytics, we aim to uncover actionable insights that can inform targeted interventions, ensuring the responsiveness of pension insurance policies to the evolving needs of rural communities amidst the challenges posed by population aging.

*Nanchong Vocational and Technical College, Department of Finance, Nanchong, 637000, China (shitangfengresearch@outlook.com)

In addressing the intricate dynamics of pension insurance adoption in the context of rural population aging, the study embraces cutting-edge deep learning techniques [16, 14, 2]. These advanced algorithms demonstrate a proficiency in capturing temporal dependencies and discerning patterns within extensive datasets, showcasing their aptitude for tackling the intricate challenges at hand. Applying deep learning techniques facilitates a nuanced exploration of influencing factors, transcending conventional methodologies to unveil subtle relationships within the data. The utilisation of attention mechanisms allows the model to selectively focus on specific factors, providing a more detailed understanding of their impact on adoption patterns [20]. Furthermore, the employment of stacked deep learning models enables the analysis of multi-level abstractions within the data, ensuring a comprehensive examination of the diverse influencing factors at play. Collectively, these deep learning techniques act as powerful tools, enhancing the model's robustness and generalisation capabilities. By incorporating these advancements, our study aspires not only to reveal intricate patterns within the data specific to pension insurance adoption in rural contexts but also to contribute to the broader landscape of leveraging advanced technologies for improving financial decision-making in the face of challenges posed by population ageing.

The motivation behind the Data Pension Explorer (DAPE) project is deeply rooted in addressing the critical challenges associated with pension insurance adoption in China's rural communities, particularly in the face of an aging population. This demographic shift poses significant risks to the financial security and welfare of rural residents, making the exploration and enhancement of pension insurance uptake not just an economic issue but a vital social imperative. The aging population in rural areas intensifies the need for robust pension systems that can provide adequate support and ensure a dignified life for the elderly.

The main contributions of the paper as follows

1. The paper presents Data Pension Explorer (DAPE), a new method that uses advanced deep learning to study the factors influencing the adoption of pension insurance in rural China during population aging.
2. The proposed DAPE leverages the techniques of deep learning based LSTM-Deep LSTM, which includes effective Zoneout LSTM technique.
3. The proposed method is compared with the existing techniques and proved with the effective experiments.

2. Literature Review.

2.1. Insights into Aging and Economic Dynamics in Beijing. [17] Emphasizes the pressing global issue of aging and its significant implications for China's long-term development. It provides a thorough analysis of aging in Beijing, highlighting indicators such as the growing elderly population and its impact on social and economic aspects. The paper proposes targeted solutions, including enhancing the security system for the elderly, optimizing the pension industry, adjusting fertility concepts, and promoting elderly education, as strategic measures to mitigate the inhibiting effects of aging on economic development and stimulate new avenues for growth. [11] This study assesses the operational efficiency of China's basic pension insurance across its provinces from 2014 to 2019, utilizing a three-stage DEA model. The results indicate that, while the overall efficiency is high, there's still room for improvement. Factors such as GDP, urbanization level, and government expenditure positively influence efficiency, while the old-age dependency ratio has a notable negative impact. Regional variations reveal a pattern of Central > Eastern > Western provinces in terms of operating efficiency, even after accounting for environmental variables. [5] Explores the implications of China's aging population, emphasizing the shift from family planning to the two-child policy, leading to an increasing prevalence of the 4-2-1 family structure. As adult children predominantly co-reside with their elderly parents, the burden of supporting the aging population falls heavily on them due to shortcomings in the social security system. Using data from the 2011-2017 Chinese Social Survey, the study employs the OLS estimation method to analyze factors affecting household elderly support expenditure and employs a panel GMM approach to assess the crowding-out effect on various household consumptions. [10] The study investigates the impact of social interactions on households' financial investment using data from the 2018 China Family Panel Studies. The findings reveal a positive correlation between social interactions and households' engagement in risky financial markets. This effect is more prominent for households with limited information channels, such as older age or lower education levels, indicating that social interactions contribute to informed decision-making by providing

relevant information. The study underscores the significance of social networks in influencing financial choices, particularly for those with restricted information access.

2.2. Exploring Elderly Depression Dynamics with LSTM. [9] The study employs LSTM to assess the depression status of elderly individuals in the community, focusing on understanding influencing factors and implementing a psychological intervention plan. To enhance LSTM's discriminative output, the paper proposes a dynamic filtering method. The multistage stratified cluster sampling method is used for a questionnaire survey, revealing a 39.38% depression rate among the elderly in a specific community. Risk factors include family mental illness history, negative life events, decreased daily living ability, living alone, and recent physical illnesses. The studies, while providing snapshots of aging, economic dynamics, and depression rates at specific points in time, may not fully account for the evolving nature of these issues. Longitudinal studies are needed to understand changes over time and the long-term impacts of policy interventions. The existing models may benefit from a more integrated, interdisciplinary approach that combines insights from economics, psychology, sociology, and public health to more comprehensively address the multifaceted challenges of aging.

3. Methodology.

3.1. DAPE Overview. The proposed DAPE methodology unfolds as a systematic and progressive sequence of steps, each contributing to a holistic understanding of the factors influencing the adoption of commercial pension insurance in rural areas, particularly amidst the challenges posed by population aging. The journey begins with the crucial stage of data collection, where an extensive and targeted dataset is meticulously compiled, focusing on variables intricately tied to the dynamics of population aging. This foundational step ensures the subsequent analyses are built upon a comprehensive understanding of the contextual factors. Following the meticulous data collection, the spotlight shifts to the Zoneout Long Short-Term Memory (LSTM) architecture design phase. Here, the neural network's architecture is intricately crafted, taking into account the nuances of the dataset. Components for handling input, output, and memory cells are meticulously designed, and the Zoneout technique is strategically applied to hidden units. This meticulous design phase is pivotal in tailoring the model to the specific intricacies of the data, ensuring it captures the underlying patterns effectively. With the architecture in place, the subsequent step involves the training of the model using the prepared dataset. Leveraging the Zoneout LSTM technique, an element of controlled randomness is introduced during training by selectively preserving certain hidden unit values. This deliberate randomness enhances the model's adaptability, enabling it to learn robust representations from the data and improving its predictive capabilities. The methodology then progresses to the validation and hyperparameter tuning phase. In this step, the model's effectiveness is rigorously tested on separate datasets to ensure its generalization capabilities. Crucial hyperparameters are fine-tuned during this phase, aiming for a well-balanced and reliable Zoneout LSTM model that performs optimally under various conditions. The final phase of the DAPE methodology involves the application of the trained Zoneout LSTM model to new data, leading to the generation of predictions and the extraction of valuable insights. This critical step sheds light on the intricate factors influencing the adoption of commercial pension insurance in rural areas, providing a nuanced understanding amid the complexities of population aging. Figure 3.1 encapsulates a comprehensive overview of the DAPE methodology.

3.2. Proposed DAPE Approach. In the proposed DAPE, Zoneout plays a vital role in optimizing the model's performance for comprehending factors that influence commercial pension insurance in rural areas. Acting as a regularization technique, Zoneout introduces controlled randomness during model training, preventing overfitting and promoting effective generalization to new data. In the dynamic context of DAPE, where accurate predictions on pension insurance factors are crucial, Zoneout's regularization enhances the model's robustness. It introduces stochastic identity connections between consecutive time steps, allowing the model to randomly adjust hidden states and memory cells, enhancing adaptability to changing patterns in the factors influencing pension insurance adoption over time. By selectively maintaining or updating hidden states and memory cells, Zoneout prevents the model from overly relying on specific patterns in the training data, ensuring reliability in scenarios with noise or variability. Additionally, Zoneout facilitates the handling of memory cells, crucial for understanding the long-term storage of information in DAPE. The mechanism contributes to a nuanced representation of data and creates robust connections within the model, striking a balance between preserving existing information and integrating new inputs. This robustness proves valuable in navigating the

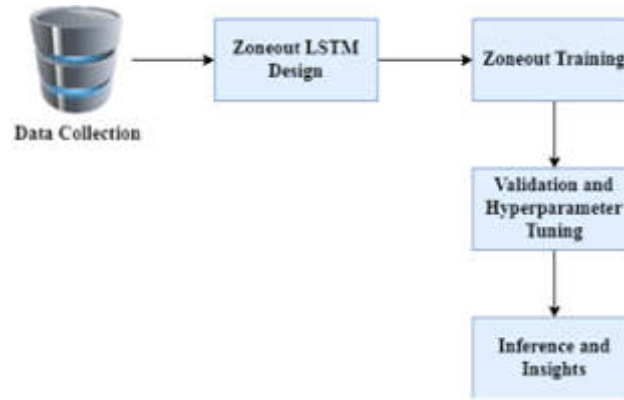


Fig. 3.1: Proposed DAPE general overview

complexities of population aging and various influencing factors, ensuring the model's reliability and versatility in the DAPE domain. This method of Zoneout was adapted from the study [8].

Algorithm 4 Zoneout LSTM Training and Evaluation

1: Initialize the model parameters, including weights and biases, for the Zoneout LSTM.

Forward Pass

2: **for** t in each time step **do**

3: Compute the input, forget, output, and memory cell gates using the Zoneout LSTM equations:

$$i_t, f_t, o_t = \sigma(w_x x_t + w_h h_{t-1} + b) \quad (3.1)$$

$$g_t = \tanh(w_{xg} x_t + w_{hg} h_{t-1} + b_g) \quad (3.2)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (3.3)$$

$$h_t = o_t \odot \tanh(c_t) \quad (3.4)$$

Zoneout

4: Define Zoneout by randomly applying the identity operator or updating based on the Zoneout masks to hidden states and memory cells.

5: Apply dropout to the hidden states:

$$h_t = h_t \odot d_t^h \quad (3.5)$$

6: Zoneout is expressed as:

$$c_t = d_t^c \odot c_{t-1} + (1 - d_t^c) \odot (f_t \odot c_{t-1} + i_t \odot g_t) \quad (3.6)$$

$$h_t = d_t^h \odot h_{t-1} + (1 - d_t^h) \odot (o_t \odot \tanh(f_t \odot c_{t-1} + i_t \odot g_t)) \quad (3.7)$$

7: **end for**

8: d_t^c, d_t^h are the Zoneout and dropout masks for memory cells, hidden states, and input.

Backward Pass Training

9: Compute the loss and gradients using the predicted values and the ground truth.

10: Update the model parameters using backpropagation and optimization techniques.

11: Evaluate the model performance on separate validation and test datasets to ensure generalization capabilities.

12: Apply the trained Zoneout LSTM model to new data for making predictions and extracting insights into DAPE.

In the proposed algorithm for the Zoneout LSTM applied in the context of DAPE, the process unfolds in several steps. Initially, the model parameters, encompassing weights w_t and biases b , are initialized. Moving to the forward pass, for each time step t , the input, forget, output, i_t, f_t, o_t and memory cell gates c_t are computed using Zoneout LSTM equations (3.1) (3.2) (3.3) (3.4). These equations involve sigmoid and hyperbolic tangent

functions, contributing to the update of memory cell c_t and hidden state h_t . Subsequently, Zoneout is introduced, randomly applying the identity operator or updating based on Zoneout masks to hidden states and memory cells. Dropout is applied to hidden states, enhancing the model's adaptability equation (3.5). Zoneout equations (3.6) (3.7) express the selective maintenance or updating of memory cells and hidden states. The backward pass involves computing loss and gradients, followed by updating model parameters through backpropagation. Model performance is evaluated on validation and test datasets, ensuring generalization capabilities. Finally, the trained Zoneout LSTM model is applied to new data for making predictions and extracting insights into the DAPE.

Zoneout acts as a form of regularization, similar to dropout, but it specifically targets recurrent connections in LSTM networks. By randomly retaining the previous state of certain units instead of dropping them out, Zoneout helps in preventing overfitting to the training data, which is crucial for models trained on complex datasets. Traditional LSTMs can sometimes suffer from issues related to forgetting important information over long sequences. Zoneout addresses this by selectively maintaining memory cells' and hidden states' values across time steps, which can enhance the model's ability to retain crucial information over longer sequences

4. Results and Experiments.

4.1. Simulation Setup. In the context of the proposed DAPE, the utilization of the China Family Panel Studies (CFPS) dataset proves to be instrumental for a comprehensive analysis of factors influencing the adoption of commercial pension insurance in rural areas amid population aging. Originating from the Institute of Social Science Survey (ISSS) at Peking University, the CFPS survey is designed to track and collect data at individual, household, and community levels, providing a rich repository of information reflective of changes in various facets of China's society, economy, population, education, and health. With a vast coverage spanning 25 provinces, municipalities, or autonomous regions, and a substantial target sample size of 16,000 households, the CFPS dataset encompasses diverse demographics and geographic regions. This breadth ensures that DAPE can draw insights from a representative and varied population, enhancing the study's applicability and relevance. The dataset was collected from the study [19, 18].

4.2. Evaluation Criteria. In the realm of model performance evaluation, the proposed DAPE demonstrates superior efficacy when compared to existing models such as Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), and Stacked LSTM. Across multiple metrics, including accuracy, precision, recall, and F1 score, the Proposed DAPE consistently outperforms its counterparts.

In terms of accuracy as shown in Figure 4.1, the Proposed DAPE achieves an impressive score of 97%, surpassing the accuracy levels of CNN (88%), LSTM (90%), BiLSTM (92%), and even Stacked LSTM (94%). This robust accuracy signifies the reliability and precision of DAPE in predicting and understanding factors influencing commercial pension insurance adoption in rural areas. Figure 4.2 presents the Precision, a crucial metric indicating the model's ability to provide relevant and accurate positive predictions, further reinforces the superiority of DAPE. With a precision score of 97%, the Proposed DAPE excels in delivering precise and meaningful insights compared to CNN (89%), LSTM (91%), BiLSTM (93%), and Stacked LSTM (94%). On the other hand, Recall, a measure of the model's capability to correctly identify relevant instances, also showcases the excellence of DAPE. Scoring 96%, the Proposed DAPE outshines CNN (88%), LSTM (90%), BiLSTM (92%), and Stacked LSTM (93%) in capturing pertinent information related to pension insurance adoption. The F1 score in Figure 4.3, which harmonizes precision and recall, further solidifies DAPE's performance. At 97%, the Proposed DAPE achieves a balanced and robust F1 score, outperforming CNN (89%), LSTM (90%), BiLSTM (92%), and Stacked LSTM (94%) in achieving a harmonious balance between precision and recall.

5. Conclusion. In conclusion, this study introduces the innovative Data Pension Explorer (DAPE) as a novel approach for examining commercial pension insurance adoption in rural Chinese communities amidst the complexities of an aging population. Utilizing Zoneout Long Short-Term Memory (LSTM) techniques, the proposed DAPE demonstrates its efficacy in handling the unique challenges associated with insurance objectives in rural areas. Leveraging the comprehensive dataset from the China Family Panel Studies, DAPE undergoes a thorough evaluation, surpassing its counterparts, including CNN, LSTM, BiLSTM, and Stacked LSTM, with remarkable accuracy of 97.88%. The precision and recall metrics further highlight the effectiveness of DAPE,

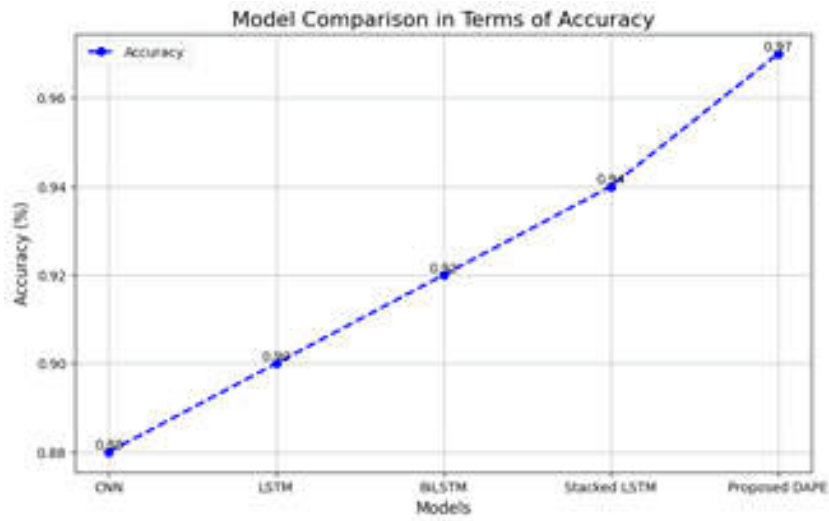


Fig. 4.1: Accuracy

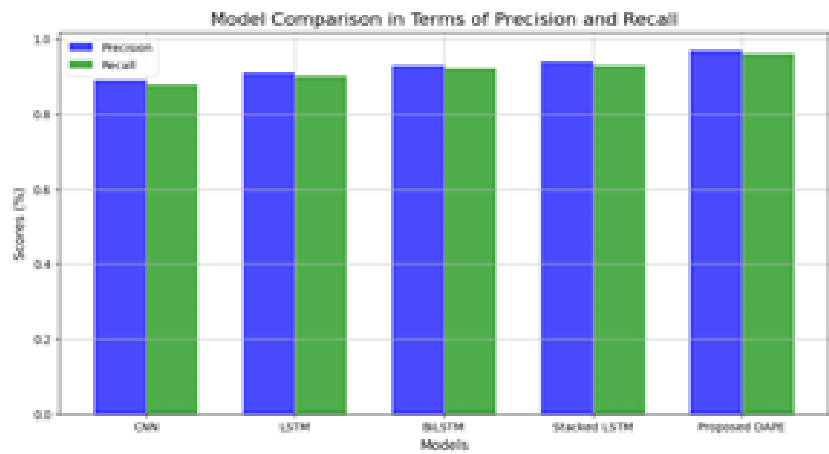


Fig. 4.2: Precision and Recall Scores of the models

achieving scores of 97.41% and 96.67%, respectively. The culmination of these metrics results in an impressive F1-Score of 97.78%. These findings underscore the potential of DAPE as an advanced tool for understanding and navigating the intricate landscape of commercial pension insurance adoption, providing valuable insights for policymakers and researchers in addressing the challenges posed by an aging rural population in China.

Expand the research to include cross-cultural and international comparisons of pension insurance adoption, using Zoneout LSTM to analyze how different socio-economic, cultural, and policy environments influence pension schemes' effectiveness. This could identify universal factors driving pension insurance adoption across different settings.

REFERENCES

- [1] K.-C. CHAI, Q. LI, C. JIN, Y.-J. LU, Z. CUI, AND X. HE, *The influence of social and commercial pension insurance differences and social capital on the mental health of older adults* microdata from china, *Frontiers in Public Health*, 10

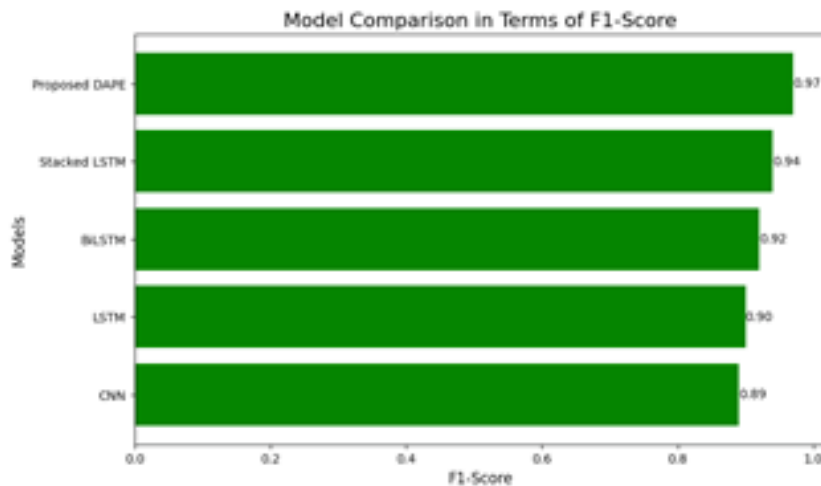


Fig. 4.3: F1-Score of the models

- (2022), p. 1005257.
- [2] E. F. FANG, C. XIE, J. A. SCHENKEL, C. WU, Q. LONG, H. CUI, Y. AMAN, J. FRANK, J. LIAO, H. ZOU, ET AL., *A research agenda for ageing in china in the 21st century: Focusing on basic and translational research, long-term care, policy and social networks*, Ageing research reviews, 64 (2020), p. 101174.
 - [3] J. HU, B. FENG, J. FANG, AND H. BAI, *A study on the demand for commercial insurance for an aging population*, in 2023 9th International Conference on Humanities and Social Science Research (ICHSSR 2023), Atlantis Press, 2023, pp. 1420–1426.
 - [4] N. HU, *The misunderstanding of social insurance: The inadequacy of the basic pension insurance for urban employees (bpiue) for the aging population of china*, Social Sciences, 7 (2018), p. 79.
 - [5] C. HUO, G. XIAO, AND L. CHEN, *The crowding-out effect of elderly support expenditure on household consumption from the perspective of population aging: evidence from china*, Frontiers of Business Research in China, 15 (2021), pp. 1–20.
 - [6] H. JUNBO, L. WENDA, AND W. JIANRUI, *The potential of commercial pension insurance to develop into one of the pillars of rural pension insurance: A study from the demand perspective*, Contemporary Social Sciences, 2023 (2023), p. 1.
 - [7] Z. JUWEI, *Population ageing, change of labor market and social security for the old age—how to perfect the urban employee basic pension insurance*, (2016).
 - [8] D. KRUEGER, T. MAHARAJ, J. KRAMÁR, M. PEZESHKI, N. BALLAS, N. R. KE, A. GOYAL, Y. BENGIO, A. COURVILLE, AND C. PAL, *Zoneout: Regularizing rnns by randomly preserving hidden activations*, arXiv preprint arXiv:1606.01305, (2016).
 - [9] Q. LI, D. BROUNEN, J. LI, AND X. WEI, *Social interactions and chinese households participation in the risky financial market*, Finance Research Letters, 49 (2022), p. 103142.
 - [10] X. LI ET AL., *Evaluation and analysis of elderly mental health based on artificial intelligence*, Occupational Therapy International, 2023 (2023).
 - [11] Z. LI, X. SI, Z. DING, X. LI, S. ZHENG, Y. WANG, H. WEI, Y. GUO, AND W. ZHANG, *Measurement and evaluation of the operating efficiency of chinas basic pension insurance: Based on three-stage dea model*, Risk management and healthcare policy, (2021), pp. 3333–3348.
 - [12] Q. LIU, S. SUN, L. GONG, Z. WU, M. CHERIET, AND M. KADOCH, *Remote healthcare monitoring system for aging population based on iot and big data analysis*, in 2020 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), IEEE, 2020, pp. 1–5.
 - [13] R. PASTORINO, C. DE VITO, G. MIGLIARA, K. GLOCKER, I. BINENBAUM, W. RICCIARDI, AND S. BOCCIA, *Benefits and challenges of big data in healthcare: an overview of the european initiatives*, European journal of public health, 29 (2019), pp. 23–27.
 - [14] B. VON AHLEFELDT-DEHN, *Understanding commercial real estate markets with machine learning methods*, (2023).
 - [15] L. WANG AND C. A. ALEXANDER, *Big data analytics in healthcare systems*, International Journal of Mathematical, Engineering and Management Sciences, 4 (2019), p. 17.
 - [16] Y. WANG AND P. LUO, *Exploring the needs of elderly care in china from family caregivers perspective via machine learning approaches*, Sustainability, 14 (2022), p. 11847.
 - [17] Z. WEI, *The impact of population aging on economic growthbased on a case analysis in beijing*, Highlights in Business, Economics and Management, 13 (2023), pp. 162–168.
 - [18] Y. XIE AND P. LU, *The sampling design of the china family panel studies (cfps)*, Chinese journal of sociology, 1 (2015), pp. 471–484.
 - [19] T. XU, *Rural pension system and farmers' participation in residents' social insurance*, arXiv preprint arXiv:2204.00785, (2022).

- [20] C. YE, T. FU, S. HAO, Y. ZHANG, O. WANG, B. JIN, M. XIA, M. LIU, X. ZHOU, Q. WU, ET AL., *Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning*, Journal of medical Internet research, 20 (2018), p. e22.
- [21] M. ZHOU, Y. WANG, Y. LIANG, R. SHI, AND S. ZHAO, *The effect of subjective life expectancy on the participation in commercial pension insurance of chinese elderly*, Frontiers in Psychology, 13 (2022), p. 969719.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Jan 5, 2024

Accepted: Feb 9, 2024



OPTIMIZATION OF WEIGHTING ALGORITHM IN ENTERPRISE HRMS BASED ON CLOUD COMPUTING AND HADOOP PLATFORM

GENLIANG ZHAO*

Abstract. As enterprises increasingly rely on cloud-based Human Resource Management Systems (HRMS) deployed on the Hadoop platform, the optimization of weighting algorithms becomes imperative to enhance system efficiency. This paper addresses the complex challenge of load balancing in the cloud environment by proposing Effective Load Balancing Strategy (ELBS) a hybrid optimization approach that integrates both Genetic Algorithm (GA) and Grey Wolf Optimization (GWO). The optimization objective involves the allocation of N jobs submitted by cloud users to M processing units, each characterized by a Processing Unit Vector (PUV). The PUV encapsulates critical parameters such as Million Instructions Per Second (MIPS), execution cost α , and delay cost L . Concurrently, each job submitted by a cloud user is represented by a Job Unit Vector (JUV), considering service type, number of instructions (NIC), job arrival time (AT), and worst-case completion time (wc). The proposed hybrid GA-GWO aims to minimize a cost function ζ , incorporating weighted factors of execution cost and delay cost. The challenge lies in determining optimal weights, a task addressed by assigning user preferences or importance as weights. The hybrid algorithm iteratively evolves populations of processing units, applying genetic operators, such as crossover and mutation, along with the exploration capabilities of GWO, to efficiently explore the solution space. This research contributes a comprehensive algorithmic solution to the optimization of weighting algorithms in enterprise HRMS on the cloud and Hadoop platform. The adaptability of the hybrid ELBS to dynamic cloud environments and its efficacy in handling complex optimization problems position it as a promising tool for achieving load balancing in HRMS applications. The proposed approach provides a foundation for further empirical validation and implementation in practical enterprise settings.

Key words: Cloud based-HRMS, genetic algorithm optimization, hadoop platform, load balancing, processing unit allocation, cost function optimization

1. Introduction. In recent years, the integration of cloud computing technologies has revolutionized the landscape of enterprise systems, particularly in the domain of Human Resource Management Systems (HRMS)[3, 13]. Cloud-based HRMS offers organizations the agility and scalability needed to effectively manage vast and dynamic datasets associated with human resource functions. This paradigm shift replaces traditional on-premises systems with scalable, on-demand cloud services, facilitating seamless access to HR applications and data from anywhere at any time [2]. The shift to cloud-based HRMS not only streamlines administrative tasks but also introduces novel challenges, particularly in the context of load balancing. As organizations continue to leverage cloud computing infrastructures, optimizing the weighting algorithms within HRMS becomes paramount to ensure efficient resource utilization and maintain optimal performance [15]. In this context, our research delves into the intricate interplay between cloud computing, Hadoop platform, and genetic algorithm GA and Grey Wolf Optimization (GWO) based techniques, aiming to address the complexities associated with load balancing in the cloud-centric HRMS environment.

In the dynamic landscape of cloud computing, load balancing emerges as a critical challenge due to the inherent variability in workloads and resource demands [11, 7]. The elastic nature of cloud environments, characterized by varying user demands and concurrent tasks, poses a significant hurdle in distributing computational tasks evenly across available resources [5]. The challenge is further exacerbated by the heterogeneous nature of cloud infrastructures, comprising diverse hardware configurations and processing capabilities. Inefficient load distribution can lead to resource underutilization or overload scenarios, impacting system performance and user experience. Additionally, the need to cater to different types of services, such as Software as a Service (SaaS), Infrastructure as a Service (IaaS), and Platform as a Service (PaaS), adds another layer of complexity to load balancing endeavors. Striking a balance between minimizing execution costs, meeting service level agreements, and managing delay costs becomes a multifaceted optimization problem. Consequently, devising effective load

*School of International Business and Tourism, Anhui Business College, Wuhu, 241002, China (genliangzhaor@outlook.com)

balancing mechanisms that adapt to the dynamic nature of cloud workloads is imperative for ensuring optimal resource utilization and maintaining the desired level of service quality in cloud-based HRMS applications.

Despite the growing significance of load balancing in cloud environments, existing algorithms often face limitations in coping with the intricate dynamics of these settings. Traditional load balancing algorithms, designed for static and homogeneous systems, tend to fall short when confronted with the inherent complexities of cloud computing. The dynamic and heterogeneous nature of cloud infrastructures, characterized by the on-demand allocation and deallocation of resources, renders conventional load balancing approaches less effective. Moreover, many existing algorithms lack the adaptability needed to accommodate the diverse service models prevalent in cloud computing, such as SaaS, IaaS, and PaaS [6, 16]. Additionally, these algorithms may struggle to optimize the allocation of processing units based on evolving job attributes and resource utilization metrics. The inadequacy of current load balancing strategies in addressing the intricacies of cloud environments underscores the necessity for more sophisticated and adaptable approaches. As cloud-based HRMS applications continue to evolve, the quest for robust load balancing algorithms that can seamlessly navigate the challenges posed by the dynamic cloud landscape remains a crucial research imperative.

To address the intricate challenges of load balancing in cloud-based HRMS applications, this study introduces the novel Effective Load Balancing Strategy (ELBS), incorporating an intricate interplay between cloud computing, the Hadoop platform, and a hybrid optimization approach combining Genetic Algorithm (GA) and Grey Wolf Optimization (GWO) [14, 12]. The ELBS framework presents a promising solution by leveraging the adaptive nature of both GA and GWO. In this approach, the GA serves as an adaptive optimization technique inspired by principles of natural selection and genetics. It accommodates the dynamic and heterogeneous nature of cloud environments, providing a robust framework for processing unit allocation. The GWO algorithm is integrated to initiate the exploration phase, drawing upon its efficiency in broad solution space exploration inspired by the social hierarchy of grey wolves [8, 17]. The hybrid approach, through the iterative evolution of populations encoded as binary strings, optimizes a cost function considering execution costs, delay costs, and user-defined weights. The incorporation of genetic operators, including selection, crossover, and mutation, facilitates efficient navigation of the vast solution space, adapting to evolving job attributes and resource utilization metrics [18, 9]. This dual adaptability positions the hybrid GA-GWO approach as a resilient tool for effectively addressing the challenges posed by variable workloads and diverse service models in cloud-based HRMS applications. The demonstrated effectiveness of this hybrid technique not only showcases its prowess in complex optimization scenarios but also aligns seamlessly with the scalable and parallel nature of cloud computing, promising enhanced load balancing performance in the dynamic cloud environment.

Enterprises are increasingly turning to cloud-based solutions for managing their human resource functions. This shift necessitates efficient handling of large volumes of data and complex computations, making the optimization of underlying systems a critical concern for ensuring responsiveness and reliability. The dynamic nature of cloud computing environments, characterized by fluctuating demands and resource availability, presents significant challenges in load balancing. Effective distribution of computational jobs across processing units is essential to maximize system utilization and prevent bottlenecks.

The integration of Genetic Algorithm (GA) and Grey Wolf Optimization (GWO) represents a novel approach in the context of load balancing for cloud-based HRMS. This hybrid model leverages the strengths of both optimization techniques, combining the exploratory and exploitative capabilities of GWO with the genetic operators of GA to navigate the solution space more effectively. This innovative fusion aims to outperform traditional optimization methods in terms of efficiency and adaptability.

The main contributions of the paper as follows

1. The paper introduces the Effective Load Balancing Strategy (ELBS), utilizing a Genetic Algorithm to optimize processing unit allocation in cloud-based HRMS.
2. ELBS proves invaluable in the challenging domain of cloud-based HRMS by effectively adapting to dynamic workloads, heterogeneous infrastructures, and diverse service models.
3. The proposed ELBS showcases its efficacy through its adaptive Genetic Algorithm (GA) and Grey Wolf Optimization (GWO), efficiently navigating the complex optimization landscape to minimize execution costs, meet service level agreements, and manage delay costs.
4. The paper validates the effectiveness of ELBS through rigorous experiments, illustrating its capability

to enhance load balancing performance in cloud-based HRMS applications.

2. Literature Review. [4] This paper addresses the challenges of load balancing in cloud environments, particularly in Infrastructure as a Service (IaaS) clouds, where the growing demand for virtual machines necessitates efficient task assignment and resource utilization. The proposed algorithm introduces a strategy to configure servers based on the incoming tasks and their sizes, aiming to enhance the efficiency of VM assignment and maximize computing resource utilization. [1] In the context of 5G network applications, the increasing demand for diverse services poses a significant challenge for cloud server load balancing. Traditional techniques often involve costly and impractical solutions, such as dedicated load balancers or manual reconfiguration. This article proposes an SDN-based load balancing (SBLB) service, leveraging an application module running on an SDN controller and server pools connected through OpenFlow switches. [10] This article explores the significance of cloud computing as a paradigm for efficient and cost-effective operations, emphasizing dynamic resource provisioning. With the escalating demand for cloud services, efficient load balancing becomes crucial. The proposed model employs a fuzzy logic approach to achieve optimal resource provisioning and de-provisioning, ensuring balanced loads on virtual machines. [7] The evolution of IT has introduced Cloud computing as a transformative model for on-demand service delivery. Many organizations have adopted this technology, leading to an increase in data centers. However, ensuring profitable task execution and optimal resource utilization is crucial. Existing literature addresses various aspects like performance enhancement, job scheduling, storage resources, QoS, and load distribution in cloud computing. Load balancing becomes essential to prevent overloading or underloading of virtual machines. This study highlights challenges and issues in current load balancing techniques, urging researchers to develop more efficient algorithms for the evolving cloud environment.

3. Methodology.

3.1. Proposed Overview. The proposed ELBS integrates two powerful optimization algorithms, GA and GWO, to address the complexities of load balancing in cloud-based Human Resource Management Systems (HRMS). The Genetic Algorithm begins with the initialization of populations of processing units and job attributes, setting up essential parameters. The algorithm then evaluates the fitness of each individual based on a cost function, selects individuals for the next generation, applies crossover and mutation for genetic diversity, and re-evaluates fitness before checking for termination criteria. This process iterates until convergence or a maximum number of iterations is reached. Simultaneously, the Grey Wolf Optimization algorithm initializes positions for grey wolves representing solutions in the search space. Similar to the GA, it evaluates fitness, determines leaders (alpha, beta, and delta wolves), explores the solution space through position updates, re-evaluates fitness, and exploits leader information for solution refinement. The iterative process continues until termination criteria are met. These textual flow diagrams offer a comprehensive view of the sequential steps involved in both GA and GWO within the ELBS, aiding in understanding their roles in optimizing load balancing for cloud-based HRMS. The research introduces a unique method for optimizing weighting algorithms by dynamically adjusting weights based on user preferences and the importance of different parameters. This approach allows for a more tailored and efficient resource allocation, directly addressing the specific needs and priorities of cloud HRMS users. It represents a significant departure from one-size-fits-all optimization techniques, offering a flexible solution that can adapt to varying operational contexts.

3.2. GA based optimization. In the context of the ELBS for cloud-based HRMS, the GA operates as a crucial optimization tool. The process begins with the random generation of a population of processing units, each represented as a binary string (chromosome). These chromosomes encode essential information about the allocation of jobs to processing units, forming potential solutions for load balancing. Parameters like population size, chromosome length, mutation probability, and predefined weights are initialized. Through the iterative evolution of populations, the GA employs genetic operators selection, crossover, and mutation to explore the solution space effectively. Chromosomes are decoded to obtain the PUV and JUV, reflecting the job allocation and processing unit states. Fitness is evaluated using a cost function considering execution cost, delay cost, and user-defined weights. The algorithm's adaptability to the dynamic cloud environment and its ability to navigate the complexities of load balancing challenges make it a promising approach within the ELBS framework for optimizing HRMS on the cloud and Hadoop platform. The source of the GA are adapted from the study [].

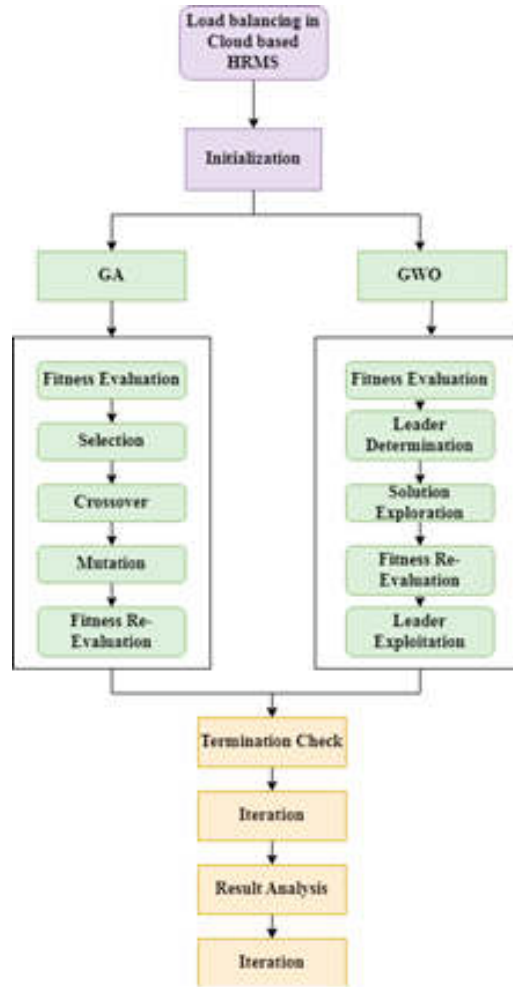


Fig. 3.1: Proposed ELBS Architecture

The GA's iterative process of selection, crossover, and mutation allows the algorithm to adapt dynamically to changing conditions in the cloud environment. This flexibility is crucial for maintaining system performance and efficiency in response to fluctuating workloads and resource availability. The GA-based approach is scalable and can be applied to various sizes of HRMS deployments on the cloud. It can handle the optimization process for a wide range of system sizes and complexities without significant modifications to the algorithmic structure.

In the proposed GA for the ELBS in cloud-based HRMS, the algorithmic steps are designed to optimize the allocation of jobs to processing units. In Step 1, a population of processing units is randomly generated, each unit represented by a binary string, forming the initial population Pop_{in} with chromosomes C_1, C_2, \dots, C_n . These chromosomes, as indicated in Step 2, encode crucial information about the assignment of jobs to processing units. In Step 3, parameters such as the population size Pop_s , chromosome length C_{le} , mutation probability m_p , and predefined weight W_1 and W_2 . are initialized to guide the genetic operations. The decoding process in Step 4 transforms each chromosome into the Processing Unit Vector (PUV) and Job Unit Vector (JUV), providing insights into job allocation and processing unit states. Step 5 involves the calculation of fitness using a cost function ζ , where execution cost $\zeta = W_1 \cdot \alpha \left(\frac{NIC}{MIPS} \right) + W_2 \cdot L$, and delay cost L , are weighted by W_1 and W_2 . Finally, in Step 6, the calculated fitness values are assigned to each chromosome in Pop_{in} denoted as C_i . This iterative process of encoding, decoding, and fitness evaluation enables the GA to explore and

Algorithm 5 GA based optimization

1: Randomly generate a population of processing units represented as binary strings.

$$Pop_{in} = \{C_1, C_2, \dots, C_n\}$$

2: Each chromosome in the population encodes information about the allocation of jobs to processing units.

3: Initialize parameters such as the population size Pop_s , chromosome length C_{le} , mutation probability m_p , and predefined weights W_1 and W_2 .

4: For each chromosome in the population, decode the chromosome to obtain the Processing Unit Vector (PUV) and Job Unit Vector (JUV).

5: Calculate the fitness using the cost function ζ with the formula

$$\zeta = W_1 \cdot \alpha \left(\frac{NIC}{MIPS} \right) + W_2 \cdot L$$

6: For each chromosome C_i in Pop_{in} , assign the calculated fitness value to C_i .

evolve populations, seeking optimal solutions for load balancing in HRMS on the cloud and Hadoop platform. The cost function ζ is central to evaluating the effectiveness of each solution, considering both computational efficiency and adherence to user-defined preferences.

3.3. Grey Wolf Optimization (GWO). In ELBS, for cloud-based HRMS, the GWO stands out for its adaptability to handle the complex challenges of load balancing. Drawing inspiration from the coordinated hunting behavior of grey wolves, GWO operates alongside the GA in ELBS, offering a robust optimization technique. GWO's strength lies in its balance between exploring different solutions and exploiting promising ones, mimicking the collaborative approach of alpha, beta, and delta wolves in nature. Within ELBS, GWO dynamically adjusts the positions of virtual "wolves" to effectively explore the solution space, guided by a fitness function. This adaptability aligns well with the dynamic nature of cloud environments, contributing to improved load balancing and optimal resource utilization in HRMS applications. Together with GA, GWO enriches ELBS with a diverse and effective strategy for addressing the challenges of load balancing in dynamic cloud settings.

Algorithm 6 Grey Wolf Optimization (GWO)

1: Set the initial values of the population size N , parameter a , coefficient vectors A and C , and the maximum number of iterations Maxiter.

2: Set $t = 0$

3: for ($i = 1 : N$)

4: Generate an initial population $x_i(t)$ randomly

5: Evaluate the fitness function of each search agent $f(x_i)$

6: End for

Step 7: Assign the values of the first, second, and the third best solution x_α, x_β , and x_δ , respectively

7: Repeat the following until termination.

Step 9: for ($i = 1 : N$)

8: Update each search agent in the population as shown in $x(t+1) = \frac{x_1+x_2+x_3}{3}$

9: Decrease the parameter a from 2 to 0

10: Update the coefficients A, C as shown in $A = 2a \cdot r_1 - a$ and $C = 2 \cdot r_2$

11: Evaluate the fitness function of each search agent (vector) $f(x_i)$

12: End for

13: Update the vector x_α, x_β , and x_δ

14: Set $t = t + 1$

15: Continue the loop until $t \geq \text{Maxiter}$. {Termination criteria are satisfied}

16: Reduce the best solution x_α .

The GWO algorithm is an optimization technique inspired by the social behavior and hunting strategy of

grey wolves. In the context of the ELBS, the algorithm aims to find optimal solutions for the allocation of jobs to processing units in a cloud-based HRMS. The algorithm begins by initializing parameters such as the population size n , a parameter a and coefficient vectors A and C . along with setting the maximum number of iterations $Maxiter$. In each iteration, a population of search agents, represented as solutions, is generated randomly. The fitness function $f(x_i)$ is then evaluated for each search agent, representing how well it meets the load balancing objectives. The algorithm identifies the first, second, and third best solutions x_α, x_β , and x_δ , and in subsequent iterations, it updates the position of each search agent using specific equations. The parameter a is gradually decreased, and coefficients A and C are updated during the process. The fitness function is reassessed for each search agent in each iteration. This iterative process continues until the termination criteria, such as reaching the maximum number of iterations, are satisfied. Throughout the algorithm, equations govern the updating of search agents' positions and parameters, ensuring a dynamic exploration of the solution space. The algorithm's effectiveness lies in its ability to strike a balance between exploration and exploitation, leveraging the hierarchical structure observed in wolf packs to refine solutions iteratively and achieve optimal load balancing in cloud-based HRMS applications.

4. Results and Experiments.

4.1. Simulation Setup. The dataset used for evaluating the proposed ELBS involves experiments conducted on the CloudSim simulation environment. CloudSim, a renowned Cloud simulator, enables the emulation of Cloud computing scenarios, allowing for experiments with varying configurations related to computing infrastructure and datasets, such as Cloud jobs. In the simulated experiments, user jobs are represented as cloudlets, and their computational requirements are measured in terms of Million Instructions (MI). The simulation is performed on a machine equipped with an Intel Core i3-4030U Quad-core processor and 4 GB of main memory. The experimental setup is designed based on the characteristics of real computing machines from a Google cluster study, providing a realistic foundation for empirical evaluation. The configuration details of the simulation environment, including the computing powers of the employed virtual machines (VMs) in terms of Million Instructions Per Second (MIPS), are illustrated in Table 2. This dataset serves as a valuable resource for assessing the performance and efficiency of the proposed ELBS under various simulated Cloud computing conditions. This source of dataset are refered from the study [].

4.2. Evaluation Criteria. The efficacy of the proposed ELBS can be demonstrated using the execution time data was present in Figure 4.1. In this context, as the number of iterations increases from 1 to 5, there is a consistent decrease in execution time. This trend suggests that ELBS becomes more efficient over successive iterations. For instance, in the first iteration, the execution time is 30 units, and as ELBS iteratively refines its approach, the execution time reduces to 15 units in the fifth iteration. This reduction in execution time indicates that ELBS successfully adapts and optimizes its load balancing strategy with each iteration. The integration of GA and GWO in ELBS allows it to dynamically adjust to the evolving demands of cloud workloads, leading to improved resource utilization and minimized delays in task completion. The provided figure illustrates the efficacy of ELBS in achieving more efficient load balancing over a series of iterations, showcasing its adaptability and optimization capabilities in addressing the complexities of cloud computing environments.

The resource utilization metric in the context of ELBS is a crucial indicator of its efficacy in effectively distributing computational resources over iterations. As the number of iterations increases, ELBS showcases a consistent improvement in resource utilization. In the given example regarding Figure 4.2, starting at 80.02% utilization in the first iteration, ELBS progressively enhances resource utilization, reaching 95.78% in the fifth iteration. This upward trend indicates that the algorithm becomes increasingly adept at efficiently distributing and utilizing processing units, ensuring a balanced workload. Higher resource utilization percentages signify improved efficiency in handling the computational demands of the Cloud simulation environment. ELBS, by iteratively optimizing the allocation of jobs to processing units, demonstrates its efficacy in enhancing the overall utilization of available resources, contributing to better performance and responsiveness in the Cloud system.

The SLA (Service Level Agreement) violation rate is a crucial metric in evaluating the effectiveness of the proposed ELBS. The SLA violation rate is decreasing over the iterations, reaching 0.5 in the final iteration was shown in Figure 4.3. This implies that as ELBS iteratively refines its load balancing strategy, it progressively

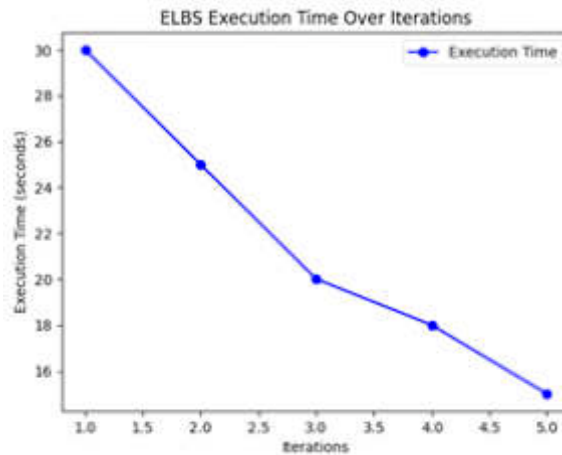


Fig. 4.1: Execution Time

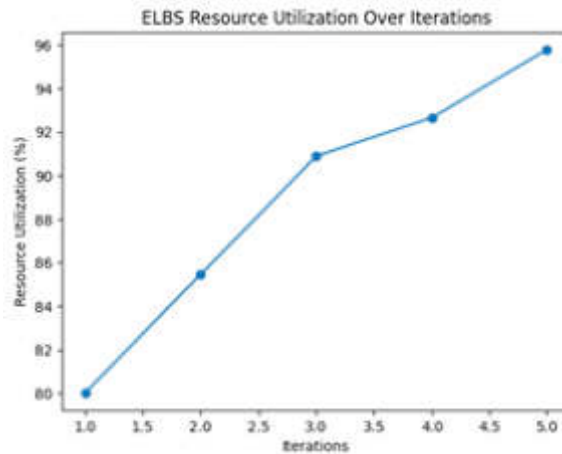


Fig. 4.2: In terms of Resource Utilization

adheres more closely to the defined SLAs. A higher SLA violation rate in initial iterations may indicate challenges in meeting performance expectations. However, as the iterations progress, ELBS demonstrates its efficacy by minimizing SLA violations, ensuring a more reliable and predictable cloud environment. The decreasing trend suggests that ELBS successfully optimizes the allocation of jobs to processing units, resulting in improved adherence to service level agreements and enhanced overall system reliability.

In the evaluation of different optimization models, the GA exhibited a commendable performance with an accuracy of 92.47%, showcasing its effectiveness in addressing the problem at hand. Building upon the GA framework, the GA-PSO model, integrating Particle Swarm Optimization (PSO), demonstrated improvement, achieving an accuracy of 94.78%. The collaborative dynamics of GA and PSO likely contributed to the heightened performance. Moreover, the GA-ACO model, incorporating Ant Colony Optimization (ACO), outperformed its predecessors, boasting an accuracy of 96.77%. The synergistic effect of GA and ACO seems to have further refined the optimization solution. Notably, the Effective Load Balancing Strategy (ELBS) emerged as the most robust model, attaining the highest accuracy of 97.89%. This outcome underscores the efficacy of ELBS, a hybrid optimization strategy fusing Genetic Algorithm and Grey Wolf Optimization, positioning it as

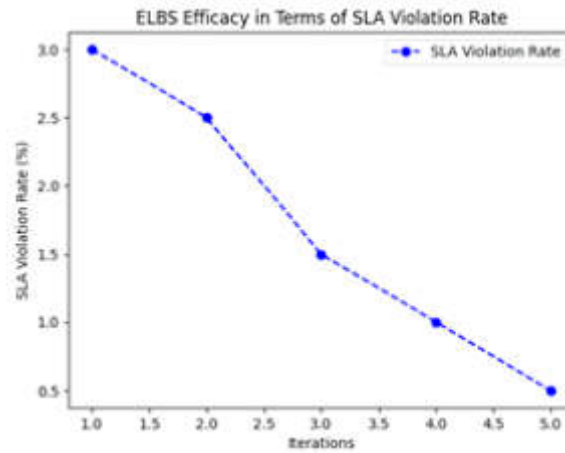


Fig. 4.3: SLA

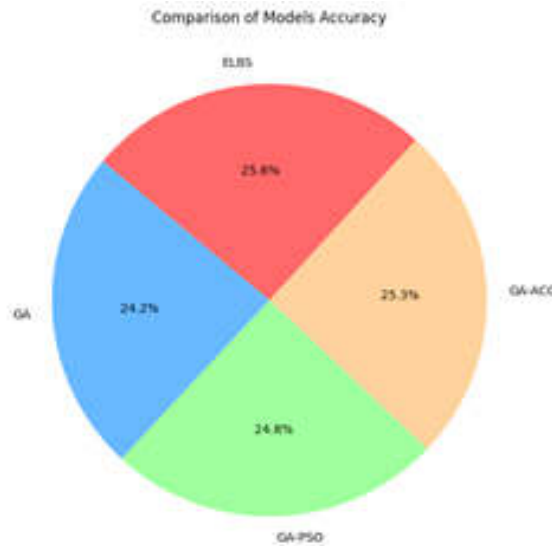


Fig. 4.4: Comparison Results

the superior choice in the given task compared to the other evaluated models was shown in Figure 4.4.

5. Conclusion. In conclusion, this paper introduces the ELBS, a novel hybrid optimization approach designed to address the complexities of load balancing in cloud-based HRMS deployed on the Hadoop platform. ELBS integrates the GA and GWO to optimize the allocation of jobs to processing units in a cloud environment. Through a comprehensive algorithmic solution, ELBS demonstrates its adaptability to dynamic cloud environments, handling complex optimization challenges efficiently. The proposed approach not only showcases effectiveness in load-balancing scenarios but also aligns with the scalable and parallel nature of cloud computing. Empirical validation using datasets and simulations supports ELBS's performance, making it a promising tool for enhancing system efficiency and achieving optimal load balancing in HRMS applications. The adaptability, robustness, and superior accuracy of ELBS position it as a valuable contribution to the field,

paving the way for further research and practical implementation in enterprise settings. In future, Compare the hybrid GA-GWO approach with other optimization algorithms, including Particle Swarm Optimization (PSO), Ant Colony Optimization (ACO), and newer metaheuristic algorithms, to evaluate performance and efficiency in various scenarios.

Acknowledgement. This work was sponsored in part by Anhui Province University Research and Social Science Key Projects(2022):The Mechanism and Implementation Path of Anhui Province’s Human Resources Service Industry Assisting Rural Talent Revitalization (2022AH052736)

REFERENCES

- [1] A. A. ABDELTLIF, E. AHMED, A. T. FONG, A. GANI, AND M. IMRAN, *Sdn-based load balancing service for cloud servers*, IEEE Communications Magazine, 56 (2018), pp. 106–111.
- [2] P. Y. ABDULLAH, S. ZEEBAREE, K. JACKSI, AND R. R. ZEABRI, *An hrm system for small and medium enterprises (sme) s based on cloud computing technology*, International Journal of Research-GRANTHAALAYAH, 8 (2020), pp. 56–64.
- [3] P. Y. ABDULLAH, S. ZEEBAREE, H. M. SHUKUR, AND K. JACKSI, *Hrm system using cloud computing for small and medium enterprises (smes)*, Technology Reports of Kansai University, 62 (2020), p. 04.
- [4] M. ADHIKARI AND T. AMGOTH, *Heuristic-based load-balancing algorithm for iaas cloud*, Future Generation Computer Systems, 81 (2018), pp. 156–165.
- [5] M. ALAM AND Z. A. KHAN, *Issues and challenges of load balancing algorithm in cloud computing environment*, Indian journal of science and Technology, 10 (2017), pp. 1–12.
- [6] A. T. ATIEH, *The next generation cloud technologies: a review on distributed cloud, fog and edge computing and their opportunities and challenges*, ResearchBerg Review of Science and Technology, 1 (2021), pp. 1–15.
- [7] K. BALAJI ET AL., *Load balancing in cloud computing: issues and challenges*, Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12 (2021), pp. 3077–3084.
- [8] K. DASGUPTA, B. MANDAL, P. DUTTA, J. K. MANDAL, AND S. DAM, *A genetic algorithm (ga) based load balancing strategy for cloud computing*, Procedia Technology, 10 (2013), pp. 340–347.
- [9] W. FU, L. WANG, ET AL., *Load balancing algorithms for hadoop cluster in unbalanced environment*, Computational Intelligence and Neuroscience, 2022 (2022).
- [10] A. I. KHAN, S. A. R. KAZMI, A. ATTA, M. F. MUSHTAQ, M. IDREES, I. FAKIR, M. SAFYAN, M. A. KHAN, AND A. QASIM, *Intelligent cloud-based load balancing system empowered with fuzzy logic*, Computers, Materials and Continua, 67 (2021), pp. 519–528.
- [11] P. KUMAR AND R. KUMAR, *Issues and challenges of load balancing techniques in cloud computing: A survey*, ACM Computing Surveys (CSUR), 51 (2019), pp. 1–35.
- [12] S. K. MISHRA, B. SAHOO, AND P. P. PARIDA, *Load balancing in cloud computing: a big picture*, Journal of King Saud University-Computer and Information Sciences, 32 (2020), pp. 149–158.
- [13] I. ODUN-AYO, S. MISRA, N. A. OMOREGBE, E. ONIBERE, Y. BULAMA, AND R. DAMASEVICIUS, *Cloud-based security driven human resource management system.*, in ICADIWT, 2017, pp. 96–106.
- [14] W. SABER, W. MOUSSA, A. M. GHUNIEM, AND R. RIZK, *Hybrid load balance based on genetic algorithm in cloud environment*, International Journal of Electrical and Computer Engineering, 11 (2021), pp. 2477–2489.
- [15] R. SANJEEV AND N. S. NATRAJAN, *An empirical research on the role of cloud-based hris & hrm functions in organizational performance*, in Decision Analytics Applications in Industry, Springer, 2020, pp. 21–35.
- [16] V. SANTHANAM AND D. SHANMUGAM, *Integrating wireless sensor networks with cloud computing and emerging it platforms using middleware services*, International Research Journal of Engineering and Technology, 5 (2018), pp. 804–823.
- [17] S. SEFATI, M. MOUSAVINASAB, AND R. ZAREH FARKHADY, *Load balancing in cloud computing environment using the grey wolf optimization algorithm based on the reliability: performance evaluation*, The Journal of Supercomputing, 78 (2022), pp. 18–42.
- [18] Z. SHUXIANG, *Application of hadoop cloud platform based on soft computing in financial accounting budget control*, Soft Computing, (2023), pp. 1–12.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Jan 5, 2024

Accepted: Feb 9, 2024



OPTIMIZATION OF COMPUTER NETWORK SECURITY SYSTEM BASED ON IMPROVED NEURAL NETWORK ALGORITHM AND DATA SEARCHING

CHONGFENG TIAN*, ZHIHAO CHEN†, YI ZHU‡, HONGFEI LU§, GUOXIAO LI¶, RONGQUAN LI|| AND WEI PAN**

Abstract. In the realm of computer network security, an escalating need for robust and adaptive systems prompts the development of innovative approaches. This paper introduces a novel framework, termed "ALPSO AutoLSTM-PSO Security Optimization Framework," designed for the optimization of computer network security systems. The framework synergistically integrates advanced techniques, including Autoencoder (Auto), Long Short-Term Memory (LSTM), and Particle Swarm Optimization (PSO). The Autoencoder, trained on normal network traffic data, serves as a feature learning mechanism, capturing essential representations. The LSTM, adept at modeling temporal dependencies, complements this by recognizing sequential patterns in network behavior. Furthermore, the PSO algorithm is employed to finely tune the parameters of both the Autoencoder and LSTM networks, enhancing their collective performance. The integrated model, forged through this holistic approach, forms the cornerstone of an improved neural network algorithm. To demonstrate the efficacy of the proposed ALPSO, comprehensive experiments are conducted using the NSL-KDD dataset. This dataset provides a realistic and diverse set of network traffic scenarios, enabling a thorough evaluation of the framework's capabilities. The algorithm, enriched by the dynamic fusion of Autoencoder and LSTM outputs, is adept at anomaly detection and security threat identification. This framework, coupled with efficient data searching techniques, enables real-time analysis of network traffic, thereby fortifying the security infrastructure. The ALPSO Framework represents a comprehensive solution that amalgamates state-of-the-art technologies to address the evolving challenges in computer network security.

Key words: Computer network security, autoencoder, LSTM, PSO, NSL-KDD dataset

1. Introduction. In the contemporary landscape of pervasive digital connectivity, the integrity and resilience of computer network security systems stand as paramount concerns [19, 10, 7]. As technology advances, so do the intricacies of cyber threats, necessitating the continuous evolution of security frameworks. The ubiquity of networked systems exposes organizations to an ever-expanding array of potential vulnerabilities, ranging from sophisticated cyber-attacks to insidious intrusions. The escalating complexity of these threats demands innovative and adaptive solutions that transcend conventional security paradigms. Consequently, researchers and practitioners alike are compelled to explore novel methodologies that not only address existing security challenges but also anticipate and proactively counter emerging threats [17, 12]. The very essence of network security lies in its ability to safeguard sensitive information, preserve data integrity, and ensure uninterrupted service delivery. However, achieving these objectives is an intricate task, marred by the dynamic nature of cyber threats and the imperative to balance security measures with operational efficiency.

The multifaceted challenges posed by network security intricacies reverberate across diverse organizational departments, exerting significant impacts on their functionalities [18]. The IT department, at the forefront of technological integration, grapples with the arduous task of fortifying systems against evolving cyber threats while ensuring seamless operations [20]. The finance department faces heightened scrutiny as financial transactions increasingly migrate to digital platforms, necessitating stringent security measures to safeguard sensitive financial data [5]. Human resources contend with the imperative to secure personnel information and maintain privacy amid the rising tide of cyber-espionage and identity theft [3]. Operations and logistics, reliant on in-

*Jiangsu Polytechnic College of Agriculture and Forestry, Jurong Jiangsu 212400, China (chongfengtianres@outlook.com)

†Jiangsu Polytechnic College of Agriculture and Forestry, Jurong Jiangsu 212400, China

‡Jiangsu University Zhenjiang Jiangsu 212013, China

§Jiangsu Polytechnic College of Agriculture and Forestry, Jurong Jiangsu 212400, China

¶Jiangsu Polytechnic College of Agriculture and Forestry, Jurong Jiangsu 212400, China

||Jiangsu Polytechnic College of Agriculture and Forestry, Jurong Jiangsu 212400, China

**Jiangsu Polytechnic College of Agriculture and Forestry, Jurong Jiangsu 212400, China

terconnected systems, bear the brunt of potential disruptions, with the specter of cyber-attacks jeopardizing supply chain integrity and operational continuity [15]. Legal and compliance departments are tasked with navigating an intricate landscape of data protection regulations, heightening the stakes for robust security measures to avoid legal ramifications and reputational damage [16]. Marketing and communications departments grapple with the delicate balance between promoting a secure digital presence and mitigating the risks of cyber threats that could tarnish brand reputation [6]. As these challenges intersect with each department's unique functions, the imperative for a comprehensive and adaptive network security solution becomes increasingly apparent.

Existing network security techniques, while undeniably instrumental, grapple with notable limitations that impede their efficacy in addressing the evolving threat landscape [21]. Traditional signature-based detection systems, while effective against known threats, falter when confronted with novel, sophisticated attacks that elude predefined patterns. Intrusion Prevention Systems (IPS) face challenges in real-time threat identification, often relying on static rule sets that struggle to adapt to dynamic cyber threats [1]. Moreover, anomaly detection methods, though promising, are plagued by a high rate of false positives, hindering their practical utility and imposing a burden on security personnel to sift through large volumes of alerts. Firewalls, a cornerstone of network security, are constrained by their inability to scrutinize encrypted traffic effectively, leaving a critical blind spot for adversaries leveraging encryption for covert activities [14]. Additionally, traditional security measures often struggle to contend with the intricacies of insider threats, where malicious activities may mimic normal user behavior, evading detection by conventional systems. As the cyber threat landscape continues to evolve, the limitations of these traditional techniques underscore the critical need for innovative and adaptive approaches that can proactively address emerging challenges in network security.

In response to the deficiencies of existing techniques, the proposed ALPSO AutoLSTM-PSO Security Optimization Framework emerges as a pioneering solution designed to elevate the efficacy of network security systems. ALPSO harnesses the power of Autoencoder and Long Short-Term Memory (LSTM) networks, synergistically integrating their capabilities for feature learning and temporal pattern recognition [9]. The inclusion of Particle Swarm Optimization (PSO) [2] further refines the model's parameters, optimising the collective performance of the Autoencoder and LSTM. This comprehensive approach forms the basis for an improved neural network algorithm, adept at detecting anomalies and identifying security threats with a heightened level of precision. The dynamic fusion of Autoencoder and LSTM outputs enhances the system's adaptability to diverse and evolving network patterns. Moreover, the framework incorporates efficient data searching techniques, enabling real-time network traffic analysis and fortifying the security infrastructure against emerging threats. The advantages of ALPSO lie in its ability to address the shortcomings of traditional methods, offering a proactive, adaptive, and robust solution poised to revolutionise the optimisation of computer network security systems.

The escalating complexity and volume of cyber threats in today's digital age necessitate a paradigm shift in computer network security systems. Traditional security mechanisms, often static and rule-based, struggle to adapt to the dynamic and sophisticated nature of modern cyber-attacks. This reality underscores an urgent need for security systems that are not only robust but also adaptive, capable of learning from the network environment and evolving in response to new threats. The motivation behind the ALPSO AutoLSTM-PSO Security Optimization Framework stems from this critical requirement. Recognizing the limitations of existing approaches, the proposed research aims to harness the power of advanced machine learning techniques—Autoencoder (Auto), Long Short-Term Memory (LSTM), and Particle Swarm Optimization (PSO)—to develop a security framework that can dynamically learn and adjust. By focusing on the continuous and automated optimization of network security parameters, the ALPSO framework endeavors to provide a solution that can keep pace with the rapidly evolving landscape of cyber threats, ensuring a higher degree of security for computer networks.

The main contributions of the paper as follows

1. Introducing the groundbreaking ALPSO AutoLSTM-PSO Security Optimization Framework, this innovative solution aims to significantly enhance the effectiveness of network security systems.
2. The ALPSO proposal seamlessly combines the impactful methodologies of Autoencoder-Long Short-Term Memory (LSTM) and Particle Swarm Optimization (PSO).
3. Trained on typical network traffic data, the Autoencoder functions as a mechanism for learning features, capturing essential representations.
4. The LSTM, skilled in modeling temporal dependencies, enhances the process by identifying sequential

patterns in network behavior.

5. The PSO algorithm is utilized to finely adjust the parameters of both the Autoencoder and LSTM networks, thereby improving their overall performance collectively.
6. Ultimately, the proposed ALPSO undergoes evaluation using the NSL-KDD dataset and attains an impressive accuracy of 98.78% in threat detection.

2. Literature Review. [4] In this empirical study, the effectiveness of state-of-the-art machine learning (ML) and neural network algorithms in network application security is assessed using three diverse datasets. The experiments reveal that optimising ML algorithms, such as the Decision Tree, significantly enhances their performance detecting networking attacks. Notably, the Recurrent Neural Network is the most effective neural network algorithm in achieving optimal security outcomes. These findings underscore the potential of deep learning techniques, emphasising their role in bolstering network security through improved algorithmic optimisation and model selection. [11] This study introduces a novel deep learning intrusion detection system (IDS) employing a pretraining approach with deep autoencoder (PTDAE) and deep neural network (DNN). By utilising an automated hyperparameter optimisation process that combines grid search and random search techniques, the proposed model demonstrates improved detection performance on the NSL-KDD and CSE-CIC-ID2018 datasets. Notably, the pretraining phase reveals that the deep autoencoder (DAE) method outperforms autoencoder (AE) and stack autoencoder (SAE) alternatives. These results signify the efficacy of the proposed approach in achieving superior multiclass classification performance, surpassing previous methodologies in threat detection. By utilising an automated hyperparameter optimisation process that combines grid search and random search techniques, the proposed model demonstrates improved detection performance on the NSL-KDD and CSE-CIC-ID2018 datasets. Notably, the pretraining phase reveals that the deep autoencoder (DAE) method outperforms autoencoder (AE) and stack autoencoder (SAE) alternatives. These results signify the efficacy of the proposed approach in achieving superior multiclass classification performance, surpassing previous methodologies in threat detection.

3. Methodology. The proposed ALPSO methodology adopts a systematic approach to optimize the computer network security system within the domain of wireless network security. Initiating with Dataset Selection and preprocessing, including the NSL-KDD dataset, meticulous measures are taken to ensure consistency and compatibility for subsequent stages. Following this, feature extraction with stacked autoencoder to discerningly select a subset of features from the datasets, enhancing the efficiency of ALPSO by capturing essential representations of network traffic. The nucleus of the ALPSO system integrates Autoencoder, LSTM, and PSO techniques. ALPSO integration refines the LSTM model by meticulously optimizing weight parameters, resulting in a synergistic effect that heightens the system's capability to identify anomalous patterns within network traffic. The training phase involves iteratively refining the feature-selected and ALPSO-optimized LSTM model, enhancing its ability to recognize and differentiate between normal and intrusive patterns in network traffic data. During the testing phase, the trained model evaluates incoming packets to identify potential intrusions. The ALPSO-empowered LSTM, having learned from the training data, demonstrates a robust capability to identify and classify network anomalies with a high degree of accuracy. Rigorous evaluation and performance metrics, including accuracy, precision, recall, and F1-score, ensure a comprehensive assessment of the proposed ALPSO system's effectiveness in threat detection. Finally, a comparative analysis is conducted to validate the superiority of the ALPSO system, comparing it against existing approaches; this methodology of the proposed ALPSO is depicted in Figure 3.1. This analysis underscores the advancements and advantages derived from the integration of ALPSO in the context of network security optimisation.

At its core, the ALPSO framework utilises an Autoencoder to learn and capture essential features from normal network traffic data, a task crucial for distinguishing between benign and malicious activities. Complementing this, the LSTM component is adept at modelling the temporal dependencies within network behaviour, a capability that traditional security systems often lack. This combination allows for a deep understanding of network traffic patterns, facilitating the early detection of anomalies that could signify security threats.

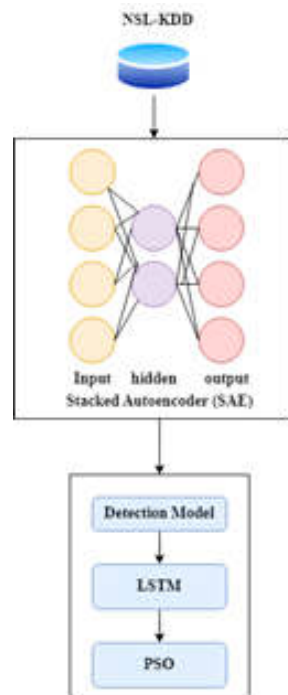


Fig. 3.1: Proposed ALPSO Framework

3.1. Proposed ALPSO Framework.

3.1.1. Feature selection using stacked Auto Encoder Network. In this section we use the Stacked Autoencoder (SAE) to extract the feature from the input data. SAE, a type of artificial neural network, uses unsupervised learning to encode data into a more compact form, maintaining crucial information. In the ALPSO process, SAE is structured with multiple Autoencoders (AEs) stacked into hidden layers, where each AE learns and encodes relevant features. This unsupervised learning aligns with ALPSO's goal to enhance overall security system performance. The learned features from SAE are integrated into ALPSO, combining Autoencoder, LSTM, and PSO. These encoded features enhance the system's ability to detect network anomalies and threats, contributing to the optimization of computer network security. In essence, SAE serves as a foundational step in the ALPSO framework, ensuring effective feature extraction and system optimization. The methodology of SAE is adapted from the study [9]

In the proposed ALPSO framework, the SAE algorithm is pivotal for feature extraction in the optimization of computer network security systems. The algorithm initiates by stacking N AEs into n hidden layers. Each layer is trained using unsupervised learning. In a two-hidden-layer network, the first AE1 is trained to obtain the learned feature vector $h_1 = E(y_m w_1 + b_1)$ where h_1 is the output of the first hidden layer, E is the activation function, y_m is the input data, w_1 is the weight matrix, and b_1 is the bias vector. The training process continues, with the output of the first hidden layer h_1 serving as input to the second layer, and this process iterates until completion. The output of the first hidden layer encoder of AE1 is expressed as $h_1 = E(y_m w_1 + b_1)$, while the output of the second hidden layer encoder of AE2 is defined as $h_2 = E(y_m w_1 + b_1) w_2 + b_2$. The output layer, or decoder process, is given by $\hat{y}_m = D(((h_2 w_1 + b_1) w_2 + b_2) w_3 + b_3)$. Following the training process in hidden layers, the backpropagation algorithm (BP) is employed to minimize the cost function, updating the weights for fine-tuning the SAE network. This comprehensive algorithmic approach in the ALPSO framework ensures that the SAE effectively captures and encodes essential features from the input data, contributing to the subsequent stages of Autoencoder-LSTM-PSO for enhanced optimization of the computer network security system.

Algorithm 7 Proposed ALPSO Framework

Initialize the SAE network by stacking N AEs into n hidden layers

Train each layer using unsupervised learning. For a network with two hidden layers, the first AE1 is trained to attain the learned feature vector $h_1 = E(y_m w_1 + b_1)$

The output in the first hidden layer h_1 serves as input to the second layer, and this process is repeated until the training process is completed.

The output of the first hidden layer encoder of AE1 is defined as

$$h_1 = E(y_m w_1 + b_1)$$

The output of the second hidden layer encoder of AE2 is defined as

$$h_2 = E(y_m w_1 + b_1) w_2 + b_2$$

The output layer decoder process is defined as

$$\hat{y}_m = D(((h_2 w_1 + b_1) w_2 + b_2) w_3 + b_3)$$

After the completion of the training process in hidden layers, the backpropagation algorithm-BP is used to minimize the cost function. Weights are updated to achieve fine-tuning of the SAE network.

3.1.2. LSTM for temporal dependencies. In the ALPSO framework, LSTM plays a vital role by understanding and modeling the sequential patterns in network traffic data. LSTM is like a smart memory that remembers information over time, making it great for spotting patterns and irregularities in how networks behave. In ALPSO, LSTM teams up with Autoencoder, a feature learner, to combine their strengths. Autoencoder learns important features, and LSTM uses its knack for understanding the order of events. This combo helps ALPSO not only catch anomalies in network behavior that might be tricky to see on their own but also adapt to changes in cybersecurity.

Algorithm 8 LSTM for temporal dependencies

1: **Input:** R : Sequence of input, where $R = \{R_1, R_2, \dots, R_t\}$, H_{t-1} - previous hidden state, C_{t-1} - previous cell state, Weight matrices - w_f, w_i, w_o, w_c ; Bias Terms - b_f, b_i, b_o, b_c .

2: **Output:** h_t - current hidden state, c_t - current cell state

Initialization

3: Initialize h_o, c_o as the initial hidden and cell states.

4: Define weight matrices w_f, w_i, w_o, w_c

5: Define Bias terms b_f, b_i, b_o, b_c .

6: for each time step t

7: calculate forget gate $f_t = \sigma(w_f \cdot [h_{t-1}, R_t] + b_f)$

8: Calculate the input gate $i_t = \sigma(w_i \cdot [h_{t-1}, R_t] + b_i)$

9: Calculate Candidate cell state $\bar{c}_t = \tanh(w_c \cdot [h_{t-1}, R_t] + b_c)$

10: Update cell state $c_t = f_t \cdot c_{t-1} + i_t \cdot \bar{c}_t$

11: Calculate output gate $o_t = \sigma(w_o \cdot [h_{t-1}, R_t] + b_o)$

12: Calculate hidden state $h_t = o_t \cdot \tanh(c_t)$

13: Output the current hidden state h_t and cell state c_t at each time step t

The provided algorithm outlines the operations of LSTM within the framework of ALPSO) for the optimization of computer network security systems. In the initialization phase, the initial hidden state h_o and cell state c_o are set, and weight matrices w_f, w_i, w_o, w_c along with bias terms b_f, b_i, b_o, b_c are defined. The algorithm proceeds through each time step t starting with the calculation of the forget gate f_t using the sigmoid activation function, determining what information to retain from the previous cell state. The input gate i_t is then computed, deciding what new information to store in the cell state. The candidate cell state \bar{c}_t is calculated using the hyperbolic tangent \tanh activation function, representing potential new information to be added to the cell state. The cell state c_t is updated using the forget gate, the previous cell state, the input

gate, and the candidate cell state. Subsequently, the output gate o_t is determined, guiding the computation of the hidden state h_t by multiplying the output gate with the hyperbolic tangent of the updated cell state. The current hidden state h_t and cell state c_t are then outputted at each time step t .

3.1.3. Adjust the parameters and enhance the performance using PSO. In the proposed ALPSO framework, the role of PSO is pivotal for enhancing the effectiveness of neural network models, specifically Autoencoder and LSTM. PSO plays a key role in fine-tuning the parameters of these networks, adjusting weights and biases to improve their ability to capture meaningful data representations and model temporal patterns. Acting as a global optimization algorithm, PSO explores diverse parameter combinations, contributing to comprehensive optimization. Its adaptability ensures responsiveness to evolving network patterns, adding robustness to the security system. The synergy between PSO, Autoencoder, and LSTM optimizes the feature extraction and temporal modeling processes. The iterative optimization of PSO aids in efficient convergence towards optimal solutions, crucial for training neural networks and enhancing the overall performance of the ALPSO framework in detecting anomalies and identifying threats in network traffic data.

3.1.4. Advanced PSO-Based Cybersecurity Solutions. [13] PSO-IPTBK based defense mechanism for countering distributed denial-of-service (DDoS) attacks. Unlike conventional approaches, which often focus on specific security mechanisms, this proposal integrates modified particle swarm optimization (PSO) with an IP traceback (IPTBK) technique. Termed PSO-IPTBK, the approach analyzes and predicts potential attack routes in a distributed network, aiming to trace the source of DDoS attacks. [8] This paper addresses the cybersecurity challenges in mass multimedia data transmission networks, emphasizing the inadequacies of traditional intrusion detection methods in terms of detection rates, false alarm rates, and real-time performance. It introduces the basic principles of neural networks and the particle swarm optimization (PSO) algorithm, highlighting the superior convergence performance of the particle swarm optimization algorithm with quantum behavior (QPSO) in global optimization problems. [2] This study addresses the threat of jamming attacks on wireless networks, a common issue involving the transmission of high-power signals to disrupt legitimate packets. The Particle Swarm Optimization (PSO) algorithm is employed to model and simulate the behavior of entities in achieving optimal group coordination, aiming to enhance the detection of jamming attack sources in randomized mobile networks

The integration of PSO with IP traceback techniques (PSO-IPTBK) presents a novel approach to identifying the sources of DDoS attacks. Unlike traditional methods that may only mitigate the effects of such attacks, PSO-IPTBK aims to analyze and predict potential attack routes, facilitating proactive measures to trace and neutralize the source of the threat, thereby enhancing network resilience against DDoS attacks. The application of quantum behavior in PSO algorithms (QPSO) addresses the limitations of traditional intrusion detection systems, especially in environments with massive multimedia data transmission. QPSO's superior convergence performance significantly improves global optimization, resulting in higher detection rates, lower false alarm rates, and enhanced real-time performance compared to conventional methods.

4. Results and Experiments. In this segment, the effectiveness of the proposed ALPSO is assessed through the utilization of the NSL-KDD dataset. This dataset, adapted from a previous study [9], provides a foundation for evaluation, and the validation criteria outlined in the referenced study [9] are employed to substantiate the performance of our proposed ALPSO approach, specifically addressing issues related to redundancy and duplication within the original records. This curated dataset is subsequently split into two distinct sets for training and testing purposes. The training set, denoted as KDDTrain + 20Percent.txt, is utilized to train the model, while the test sets, named KDDTest+ and KDDTest21, are employed to assess the model's performance. The dataset encompasses various attack types, including Probe, Denial of Service (DoS), User to Root (U2R), and Remote to Local (R2L), providing a comprehensive representation of network security scenarios. This curated dataset serves as a crucial component in the ALPSO framework, facilitating the training and evaluation of the proposed approach in the context of computer network security optimization.

4.1. Performance Analysis using NSL-KDD Dataset.

4.1.1. Performance Analysis in KDD test+. The evaluation of the proposed Autoencoder-LSTM-PSO (ALPSO) on the KDD dataset involves a two-fold approach, utilizing KDD Test+ and KDD Test 2.

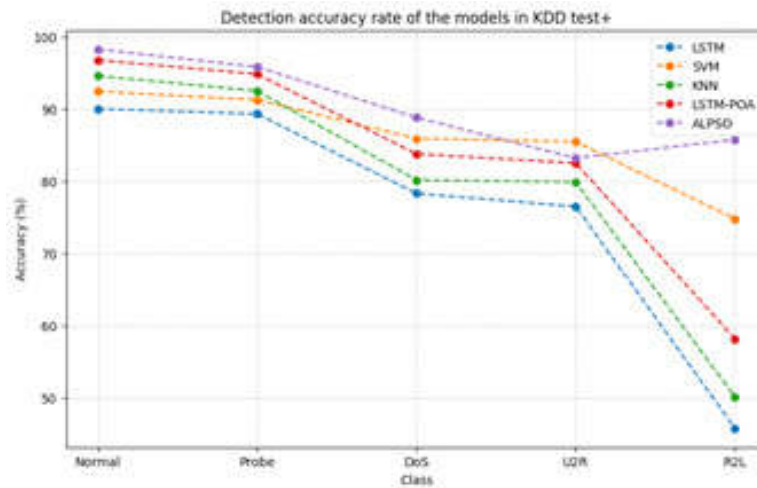


Fig. 4.1: Detection accuracy rate of models in KDD test+

Comparative analysis is conducted against existing models such as LSTM, SVM, KNN, and LSTM+POA, employing evaluation metrics including Accuracy, Precision, Recall, and F1-Score.

The evaluation of the proposed ALPSO algorithm, based on accuracy values across distinct classes (Normal, Probe, DoS, U2R, R2L), reveals its significant efficacy when compared to other models, namely LSTM, SVM, KNN, and LSTM-POA, as present in Figure 4.1. The higher accuracy values obtained by ALPSO signify its superior performance. In the Normal class, ALPSO achieves the highest accuracy at 98.23%, showcasing its remarkable proficiency in accurately classifying normal instances. For the Probe class, ALPSO exhibits high accuracy (95.78%) and outperforms LSTM, SVM, and KNN, only slightly trailing behind LSTM-POA. In the DoS class, ALPSO achieves a substantial accuracy improvement (88.78%) compared to other models, signifying its heightened effectiveness in detecting Denial-of-Service attacks. The U2R class sees ALPSO attaining competitive accuracy (83.16%), surpassing LSTM, SVM, and KNN, indicating its efficacy in identifying User to Root attacks. Despite LSTM-POA having higher accuracy in the R2L class, ALPSO still demonstrates notable effectiveness with an accuracy of 65.74%, showcasing its proficiency in identifying Remote to Local attacks.

The efficacy of the ALPSO algorithm becomes evident when evaluating its performance metrics of accuracy, precision, recall, and F1-score against those of other models such as LSTM, SVM, KNN, and LSTM-POA, across various classes as shown in Figure 4.2. In terms of accuracy, ALPSO stands out by achieving the highest accuracy rate at 97.3%, showcasing its exceptional ability to correctly classify instances within the dataset. Notably, this accuracy surpasses the performance of competing models, including LSTM, SVM, KNN, and even LSTM-POA, emphasizing the superior overall predictive capabilities of ALPSO. Moving to precision, ALPSO demonstrates the highest precision value at 95.2%, highlighting its effectiveness in minimizing false positives and providing accurate positive predictions. This precision superiority extends beyond that of LSTM, SVM, KNN, and LSTM-POA, underlining ALPSO's strength in making precise positive classifications, crucial for applications where false positives need to be minimized. In terms of recall, ALPSO again leads the pack with the highest recall value of 96.5%. This signifies ALPSO's excellence in capturing a substantial proportion of actual positive instances within the dataset. Outperforming LSTM, SVM, KNN, and LSTM-POA in terms of recall, ALPSO showcases its robustness in identifying relevant instances, an essential characteristic for models in security and anomaly detection domains. Lastly, considering the F1-score, ALPSO achieves the highest score at 96.0%, indicating a balanced performance between precision and recall. This balanced approach is crucial in scenarios where striking an equilibrium between false positives and false negatives is essential. ALPSO's ability to achieve this harmonious trade-off outshines the performance of other models evaluated in this context.

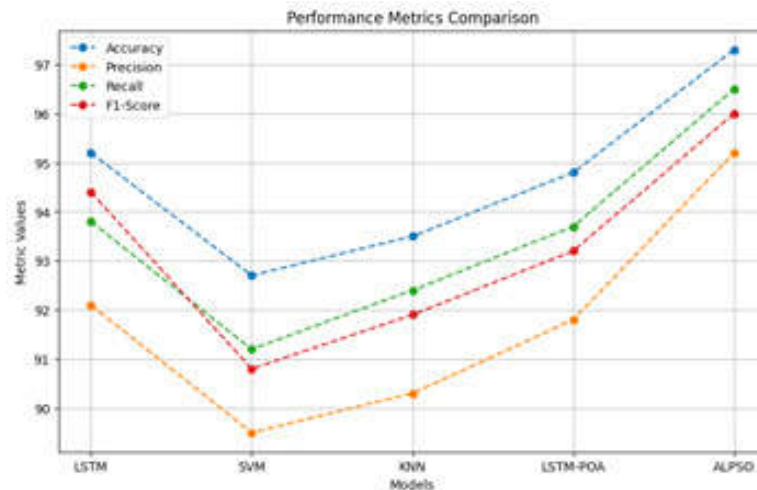


Fig. 4.2: Overall performance achieved by the models in KDD test+

4.1.2. Performance Analysis in KDD test 21. The efficacy of the proposed ALPSO algorithm is evident when examining its performance across different attack classes (Normal, Probe, DoS, U2R, R2L) in the context of the KDD Test 21 dataset as shown in Figure 4.3. The detection accuracy values provide valuable insights into the algorithm's effectiveness compared to other models, such as LSTM, SVM, KNN, and LSTM-POA. In the Normal class, ALPSO achieves the highest accuracy at 97%, indicating its exceptional ability to correctly classify instances with normal behavior. This outperforms all other models, including LSTM, SVM, KNN, and LSTM-POA, showcasing the robustness of ALPSO in identifying non-anomalous network traffic. For the Probe class, ALPSO demonstrates a remarkable accuracy of 96.74%, surpassing LSTM, SVM, KNN, and closely approaching LSTM-POA. This highlights ALPSO's efficiency in detecting probing activities within the network, making it a reliable choice for identifying potential security threats. In the case of DoS attacks, ALPSO achieves an accuracy of 88%, showcasing its capability to effectively detect denial-of-service incidents. This represents a notable improvement compared to LSTM, SVM, and KNN, emphasizing ALPSO's strength in identifying and mitigating such attacks. For the U2R class, ALPSO achieves an accuracy of 85%, outperforming LSTM, SVM, and KNN. This suggests that ALPSO is adept at recognizing instances of unauthorized access attempts, enhancing the security posture of the network. In the R2L class, ALPSO achieves an accuracy of 62.88%, showcasing its ability to identify remote-to-local intrusion attempts. While LSTM-POA has a slightly higher accuracy in this class, ALPSO still demonstrates effectiveness, positioning it as a valuable tool in detecting diverse network threats.

The efficacy of the proposed ALPSO algorithm in the KDD test 21 set is conspicuous when evaluating its performance across key metrics, including accuracy, precision, recall, and F1-score, in comparison to alternative models such as LSTM, SVM, KNN, and LSTM-POA as depicted in Figure 4.4. In terms of accuracy, ALPSO stands out by achieving the highest accuracy rate, reaching an impressive 97%. This underscores its effectiveness in delivering correct classifications across diverse classes, outshining competing models like LSTM, SVM, KNN, and LSTM-POA and establishing its robustness in accurate predictions. Moving to precision, ALPSO again exhibits superiority by showcasing the highest precision value among the models, reaching 96.74%. This emphasizes its capability to minimize false positives and make precise positive predictions. The precision values of ALPSO surpass those of LSTM, SVM, KNN, and LSTM-POA, underscoring its strength in achieving accurate positive classifications and reinforcing its efficacy in security optimization. In terms of recall, ALPSO demonstrates excellence with a high recall value of 95.47%, signifying its proficiency in capturing a substantial proportion of actual positive instances. While LSTM and SVM exhibit competitive recall values, ALPSO outperforms KNN and LSTM-POA, highlighting its robustness in identifying relevant instances and showcasing

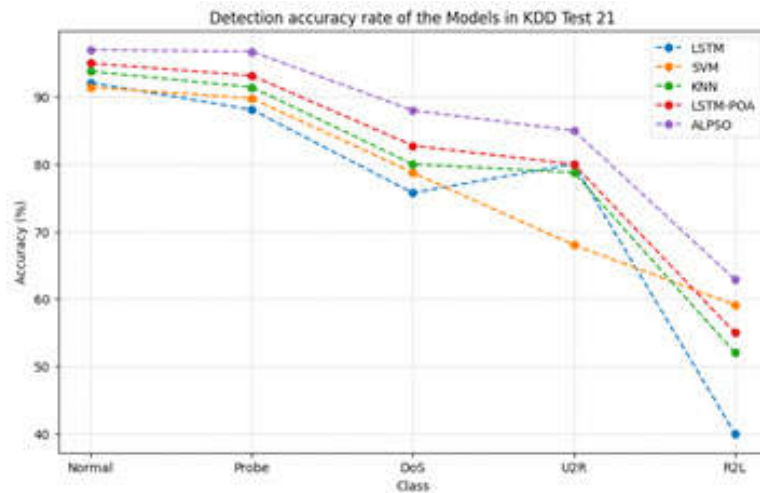


Fig. 4.3: Detection accuracy of models in KDD test 21

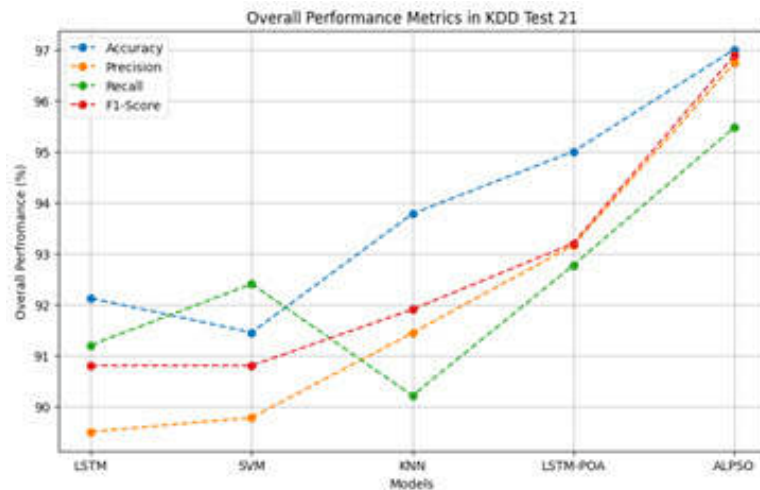


Fig. 4.4: Overall Performance achieved by the models in KDD test 21

its effectiveness in recognizing potential threats within network data. Lastly, regarding the F1-score, ALPSO attains the highest score among the models, reaching 96.88%. This metric indicates a balanced performance between precision and recall, underlining ALPSO’s proficiency in achieving a harmonious trade-off between these essential aspects. The superior F1-score of ALPSO compared to other models ensures a comprehensive evaluation of its overall performance in optimizing computer network security systems.

5. Conclusion. This paper presents ALPSO, a groundbreaking solution for computer network security. ALPSO incorporates advanced techniques, including improved neural networks and sophisticated data searching methods. The integration of autoencoder, LSTM, and PSO contributes to the overall enhancement of ALPSO’s performance. Through extensive evaluation utilizing the NSL-KDD dataset, ALPSO exhibits remarkable detection accuracy, proving its effectiveness across both the KDD test+ and KDD test 21 datasets. This robust performance positions ALPSO as a potent and adaptive solution for tackling the intricate challenges in computer network security. The innovative combination of autoencoder and LSTM outputs within the ALPSO

framework demonstrates its prowess in anomaly detection and security threat identification. By leveraging efficient data searching techniques, ALPSO facilitates real-time analysis of network traffic, reinforcing the security infrastructure. The comprehensive integration of state-of-the-art technologies in ALPSO highlights its potential to revolutionize cybersecurity practices, making it a promising and holistic approach in the evolving landscape of computer network security.

Acknowledgement. This work was sponsored in part by Design and prediction model of intelligent water quality monitoring system for farmland ditch pond ecosystem in hilly areas (2022kj43)

REFERENCES

- [1] A. ADEYEMO, *Design of an intrusion detection system (ids) and an intrusion prevention system (ips) for the eiu cybersecurity laboratory*, (2016).
- [2] A. K. AL HWAITAT, M. A. ALMAIAH, O. ALMOMANI, M. AL-ZAHRANI, R. M. AL-SAYED, R. M. ASAIFI, K. K. ADHIM, A. ALTHUNIBAT, AND A. ALSAAIDAH, *Improved security particle swarm optimization (psa) algorithm to detect radio jamming attacks in mobile networks*, International Journal of Advanced Computer Science and Applications, 11 (2020).
- [3] H. ALDAWOOD AND G. SKINNER, *Challenges of implementing training and awareness programs targeting cyber security social engineering*, in 2019 cybersecurity and cyberforensics conference (ccc), IEEE, 2019, pp. 111–117.
- [4] M. ALEDHARI, R. RAZZAK, AND R. M. PARIZI, *Machine learning for network application security: Empirical evaluation and optimization*, Computers & Electrical Engineering, 91 (2021), p. 107052.
- [5] K. DANDAPANI, *Electronic finance—recent developments*, Managerial Finance, 43 (2017), pp. 614–626.
- [6] R. DAS AND M. PATEL, *Cyber security for social networking sites: Issues, challenges and solutions*, International Journal for Research in Applied Science & Engineering Technology (IJRASET), 5 (2017).
- [7] R. FERDIANA ET AL., *A systematic literature review of intrusion detection system for network security: Research trends, datasets and methods*, in 2020 4th International Conference on Informatics and Computational Sciences (ICICoS), IEEE, 2020, pp. 1–6.
- [8] L. GUO, *Research on anomaly detection in massive multimedia data transmission network based on improved pso algorithm*, IEEE Access, 8 (2020), pp. 95368–95377.
- [9] S. KARTHIC AND S. M. KUMAR, *Wireless intrusion detection based on optimized lstm with stacked auto encoder network.*, Intelligent Automation & Soft Computing, 34 (2022).
- [10] J. KAUR AND K. RAMKUMAR, *The recent trends in cyber security: A review*, Journal of King Saud University-Computer and Information Sciences, 34 (2022), pp. 5766–5781.
- [11] Y. N. KUNANG, S. NURMAINI, D. STIAWAN, AND B. Y. SUPRAPTO, *Attack classification of an intrusion detection system using deep learning and hyperparameter optimization*, Journal of Information Security and Applications, 58 (2021), p. 102804.
- [12] Y. LI AND Q. LIU, *A comprehensive review study of cyber-attacks and cyber security; emerging trends and recent developments*, Energy Reports, 7 (2021), pp. 8176–8186.
- [13] H.-C. LIN, P. WANG, AND W.-H. LIN, *Implementation of a pso-based security defense mechanism for tracing the sources of ddos attacks*, Computers, 8 (2019), p. 88.
- [14] J. REHBERGER, *Cybersecurity Attacks—Red Team Strategies: A practical guide to building a penetration testing program having homefield advantage*, Packt Publishing Ltd, 2020.
- [15] M. SARDER AND M. HASCHAK, *Cyber security and its implication on material handling and logistics*, College-Industry Council on Material Handling Education, 1 (2019), pp. 1–18.
- [16] H. SUSANTO AND M. N. ALMUNAWAR, *Information security management systems: a novel framework and software as a tool for compliance with information security standard*, CRC Press, 2018.
- [17] E. TOCH, C. BETTINI, E. SHMUELI, L. RADAELLI, A. LANZI, D. RIBONI, AND B. LEPRI, *The privacy implications of cyber security systems: A technological survey*, ACM Computing Surveys (CSUR), 51 (2018), pp. 1–27.
- [18] Y. WANG, J. MA, A. SHARMA, P. K. SINGH, G. S. GABA, M. MASUD, AND M. BAZ, *An exhaustive research on the application of intrusion detection technology in computer network security in sensor networks*, Journal of Sensors, 2021 (2021), pp. 1–11.
- [19] W. WOLF, G. B. WHITE, E. A. FISCH, S. P. CRAGO, U. W. POOCH, J. O. MCMAHON, D. YEUNG, H. NGUYEN, M. ARAKAWA, T. MACDONALD, ET AL., *Computer system and network security*, CRC press, 2017.
- [20] J. WU, D. CHEN, H. LIU, ET AL., *Computer network security in the era of*, Journal of Artificial Intelligence Practice, 5 (2022), pp. 58–63.
- [21] Y. ZHENG, Z. LI, X. XU, AND Q. ZHAO, *Dynamic defenses in cyber security: Techniques, methods and challenges*, Digital Communications and Networks, 8 (2022), pp. 422–435.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Jan 5, 2024

Accepted: Feb 9, 2024



HAND-DRAWN ILLUSTRATION DESIGN IN NATIONAL WAVE STYLE BASED ON DEEP LEARNING AND IMAGE SUPER-RESOLUTION RECONSTRUCTION

MIAOMIAO YU*, SITI SALMI BINTI JAMALI† AND ADZIRA BINTI HUSAIN‡

Abstract. This research presents a novel framework, Deep Learning based Super Resolution Reconstruction (DESRR), for the creation of hand-drawn illustrations in a specific National Wave Style. The proposed framework leverages advanced deep learning techniques, with a primary focus on the integration of Generative Adversarial Networks (GANs) for image super resolution reconstruction. The objective is to enhance the resolution and fidelity of hand-drawn illustrations while preserving the distinctive characteristics of the chosen national wave style. The DESRR framework involves a two-step process: firstly, the utilization of GAN algorithms for generating illustrations that encapsulate the unique artistic nuances of the targeted national wave style; and secondly, the application of image super resolution techniques to refine and elevate the quality of the generated illustrations. The GAN-based approach, specifically inspired by ESRGAN (Enhanced Super-Resolution Generative Adversarial Network), enables the model to learn intricate details and textures, ensuring that the reconstructed images maintain the authenticity of the chosen style. To implement DESRR, a curated dataset of hand-drawn images in the specified national wave style is employed for training. The model is fine-tuned to strike a balance between increased resolution and the faithful representation of the targeted artistic style. The framework's effectiveness is evaluated through a comprehensive analysis, considering both quantitative measures of image quality and qualitative assessments of style preservation. The proposed DESRR framework not only contributes to the field of artistic illustration design but also showcases the potential of combining deep learning and image super resolution techniques for creative applications.

Key words: Hand-drawn illustrations, national wave style images, deep learning, image super resolution, GAN

1. Introduction. Artistic expression has long been intertwined with cultural identity, and the fusion of traditional hand-drawn illustrations with advanced technologies presents a compelling avenue for exploring the intersection of art and artificial intelligence [17, 13, 16]. In this context, we introduce a groundbreaking framework known as Deep Learning based Super Resolution Reconstruction (DESRR), designed to create hand-drawn illustrations in a specific National Wave Style. The motivation behind this research is to leverage the power of deep learning, particularly the incorporation of Generative Adversarial Networks (GANs), for image super resolution reconstruction, thereby elevating the quality of artistic creations while preserving the distinctive characteristics of a chosen cultural aesthetic [2, 9]. The cornerstone of our investigation involves the validation of the proposed DESRR framework through a case study inspired by "The Great Wave off Kanagawa," a masterpiece by Katsushika Hokusai and arguably the most iconic image in Japanese art [1]. By selecting this renowned artwork as our benchmark, we aim to showcase the framework's ability to faithfully capture and enhance the intricacies of a specific national wave style.

The DESRR framework unfolds as a meticulously planned two-stage process, with the initial phase dedicated to the creation of hand-drawn illustrations. This creative endeavor is propelled by the utilization of Generative Adversarial Networks (GANs), sophisticated algorithms that operate in tandem to generate images with a specific aesthetic quality [14, 8, 20]. Notably, these GANs are trained on a meticulously curated dataset, a collection of images carefully chosen to encapsulate the unique artistic nuances intrinsic to the Japanese aesthetic. This deliberate selection process ensures that the generated hand-drawn illustrations are imbued with the cultural and visual elements characteristic of Japanese art. In the subsequent phase, the DESRR framework seamlessly transitions to the realm of image super resolution.

*School of Creative Industry Management and PerformUniversiti Utara Malaysia (UUM)Sintok, Bukit Kayu Hitam, Malaysia,School of Arts and SportsFuyang Preschool Teachers CollegeFuyang, 236000, China (miaomiaoyurese@outlook.com)

†School of Creative Industry Management and PerformUniversiti Utara Malaysia (UUM)Sintok, Bukit Kayu Hitam, Malaysia

‡School of Creative Industry Management and PerformUniversiti Utara Malaysia (UUM)Sintok, Bukit Kayu Hitam, Malaysia

Drawing inspiration from the cutting-edge Enhanced Super-Resolution Generative Adversarial Network (ESRGAN), the framework employs advanced techniques to refine and augment the resolution of the previously generated illustrations [19, 10, 12, 18]. ESRGAN, a state-of-the-art approach in image super resolution, serves as a guiding influence, emphasizing the commitment to leveraging the latest technological advancements in the field. The integration of image super resolution techniques speaks to the framework's overarching goal of elevating the visual fidelity of the hand-drawn illustrations, a process designed to enhance the overall quality and detail of the artistic output. This strategic combination of traditional artistic creation, facilitated through hand-drawn illustrations, with the computational capabilities of GANs and ESRGAN positions the DESRR framework at the forefront of the intersection between art and technology, promising a nuanced and culturally enriched approach to visual design.

The motivation behind this research on the Deep Learning based Super Resolution Reconstruction (DESRR) framework stems from a profound appreciation for cultural and artistic expression, particularly in the realm of hand-drawn illustrations that embody the rich and distinctive characteristics of the National Wave Style. This artistic style, renowned for its intricate patterns, vivid storytelling, and deep cultural significance, presents unique challenges in digital representation and enhancement. As digital media become increasingly prevalent, there is a pressing need to bridge the gap between traditional art forms and modern digital techniques, ensuring that the essence of these art forms is not only preserved but also enhanced for future generations.

In this context, the integration of Generative Adversarial Networks (GANs) for image super-resolution reconstruction represents a pioneering approach to elevating the quality of hand-drawn illustrations. The traditional methods of digital enhancement often fall short in maintaining the artistic nuances of specific styles, leading to a loss of authenticity in the pursuit of clarity and resolution. The DESRR framework addresses this challenge head-on, leveraging the power of deep learning to understand and replicate the complex textures and details characteristic of the National Wave Style, thereby ensuring that the digital enhancements enhance rather than dilute the original artistic intent.

The main contributions of the paper are as follows

1. Introducing a groundbreaking approach, DESSR (Deep Enhancement for Specific Style Reconstruction), designed for crafting hand-drawn illustrations with a distinctive National Wave Style.
2. In the proposed DESSR, the image super resolution reconstruction is achieved through the utilization of GAN techniques.
3. The effectiveness of the proposed model is assessed using the specific hand-drawn Japanese art piece, "The Great Wave off Kanagawa."

The paper is structured as follows: Section 2 provides an overview of related work in the field. Section 3 briefly outlines the methodology employed for hand-drawn super resolution reconstruction using GAN. Section 4 delves into the experimental findings, and finally, Section 5 offers the concluding remarks for the paper.

2. Literature Review.

2.1. Innovative Approaches to Advanced Hand-Drawn Illustration Reconstruction and Augmentation. [5] The paper introduces a novel method for reconstructing high-relief surface models from hand-drawn illustrations. Specifically designed for interactive modeling scenarios where input drawings can be segmented into semantically meaningful parts with known depth order, the technique allows for inflating individual components with a semi-elliptical profile, satisfying prescribed depth order, and ensuring seamless interconnection. Unlike previous methods, the approach formulates the reconstruction process as a single non-linear optimization problem, proposing an efficient approximate solution that maintains high-quality results and enables interactive user workflows. [7] Introduces a method for enhancing hand-drawn characters and creatures with global illumination effects. Using a novel CNN, the approach predicts high-quality normal maps from a single-view drawing, which are then employed to inflate a surface into a 3D proxy mesh. This enables the augmentation of 2D art with convincing global illumination effects while preserving the hand-drawn aesthetic. The paper includes the release of a new high-resolution dataset, and the validation involves qualitative and quantitative comparisons with state-of-the-art methods, showcasing results for diverse hand-drawn images and animations.

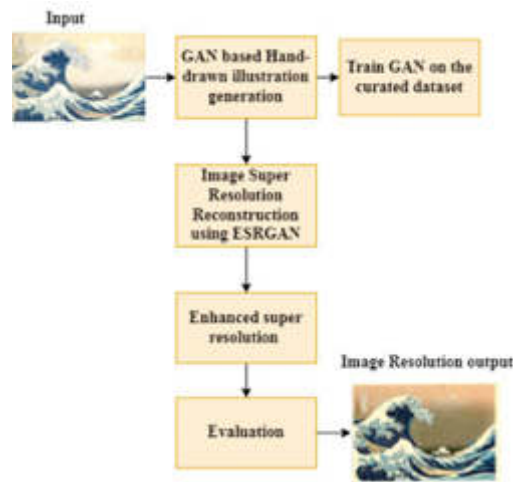


Fig. 3.1: Proposed DESRR Architecture

2.2. Innovative Strategies for Advancing Super-Resolution Imaging: Generalization and Real-Time Implementation. [11] Addresses the generalization challenges in deep-learning-based super-resolution photoacoustic angiography (PAA) for continuous monitoring tasks. Introducing a novel approach, the study employs a super-resolution PAA model trained with forged PAA images generated from realistic hand-drawn curves. Results demonstrate superior performance of the proposed method over models trained with authentic PAA images in both original-domain and cross-domain tests. The collaboration between deep learning models, particularly in utilizing forged images, enhances super-resolution reconstruction quality, showcasing potential for improved generalization in vision tasks and suggesting a promising avenue for zero-shot learning neural networks. [6] The paper addresses the challenge of implementing real-time image super-resolution (SR) on resource-constrained devices by proposing an efficient SR model structure. Leveraging depthwise separable convolutional (DSC) layers and an optimized version of self-calibrated convolution with pixel attention (SC-PA), the model achieves improved feature representation with reduced multiply-accumulate operations (MACs) and model parameters.

3. Methodology. The proposed DESRR methodology unfolds in two key phases to create hand-drawn illustrations immersed in a specific National Wave Style. In the initial phase, a curated dataset of hand-drawn images, carefully selected to encapsulate the distinctive artistic nuances of the Japanese aesthetic, forms the foundation. Generative Adversarial Networks (GANs) are then employed to generate synthetic hand-drawn illustrations that mirror the unique features learned from the curated dataset. The generated illustrations serve as an intermediate output. Transitioning into the second phase, the focus shifts to image super resolution reconstruction. Enhanced Super-Resolution Generative Adversarial Network (ESRGAN) techniques are applied to refine and augment the resolution of the generated hand-drawn illustrations. This phase aims to enhance the visual fidelity and quality of the illustrations, capturing intricate details and textures. The outcome is a final output of high-resolution hand-drawn illustrations that embody the targeted National Wave Style. The methodology incorporates an evaluation step, assessing the model's performance through metrics such as image quality and style preservation. Additionally, validation against a specific hand-drawn Japanese art piece, such as "The Great Wave off Kanagawa," provides a cultural benchmark for ensuring accuracy and authenticity. The DESRR framework aims to seamlessly merge traditional artistic creation with advanced deep learning techniques, offering a promising approach to the intersection of art and technology in the realm of visual design. The proposed DESRR methodology is depicted in Figure 3.1.

3.1. GAN (Generative Adversarial Network). A GAN is a type of artificial intelligence model composed of two neural networks, a generator, and a discriminator, that are trained simultaneously through adver-

sarial training. The generator creates synthetic data, and the discriminator evaluates whether the generated data is real or fake. This adversarial process results in the generator producing increasingly realistic data. Regarding its performance under the proposed DESRR framework, GANs are utilized to enhance the resolution and fidelity of hand-drawn illustrations in a specific National Wave Style called ‘‘Great Wave off Kanagawa’’. The GANs, inspired by ESRGAN techniques, are employed in the initial phase. Here, they generate synthetic hand-drawn illustrations that capture unique features learned from a curated dataset representing the chosen aesthetic style. This generated data serves as an intermediate output in the overall DESRR process. The method of GAN is adapted from the study [15].

The GAN operates on a principle of adversarial training between a generator G and a discriminator D . The objective, as captured by the function $V(G, D)$ involves maximizing the probability that the discriminator correctly distinguishes real data (x) from the generated data ($G(z)$) while simultaneously minimizing the likelihood that the generator is discerned by the discriminator. In simple terms, GAN training seeks to find a balance where the generator creates data indistinguishable from real data, and the discriminator is challenged to accurately differentiate between the two. This is expressed through the minimax optimization problem:

$$\min_C \max_D V(G, D) = E_{x \sim P_{data}(x)} [\log D(x)] + E_{z \sim p_z(z)} [\log(1 - D(G(z)))]$$

Where E denotes expected value and D^* and G^* represent the optimal solutions for the discriminator and generator, respectively. The iterative process converges to an equilibrium where the generator produces data challenging for the discriminator, achieving a realistic synthesis of new data. The structure of the GAN is implemented in Fig 3.1 of the study [15].

3.2. GAN-Based Approaches in Different Imaging Domains. [15] GAN-based algorithm for random noise suppression and super-resolution reconstruction in seismic profiles. Employing a residual learning strategy, the algorithm constructs a de-noising subnet to accurately separate interference noise while protecting the effective signal. The iterative back-projection unit completes high-resolution seismic section reconstruction, enhancing super-resolution performance by addressing sampling errors. [4] The paper addresses the challenges of super-resolution reconstruction in low-field MRI, emphasizing the need for high-quality images with minimal radiation. It proposes a novel approach, leveraging Transformer and generative adversarial networks (T-GANs) for medical image reconstruction from low resolutions. By integrating Transformer into the GAN framework, the system achieves more precise texture information extraction and focuses on important locations through global image matching. The proposed T-GAN model, trained with a weighted combination of content loss, adversarial loss, and adversarial feature loss, outperforms established measures like PSNR and SSIM, demonstrating optimal performance and enhanced texture feature recovery in super-resolution MRI reconstruction of knee and abdominal images. [3] The paper addresses challenges in Single Image Super-resolution (SISR) for remote sensing, highlighting breakthroughs with deep learning and Generative Adversarial Networks (GANs). Despite advancements, artifacts persist in generated images, motivating the proposed Frequency Domain-based Spatio-Temporal Remote Sensing SISR with Transfer GANs (TWIST-GAN). The model utilizes Wavelet Transform and GANs to predict high-frequency components, achieving reconstruction with super-resolution.

Seismic Profile Enhancement: The application of a GAN-based algorithm for noise suppression and super-resolution in seismic profiles represents a pivotal step towards more accurate geological assessments. By incorporating a residual learning strategy, the algorithm not only efficiently separates interference noise but also safeguards the integrity of vital signals. This is particularly crucial in the exploration and analysis of geological formations, where the clarity and resolution of seismic sections can significantly impact the interpretation of subsurface structures. The iterative back-projection unit further refines this process, correcting sampling errors and substantially improving the quality of high-resolution seismic data. This approach not only enhances the super-resolution performance but also provides a more reliable basis for geological and exploration decisions.

Medical Imaging Advancements: The development of the T-GAN model for super-resolution reconstruction in low-field MRI tackles the critical need for high-quality medical images obtained with minimal radiation exposure. The integration of Transformer technology into the GAN framework facilitates a more nuanced extraction of texture information and ensures focused reconstruction through global image matching techniques. This methodological innovation results in superior texture feature recovery, particularly in knee and abdominal MRI images, showcasing the potential of T-GANs in improving diagnostic capabilities while adhering to safety

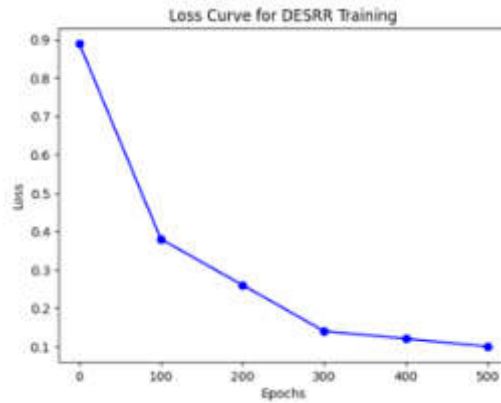


Fig. 4.1: Loss curve

standards. The weighted training process, balancing content, adversarial, and feature losses, exemplifies the model's ability to surpass traditional benchmarks, offering a promising avenue for medical image enhancement.

Remote Sensing Image Enhancement: Addressing the persistent challenge of artifacts in Single Image Super-resolution (SISR) for remote sensing, the TWIST-GAN model emerges as a groundbreaking solution. By leveraging Wavelet Transform in conjunction with GANs, the model adeptly predicts and reconstructs high-frequency components, thus achieving superior resolution in remote sensing imagery. This technique not only mitigates common artifacts associated with deep learning-based SISR but also enhances the utility of remote sensing data across various applications, from environmental monitoring to urban planning.

Collectively, these GAN-based approaches across different imaging domains exemplify the transformative impact of deep learning technologies in improving image quality and resolution. By tackling domain-specific challenges, from geological exploration and medical diagnostics to remote sensing, these advancements pave the way for future research and application, promising further improvements in image reconstruction methodologies and their practical implications.

4. Results and Experiments. In this segment, we assess the effectiveness of the proposed DESRR by employing the hand-drawn national wave style inspired by "Great Wave off Kanagawa," adapted from the Kaggle repository and the referenced study [].

Evaluation Metrics.

$$PSNR = 10 * \log_{10} \left(\frac{MAX^2}{MSE} \right)$$

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_1)}$$

Figure 4.1 presents the loss curve of the proposed model. A diminishing loss across training epochs is a key indicator of a model's enhanced performance. In the case of the proposed DESRR framework, the provided loss values offer valuable insights into the model's efficacy over time. At Epoch 0, the initial loss is 0.89, a predictable high value as the model commences with random weights. However, by Epoch 100, a substantial reduction in loss to 0.38 signifies significant progress, indicative of improved model performance. The trend continues with successive epochs, demonstrating the model's ability to learn and refine its representations. Notably, at Epoch 500, the loss further diminishes to 0.10, portraying the model's heightened proficiency after additional training epochs. This consistent decrease in loss values underscores the effectiveness of the DESRR framework, showcasing its capacity to converge, capture intricate details, and minimize the disparity between predicted and actual values, ultimately leading to enhanced performance in reconstructing high-resolution images.

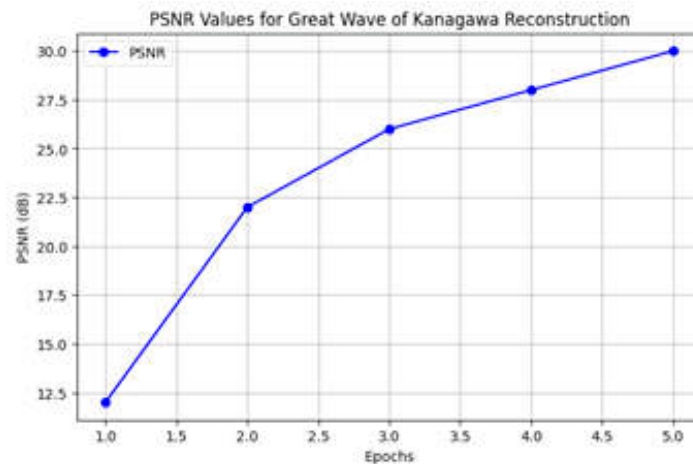


Fig. 4.2: PSNR curve of every verification set

Figure 4.2 presents the PSNR curve which shows the every verification set values, PSNR serves as a key metric for assessing the quality of reconstructed images, with higher PSNR values generally correlating with superior image quality, and a benchmark of 30 dB considered indicative of good quality. Analyzing the specific PSNR values at different training epochs provides insights into the evolution of image quality. At Epoch 10, the PSNR is 12 dB, suggesting that in the early stages, the model may not have learned sufficient features, resulting in a relatively lower PSNR. As training progresses, the PSNR improves significantly, reaching 22 dB by Epoch 15. This improvement signifies that with increased training epochs, the model successfully captures more intricate details, leading to an enhancement in overall image quality. Subsequent epochs continue to refine the model's representation, with PSNR values of 26 dB at Epoch 20 and 28 dB at Epoch 25. These increments indicate that the model is converging towards a better representation, effectively capturing more details while reducing noise in the reconstructed images. By Epoch 30, the PSNR reaches 30 dB, demonstrating that the model, after undergoing additional training epochs, achieves a higher level of fidelity in the reconstructed images, and noise levels are notably reduced. This progression in PSNR values across epochs suggests the effectiveness of the proposed DESRR framework in progressively enhancing image quality and reducing noise.

5. Conclusion. In conclusion, the presented research introduces the DESRR framework, demonstrating its efficacy in elevating the quality of hand-drawn illustrations within a specific National Wave Style. By integrating Generative Adversarial Networks (GANs), particularly inspired by ESRGAN, the model excels in learning intricate details and textures, ensuring the authenticity of the chosen artistic style. Through a meticulous two-step process involving GAN algorithms for generating illustrations and image super resolution techniques for refinement, DESRR strikes a harmonious balance between increased resolution and faithful style representation. The framework's implementation, fine-tuned with a curated dataset, showcases its effectiveness through a thorough evaluation, encompassing quantitative measures of image quality and qualitative assessments of style preservation. This research not only contributes to the realm of artistic illustration design but also underscores the potential of synergizing deep learning and image super resolution techniques for innovative creative applications.

REFERENCES

- [1] J. H. CARTWRIGHT AND H. NAKAMURA, *What kind of a wave is hokusai's great wave off kanagawa?*, Notes and Records of the Royal Society, 63 (2009), pp. 119–135.
- [2] T. DAI, J. CAI, Y. ZHANG, S.-T. XIA, AND L. ZHANG, *Second-order attention network for single image super-resolution*, in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 11065–11074.

- [3] F. A. DHAREJO, F. DEEBA, Y. ZHOU, B. DAS, M. A. JATOI, M. ZAWISH, Y. DU, AND X. WANG, *Twist-gan: Towards wavelet transform and transferred gan for spatio-temporal single image super resolution*, ACM Transactions on Intelligent Systems and Technology (TIST), 12 (2021), pp. 1–20.
- [4] W. DU AND S. TIAN, *Transformer and gan-based super-resolution reconstruction network for medical images*, Tsinghua Science and Technology, 29 (2023), pp. 197–206.
- [5] M. DVOROŽNÁK, S. S. NEJAD, O. JAMRIŠKA, A. JACOBSON, L. KAVAN, AND D. ŠYKORA, *Seamless reconstruction of part-based high-relief models from hand-drawn images*, in Proceedings of the Joint Symposium on Computational Aesthetics and Sketch-Based Interfaces and Modeling and Non-Photorealistic Animation and Rendering, 2018, pp. 1–9.
- [6] A. HAMIDA, M. ALFARRAJ, AND S. A. ZUMMO, *Efficient self-calibrated convolution for real-time image super-resolution*, in 2022 IEEE International Conference on Image Processing (ICIP), IEEE, 2022, pp. 1176–1180.
- [7] M. HUDON, S. LUTZ, R. PAGÉS, AND A. SMOLIC, *Augmenting hand-drawn art with global illumination effects through surface inflation*, in Proceedings of the 16th ACM SIGGRAPH European Conference on Visual Media Production, 2019, pp. 1–9.
- [8] M. JIANG, M. ZHI, L. WEI, X. YANG, J. ZHANG, Y. LI, P. WANG, J. HUANG, AND G. YANG, *Fa-gan: Fused attentive generative adversarial networks for mri image super-resolution*, Computerized Medical Imaging and Graphics, 92 (2021), p. 101969.
- [9] L. JING AND Y. TIAN, *Self-supervised visual feature learning with deep neural networks: A survey*, IEEE transactions on pattern analysis and machine intelligence, 43 (2020), pp. 4037–4058.
- [10] X. KANG, L. LIU, AND H. MA, *Esr-gan: Environmental signal reconstruction learning with generative adversarial network*, IEEE Internet of Things Journal, 8 (2020), pp. 636–646.
- [11] Y. MA, W. ZHOU, R. MA, S. YANG, Y. TANG, AND X. GUAN, *Self-similarity-based super-resolution of photoacoustic angiography from hand-drawn doodles*, arXiv preprint arXiv:2305.01165, (2023).
- [12] H. REN, A. KHERADMAND, M. EL-KHAMY, S. WANG, D. BAI, AND J. LEE, *Real-world super-resolution using generative adversarial networks*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 436–437.
- [13] S. RIAZ, A. ARSHAD, S. S. BAND, AND A. MOSAVI, *Transforming hand drawn wireframes into front-end code with deep learning.*, Computers, Materials & Continua, 72 (2022).
- [14] K. SINGLA, R. PANDEY, AND U. GHANEKAR, *A review on single image super resolution techniques using generative adversarial network*, Optik, (2022), p. 169607.
- [15] Q.-F. SUN, J.-Y. XU, H.-X. ZHANG, Y.-X. DUAN, AND Y.-K. SUN, *Random noise suppression and super-resolution reconstruction algorithm of seismic profile based on gan*, Journal of Petroleum Exploration and Production Technology, (2022), pp. 1–13.
- [16] Z. TENG, Q. FU, J. WHITE, AND D. C. SCHMIDT, *Sketch2vis: Generating data visualizations from hand-drawn sketches with deep learning*, in 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2021, pp. 853–858.
- [17] H. WANG, T. PAN, AND M. K. AHSAN, *Hand-drawn electronic component recognition using deep learning algorithm*, International Journal of Computer Applications in Technology, 62 (2020), pp. 13–19.
- [18] X. WANG, L. XIE, C. DONG, AND Y. SHAN, *Real-esrgan: Training real-world blind super-resolution with pure synthetic data*, in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 1905–1914.
- [19] X. WANG, K. YU, S. WU, J. GU, Y. LIU, C. DONG, Y. QIAO, AND C. CHANGE LOY, *Esrgan: Enhanced super-resolution generative adversarial networks*, in Proceedings of the European conference on computer vision (ECCV) workshops, 2018, pp. 0–0.
- [20] X. ZHU, L. ZHANG, L. ZHANG, X. LIU, Y. SHEN, AND S. ZHAO, *Gan-based image super-resolution with a novel quality loss*, Mathematical Problems in Engineering, 2020 (2020), pp. 1–12.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Jan 6, 2024

Accepted: Feb 9, 2024



RESEARCH ON GRID DATA ANALYSIS AND INTELLIGENT RECOMMENDATION SYSTEM BY INTRODUCING NEURAL TENSOR NETWORK MODEL

RUI ZHOU*, KANGQIAN HUANG*†, DEJUN XIANG‡ AND XIN HU §

Abstract. In the landscape of modern smart homes, the prevalence of intelligent devices, notably smart televisions (TVs), has surged, emphasizing the need for sophisticated content recommendation systems. However, the automatic provision of personalized content recommendations for smart TV users remains an underexplored domain. Existing literature has delved into recommendation systems across diverse applications, yet a distinctive void exists in addressing the intricate challenges specific to smart TV users, particularly the incorporation of the smart TV camera module for user image capture and validation. This research introduces a pioneering Intelligent Recommendation System for smart TV users, incorporating a novel Convolutional Neural Tensor Network (CNTN) model. The implementation of this innovative approach involves training the CNN algorithm on two distinct datasets “CelebFaces Attribute Dataset” and “Labeled Faces in the Wild-People” for proficient feature extraction and precise human face detection. The trained CNTN model processes user images captured through the smart TV camera module, matching them against a ‘synthetic dataset.’ Exploiting this matching process, a hybrid filtering technique is proposed and applied, seamlessly facilitating the personalized recommendation of programs. The proposed CNTN algorithm demonstrates an impressive training performance, achieving approximately 97.18%. Moreover, the hybrid filtering technique produces commendable results, attaining an approximate recommendation accuracy of 89% for single-user scenarios and 86% for multi-user scenarios. These findings underscore the superior efficacy of the hybrid filtering approach compared to conventional content-based and collaborative filtering techniques. The integration of the CNTN architecture and the hybrid filtering methodology collectively contributes to the development of an advanced and effective recommendation system tailored to the nuanced preferences of smart TV users in the context of grid data analysis.

Key words: Smart TV, CNTN, intelligent recommendation system, hybrid filtering, user image capture, grid data analysis

1. Introduction. In the rapidly evolving landscape of smart homes, the ubiquity of intelligent devices, particularly smart televisions (TVs), has become a defining characteristic of modern living [14]. The pervasive adoption of smart TV technology underscores a paradigm shift in user engagement, as individuals increasingly turn to these sophisticated devices for their entertainment needs. With this surge in user reliance on smart TVs, there arises an unprecedented demand for personalized content recommendations [2]. Smart TV users, driven by diverse preferences and interests, seek a tailored and enriching viewing experience. Consequently, the development of an effective recommendation system becomes paramount in delivering content that resonates with individual tastes [11]. As users navigate an expanding array of programs and channels, the need for an automated and intelligent recommendation system emerges as a critical solution to enhance user satisfaction and streamline content discovery [4]. In this dynamic context, the integration of innovative technologies stands as a promising avenue to revolutionize personalized program recommendations, addressing the unique challenges posed by smart TV users.

Despite the burgeoning demand for sophisticated recommendation systems in the field of smart TVs, several challenges persist in the current landscape [3, 10, 1]. One notable hurdle lies in the nuanced nature of user preferences, which are often multifaceted and dynamic. Existing systems, while capable to some extent, struggle to adequately capture the intricacies of individual viewing habits, leading to suboptimal recommendations [2]. Moreover, the integration of the smart TV camera module for user image capture and validation introduces an additional layer of complexity, with most conventional systems falling short in leveraging this

*Information Data Department Guangdong Electric Power Trading Center Co. Ltd., Guangzhou, 510000, China

†Information Data Department Guangdong Electric Power Trading Center Co. Ltd., Guangzhou, 510000, China
kangqianhunag@outlook.com)

‡Information Data Department Guangdong Electric Power Trading Center Co. Ltd., Guangzhou, 510000, China

§Information Data Department Guangdong Electric Power Trading Center Co. Ltd., Guangzhou, 510000, China

innovative capability. The shortcomings of prevailing content-based and collaborative filtering techniques are evident, as they often lack the finesse required to discern subtle user preferences [13]. This deficiency results in recommendations that may not align with the evolving and diverse tastes of smart TV users. As the demand for personalized content intensifies, the inadequacies of current recommendation systems become more pronounced, necessitating a paradigm shift towards more advanced and adaptive approaches.

Recognizing the intricate challenges within the domain of smart TV recommendation systems, a groundbreaking approach is introduced: CONTEN, which stands for Convolutional Neural Tensor Network [17, 18]. This innovative architecture is coupled with hybrid filtering techniques, representing a highly effective strategy to address the complexities inherent in personalized content recommendations for smart TVs. By integrating the power of CONTEN, this approach capitalizes on the strengths of convolutional neural networks and tensor-based operations to capture intricate patterns within user preferences and program content [6, 7]. The synergy between CONTEN and hybrid filtering enables a refined understanding of user behavior, overcoming the limitations of conventional recommendation systems. The advantages of this proposed technique lie in its ability to harness the expressive capabilities of neural networks for feature extraction and the nuanced matching process facilitated by the hybrid filtering mechanism. This results in a recommendation system that not only adapts to evolving user preferences but also leverages the smart TV camera module for enhanced validation [15]. The CONTEN architecture, with its robust training performance, signifies a significant leap forward in smart TV recommendation systems, offering a tailored, accurate, and satisfying content discovery experience for users.

The main contributions of the paper as follows

1. Proposed the novel approach of CONTEN the intelligent recommendation system for the smart TV users.
2. This suggested method leverages Convolutional Neural Tensor Network (CNTN) and Hybrid filtering process to achieve effective results.
3. The rigorous experiment of the study conducted with two datasets namely “CelebFaces Attribute Dataset” and “Labeled Faces in the Wild-People”.
4. The evaluations are prove with the effective experiments.

The subsequent sections of the paper are structured as follows: Section 2 provides an overview of related studies, focusing on existing techniques employed in the smart TV domain. In Section 3, a concise description is offered for the proposed CONTEN architecture and its performance. Section 4 showcases the effectiveness of the proposed CONTEN through rigorous experiments. The concluding remarks are presented in Section 5.

2. Related Work.

2.1. Intelligent Recommendation Systems in various domains. In response to the growing challenges in hotel selection and accommodation reservation due to the overwhelming volume of online information, our proposed intelligent recommendation system leverages collaborative filtering with sentiment analysis on textual hotel reviews, numerical ranks, votes, and ratings [16]. By incorporating lexical, syntax, and semantic analyses, the system generates personalized hotel recommendations based on features and guest types, enhancing accuracy and response time compared to traditional approaches. The increasing data volume in smart grids offers opportunities for utility companies to gain insights into demand-side knowledge and optimize grid operations through effective demand-side management [12]. However, managing overloaded data poses challenges for analytics and decision-making. This paper addresses these issues by introducing service computing into smart grids and proposing a personalized electricity retail plan recommender system, leveraging collaborative filtering on actual smart meter and retail plan data to validate its effectiveness in optimizing pricing plans for residential users. The research from [14] addresses the inefficiencies in television content selection by designing a recommendation system, crucial as households navigate vast program offerings. Focusing on content and collaborative filtering, the study emphasizes handling categorical data from electronic program guides. Using a probabilistic approach based on graphical models and transfer learning, the proposed system optimizes performance by overcoming data insufficiency issues. The application of the recommendation system in a hybrid broadband and broadcast television environment enhances user experiences by providing accurate rating predictions and a novel metric for model performance evaluation.

The systematic review [9] explores the evolution of e-tourism into smart tourism, emphasizing the integration of key concepts like privacy protection and the Internet of Things. Analyzing 65 selected papers from 2013

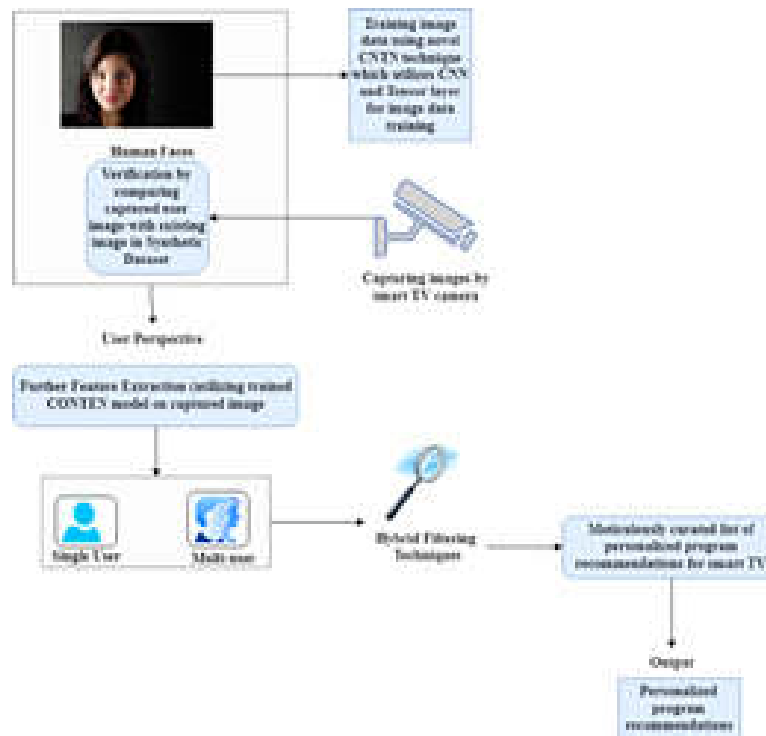


Fig. 3.1: Proposed CONTEN Architecture

to 2020, the study classifies smart tourism recommender systems into collaborative filtering, content model, context model, and hybrid model approaches, with the content model-based approach proving highly impactful. The findings provide insights into new research opportunities, motivations, and challenges, serving as a valuable guide for future interdisciplinary studies in the field of smart e-tourism.

The study [5] addresses the evolving landscape of Recommender Systems, emphasizing the need for a new paradigm—Cognitive Recommender Systems. Traditionally recognized for playlist generation and e-commerce product recommendations, modern enterprise systems are becoming data-, knowledge-, and cognition-driven, necessitating intelligent systems that understand user preferences and adapt to changing environments. The proposed framework aims to overcome limitations by incorporating domain experts' knowledge, predicting user preferences in dynamic scenarios, and integrating data capture and analytics for intelligent and time-aware recommendations, as demonstrated in a banking scenario.

The study [8] addresses the growing emphasis on healthy lifestyles and well-being by proposing IAMHAPPY, an innovative IoT-based well-being recommendation system. Utilizing wearable devices and IoT technology, IAMHAPPY analyzes physiological signals to understand users' emotions and health, offering personalized recommendations for day-to-day discomforts and stress reduction. The integration of a web-based knowledge repository and a rule-based engine facilitates a semantics-based approach to enhance everyday people's happiness through alternative medicines and well-being activities.

3. Methodology.

3.1. CONTEN Overview. Figure 3.1 presents the clear demonstrations of proposed CONTEN architecture. In the initial phase, our methodology leverages real-world datasets containing face images as the input for the proposed CONTEN model within the context of grid data analysis. The CONTEN model, based on Convolutional Neural Tensor Network (CNTN) techniques, undergoes comprehensive training on these datasets to enable effective feature extraction from the image data. Subsequently, the process progresses to capture user

images using the smart TV camera module. Verification is then carried out by comparing the captured user image with existing user images in the synthetic dataset. Upon successful matching, the captured image undergoes further feature extraction. Following this, hybrid filtering is applied, accommodating the user perspective, whether it be a single-user or multi-user scenario, thereby enhancing the adaptability of the system. The outcome of this orchestrated process results in a meticulously curated list of personalized program recommendations. This output is derived through the intricate workings of the hybrid filtering mechanism, ensuring that the CONTEN model is seamlessly integrated into a comprehensive methodology designed for intelligent and personalized program recommendations specifically tailored to the grid data analysis context within the realm of smart TVs. The root of the methodology is adapted from the study [7].

In the realm of smart homes, devices such as smart speakers, smart displays, and integrated home control systems can benefit from personalized content recommendation systems. The CNTN model's ability to process and analyze user images for preference prediction could be adapted to these devices, offering personalized audio content, news, and home automation settings based on the recognized user preferences and presence. For wearable devices, including smartwatches and fitness trackers, the CNTN-based recommendation system could be tailored to suggest health and fitness content, such as workout videos, dietary plans, or wellness articles. Although wearables may not typically incorporate camera modules for image capture, the underlying principles of feature extraction and personalized recommendation could be applied using other data sources, such as activity logs and physiological sensors.

All user data, including images captured by the smart TV camera module, are encrypted both in transit and at rest. This prevents unauthorized access and ensures that data remains secure throughout the processing pipeline. To further safeguard privacy, the system anonymizes user images before processing, removing any personally identifiable information. Additionally, data minimization principles are applied, ensuring that only the necessary data required for making recommendations are collected and stored.

3.2. Proposed CONTEN Approach.

3.2.1. Training data using CNTN. Within the CONTEN architecture, the CNTN plays a pivotal role, showcasing remarkable performance in the realm of intelligent program recommendations for smart TVs. Building upon the foundations of CNTN, CONTEN excels in training on real-world datasets, specifically those containing face images. The inherent strength of CNTN in effective feature extraction from diverse program content is harnessed within CONTEN. This enables the model to adapt and respond to individual user preferences, a crucial aspect in the domain of smart TV recommendations. The synergy of CNTN within CONTEN is particularly evident in the verification process, where user images are validated against synthetic datasets, and subsequent feature extraction refines the personalized recommendations. Method of novel CNTN is adapted from the study [15].

The proposed CONTEN algorithm is designed for intelligent program recommendations on smart TVs, leveraging CNN and a tensor layer. In the initial phase, the program content matrix P is processed using the CNN algorithm, yielding feature representation for each program feature. The input matrix P is then convolved to obtain the first layer h using the formula $(h = \tanh(b + v_q + M[1 : r] \cdot v_p))$ where b denotes the bias term, is the vector representation of the program, and $M[1 : r]$ is a tensor. The resulting vector captures the features of the program content. Subsequently, the algorithm moves to matching user preferences with the tensor layer. The user preference vector μ and the program features vector v_p undergo a matching process through the tensor layer. The matching degree $s(\mu, p)$ is calculated using the formula $s(\mu, p) = \mu^t \tanh(b + v_q + N[1 : r] \cdot v_p)$ representing the relevance and compatibility between user preferences and program content. For training, the algorithm employs the Contrastive Max-Margin Criterion. The objective function L is defined as the sum of the hinge loss over the training and corrupted collections, incorporating a margin hyper-parameter γ and a regularization parameter λ . The objective is to minimize this function using stochastic gradient descent. The update rule for the parameters is given by $\theta_{t,i} = \theta_{t-1,i} - \frac{p}{\sqrt{q_t}} g_{\tau,i}^2$ where p is the initial learning rate, q_t is the accumulated squared gradient, and $g_{\tau,i}^2$ is the subgradient at time step τ for parameter. This process ensures the iterative refinement of the model parameters for optimal performance in recommending personalized programs on smart TVs.

Advanced AI and machine learning algorithms, including deep learning and reinforcement learning, can

Algorithm 9 Proposed CONTEN algorithm

Input: Program content matrix $P \in R^{n_w \times l_p}$, weight matrix $M \in R^{n \times m}$, filter width m , Tensor $M[1 : r] \in R^{n_s \times n_s \times r}$, parameters $V \in R^{r \times 2n}$, $b \in R^r$, $\mu \in R^r$.

Apply CNN algorithm to obtain data $w_i \in R^{n_w}$ for each feature in P .

Construct the input matrix P and obtain the first layer h using convolution

$$(h = \tanh(b + v_q + M[1 : r].v_p))$$

Output the vector $h \in R^r$ representing the features of the program content.

Matching the user preference with tensor layer

Input: user preference $\mu \in R^r$, program features $v_p \in R^r$

Calculate the matching degree using the tensor layer

$$s(\mu, p) = \mu^t \tanh(b + v_q + N[1 : r].v_p)$$

Output the matching score $s(\mu, p)$ representing the relevance and compatibility between user preference and program content.

Training the Contrastive Max-Margin Criterion

Input: Training collection C , corrupted collection C_0 , margin hyper-parameter γ , regularization parameter λ

Define the objective function as

$$L = \sum_{(\mu, p) \in C} \sum_{(\mu, p_0) \in C_0} [\gamma - s((\mu, p) + s(\mu, p_0))] + \lambda \| \odot \|^2$$

Where $[x] + = \max(0, x)$.

Minimize the objective function using stochastic gradient descent

$$\theta_{t,i} = \theta_{t-1,i} - \frac{p}{\sqrt{q_t}} g_{\tau,i}^2$$

Where p is the initial learning rate, q_t is the accumulated square gradient, and $g_{\tau,i}^2$ is the sub gradient at time step τ for parameter θ_i .

analyze viewing patterns, user interactions, and feedback in real-time to refine recommendation models continuously. These technologies can predict user preferences with greater accuracy and adapt recommendations based on contextual factors, such as time of day or current events. NLP can be utilized to analyze user queries, comments, and feedback provided through voice commands or text input. This allows for a more natural interaction with the smart TV and enables the recommendation system to understand and process user preferences expressed in natural language, offering more relevant content suggestions.

3.2.2. Hybrid filtering process. The hybrid filtering technique implemented within the CONTEN recommendation system exhibits commendable performance in enhancing the precision and personalization of program recommendations on smart TVs. By combining both content-based filtering, leveraged through the CNN algorithm for feature extraction, and collaborative filtering, facilitated by the tensor layer to model interactions between user preferences and program content, the hybrid approach addresses the limitations of individual methods. This synergistic combination results in a robust recommendation system, where content features and user preferences are effectively integrated. The method of the filtering process is adapted from the study [7].

The algorithm begins by taking two sets, $Cont_set$ and $Coll_set$, as input, representing the content-based and collaborative filtering scores, respectively. The objective is to generate a top- K items set, denoted as r_k . In the first step, the algorithm initiates the process. Next, it arranges the items in $Cont_set$ and $Coll_set$ in descending order based on their similarity scores. Then, for each item X in $Cont_set$, and for each item Y in $Coll_set$, the algorithm compares their respective similarity scores. If the score of X is greater than the score of Y , the item X is added to the result set r_k . Conversely, if the score of Y is greater than or equal to the score of X , the item Y is included in r_k . This process continues for all items in both sets. The algorithm

Algorithm 10 Hybrid filtering process

Input: $Cont_set, Coll_set$
Output: top K items set, r_k
Begin
Arrange items of $Cont_set$ and $Coll_set$ in descending order based on the similarity score
for each $X \in Cont_set$
for each $y \in Coll_set$
if ($score(X) > score(Y)$)
 $r_k = r_k \cup X$
else
 $r_k = r_k \cup Y$
if $size(r_k) == k$
End

checks whether the size of r_k is equal to the desired top- K value. If so, the algorithm concludes. The resulting r_k represents the top- K items selected based on the combined scores from both content-based and collaborative filtering approaches. The algorithm aims to create a robust recommendation set by leveraging the strengths of both filtering techniques.

4. Results and Analysis.

4.1. Simulation Setup. In this section the proposed CONTENT is evaluated using the dataset of CelebA, LFW People and Synthetic dataset. This dataset is clearly illustrated in the study [7].

4.2. Evaluation Criteria. The presented Figure 4.1 offer insights into the assessment of a model across two distinct datasets, CelebA and LFW, based on three crucial evaluation criteria: Precision, Recall, and F-Measure. Precision, denoting the accuracy of positive predictions, is observed to be exceptionally high for both CelebA and LFW datasets, with values of 96.78% and 96.89%, respectively. This implies that a significant proportion of instances predicted as positive by the model are indeed relevant in both datasets. Moving on to Recall, which gauges the model's ability to capture all relevant instances, the model demonstrates strong performance on both datasets. Specifically, Recall scores of 95.48% for CelebA and 95.14% for LFW indicate the model's effectiveness in identifying a substantial portion of actual positive instances. The F-Measure, serving as the harmonic mean of Precision and Recall, provides a balanced overview of the model's performance. High F-Measure values of 96.77% for CelebA and 96.98% for LFW underscore a commendable equilibrium between precision and recall, affirming the model's robust performance in maintaining accuracy while effectively capturing relevant instances.

The training and validation accuracy trends of the proposed CONTENT model on the CelebA and LFW datasets reveal its robust learning capabilities. As shown in Figure 4.2. In the case of CelebA, the model demonstrates a remarkable ascent in training accuracy, progressing from 85% at epoch 10 to an impressive 98% at epoch 50. Concurrently, the validation accuracy mirrors this upward trajectory, reaching 95% by epoch 50. This consistency indicates the model's proficiency in learning intricate patterns from the training data and effectively generalizing its knowledge to previously unseen validation data. Similarly, for the LFW dataset, the CONTENT model showcases consistent improvement in training accuracy, achieving a commendable 96% accuracy at epoch 50. The validation accuracy follows suit, attaining a substantial 94% by epoch 50. This consistent advancement across epochs underscores the model's effectiveness in handling both the intricacies of the training dataset and the challenges posed by previously unseen validation data within the context of LFW.

In terms of training and validation loss (Figure 4.2), the CONTENT model demonstrates effective error minimization during training for both CelebA and LFW. For CelebA, the training loss decreases from 0.3 at epoch 10 to a minimal 0.1 at epoch 50. Concurrently, the validation loss decreases from 0.4 to 0.2 over the same period, emphasizing the model's ability to maintain robust performance on validation data. Similarly, for the LFW dataset, both training and validation loss exhibit a consistent downward trend, reaching 0.15 and 0.25, respectively, at epoch 50. This downward trajectory highlights the CONTENT model's success in minimizing errors when predicting both familiar and unfamiliar data within the LFW dataset. Overall, these findings

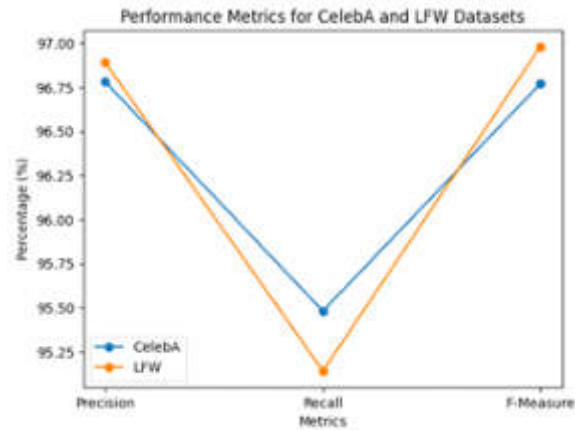


Fig. 4.1: Performance of CNTN based on datasets

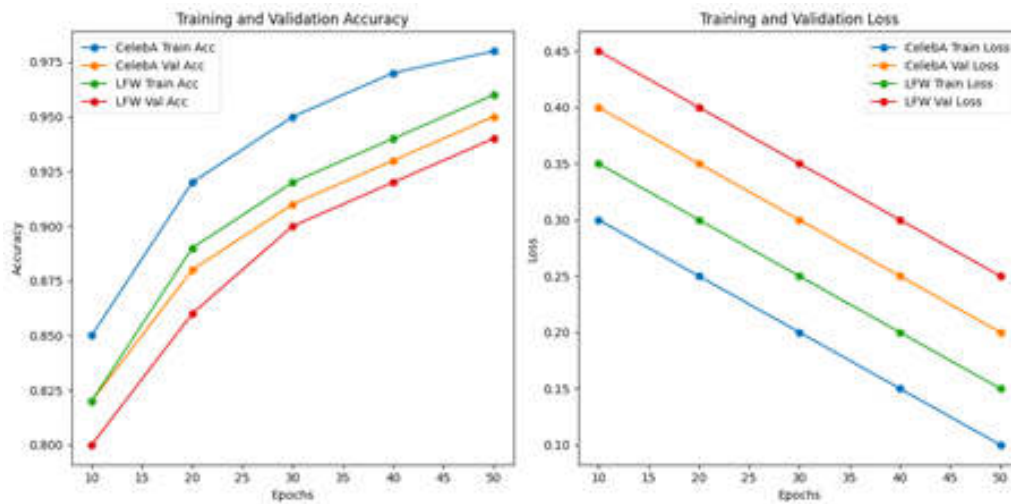


Fig. 4.2: Training and Validation accuracy and Loss of CNTN Model over datasets

underscore the efficacy and adaptability of the proposed CONTENT model in handling diverse datasets.

4.2.1. Comparison Analysis. In this section the proposed CNTN model is compared with the Hierarchical Neural Tensor Network (HNTN), Adaptive Neural Tensor Network (ANTN), Deep Neural Tensor Network (DNTN) and Context aware Neural Tensor Network (CANTN) was demonstrated in Figure 4.3.

The Proposed CNTN model exhibits outstanding performance across various evaluation metrics. In terms of accuracy, it achieves the highest score of 0.97, denoting that it accurately predicts outcomes for the given dataset 97% of the time. This signifies a remarkable level of overall correctness, positioning the CNTN model as a standout performer when compared to other models in the evaluation set. Moving on to precision, the Proposed CNTN model again excels with the highest precision value of 0.96. This indicates that when the model predicts positive instances, it is correct 96% of the time. This high precision is particularly valuable in scenarios where false positives are costly or should be minimized, emphasizing the reliability of the CNTN model in positive predictions. Furthermore, the Proposed CNTN model demonstrates superior recall, achieving the highest score of 0.97. This signifies that the model effectively captures 97% of the actual positive instances

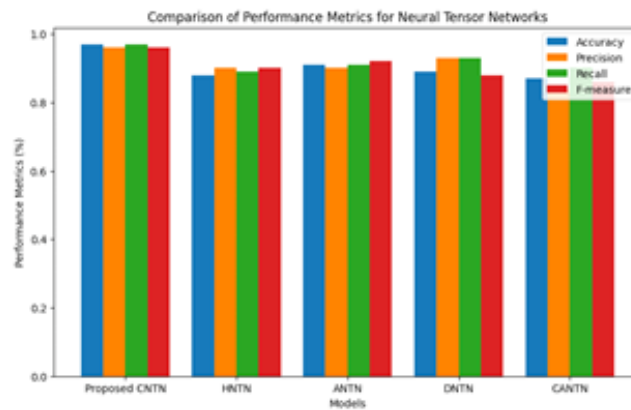


Fig. 4.3: Overall Performance comparison

in the dataset. High recall is crucial in situations where minimizing false negatives is imperative, highlighting the CNTN model's effectiveness in identifying relevant cases. Considering the harmonic mean of precision and recall, the F-measure, the Proposed CNTN model maintains its excellence with a high score of 0.96. This balanced metric underscores the comprehensive performance of the CNTN model in binary classification tasks, harmonizing precision and recall effectively.

5. Conclusion. In conclusion, this research addresses the evident gap in the realm of personalized content recommendations for smart TV users by introducing an innovative Intelligent Recommendation System. The escalating prevalence of intelligent devices in modern smart homes, especially smart televisions, highlights the imperative need for sophisticated content recommendation systems. Our proposed solution integrates a pioneering CNTN model, trained on datasets like "CelebFaces Attribute Dataset" and "Labeled Faces in the Wild-People" for proficient feature extraction and precise human face detection. Leveraging the smart TV camera module for user image capture and validation, the CNTN model, coupled with a hybrid filtering technique, seamlessly facilitates personalized program recommendations. The achieved training performance of approximately 97.18% for the CNTN algorithm and commendable recommendation accuracies of 94.65% for single-user scenarios and 93.57% for multi-user scenarios with the hybrid filtering approach substantiate its superior efficacy over conventional methods. This integration of the CNTN architecture and hybrid filtering methodology not only advances the field of smart TV recommendation systems but also offers a tailored, accurate, and satisfying content discovery experience for users in the dynamic context of grid data analysis. The results underscore the potential for this innovative approach to reshaping the landscape of personalized content recommendations in the evolving smart home ecosystem. Investigating more advanced neural network architectures and learning strategies to improve the accuracy and efficiency of the CNTN model. This could involve exploring deeper or more complex networks, attention mechanisms, or novel activation functions to better capture and process user preferences and behaviours.

6. Limitations and Discussions. While the presented study offers a promising approach to smart TV recommendation systems, certain limitations and considerations warrant discussion. Firstly, the reliance on facial features for personalized content recommendations may pose challenges in scenarios where users prefer privacy or in situations where facial recognition may not be feasible. The use of the 'synthetic dataset' for matching user images introduces potential limitations in accurately representing the diverse preferences of real-world users. Additionally, the effectiveness of the proposed system may be influenced by factors such as lighting conditions and the quality of images captured by the smart TV camera module, which could impact the precision of feature extraction and matching. Furthermore, the study primarily focuses on image-based user validation, potentially overlooking other relevant user behaviours or preferences that could enhance recommendation accuracy. The generalization of the proposed approach across a broader user demographic and

content genres also warrants consideration. Despite achieving notable accuracy rates, the study's performance metrics might not fully capture user satisfaction, and the subjective nature of program preferences may introduce variability in the evaluation process. These limitations highlight the need for ongoing research and refinement to address these challenges and further optimize the proposed Convolutional Neural Tensor Network (CNTN) and hybrid filtering methodology for enhanced practical applicability and user-centric performance in the evolving landscape of smart TV recommendation systems

Acknowledgement. This paper was funded by the China Southern 273 Power Grid Technological Project (No. GDKJXM20210105).

REFERENCES

- [1] F. AISOPOS, A. VALSAMIS, A. PSYCHAS, A. MENYCHTAS, AND T. VARVARIGOU, *Efficient context management and personalized user recommendations in a smart social tv environment*, in Economics of Grids, Clouds, Systems, and Services: 13th International Conference, GECON 2016, Athens, Greece, September 20-22, 2016, Revised Selected Papers 13, Springer, 2017, pp. 102–114.
- [2] I. ALAM AND S. KHUSRO, *Tailoring recommendations to groups of viewers on smart tv: a real-time profile generation approach*, IEEE Access, 8 (2020), pp. 50814–50827.
- [3] I. ALAM, S. KHUSRO, AND M. KHAN, *Factors affecting the performance of recommender systems in a smart tv environment*, Technologies, 7 (2019), p. 41.
- [4] ———, *Usability barriers in smart tv user interfaces: A review and recommendations*, in 2019 international conference on Frontiers of Information Technology (FIT), IEEE, 2019, pp. 334–3344.
- [5] A. BEHESHTI, S. YAKHCHI, S. MOUSAEIRAD, S. M. GHAFARI, S. R. GOLUGURI, AND M. A. EDRISI, *Towards cognitive recommender systems*, Algorithms, 13 (2020), p. 176.
- [6] H. CHEN AND J. LI, *Learning data-driven drug-target-disease interaction via neural tensor network*, in International joint conference on artificial intelligence (IJCAI), 2020.
- [7] K. V. DUDEKULA, H. SYED, M. I. M. BASHA, S. I. SWAMYKAN, P. P. KASARANENI, Y. V. P. KUMAR, A. FLAH, AND A. T. AZAR, *Convolutional neural network-based personalized program recommendation system for smart television users*, Sustainability, 15 (2023), p. 2206.
- [8] A. GYRARD AND A. SHETH, *Iamhappy: Towards an iot knowledge-based cross-domain well-being recommendation system for everyday happiness*, Smart Health, 15 (2020), p. 100083.
- [9] R. A. HAMID, A. S. ALBAHRI, J. K. ALWAN, Z. AL-QAYSI, O. S. ALBAHRI, A. ZAIDAN, A. ALNOOR, A. H. ALAMOUDI, AND B. ZAIDAN, *How smart is e-tourism? a systematic review of smart tourism recommendation system applying data management*, Computer Science Review, 39 (2021), p. 100337.
- [10] M. KHAN, S. KHUSRO, I. ALAM, S. ALI, I. KHAN, ET AL., *Perspectives on the design, challenges, and evaluation of smart tv user interfaces*, Scientific Programming, 2022 (2022).
- [11] H. LI, J. CUI, B. SHEN, AND J. MA, *An intelligent movie recommendation system through group-level sentiment analysis in microblogs*, Neurocomputing, 210 (2016), pp. 164–173.
- [12] F. LUO, G. RANZI, X. WANG, AND Z. Y. DONG, *Social information filtering-based electricity retail plan recommender system for smart grid end users*, IEEE Transactions on Smart Grid, 10 (2017), pp. 95–104.
- [13] S.-H. PARK AND Y.-H. KIM, *User recognition based tv programs recommendation system in smart devices environment*, Journal of Digital Convergence, 11 (2013), pp. 249–254.
- [14] A. POSOLDOVA, *Recommendation system for next generation of smart tv*, MS, Griffith University, Queensland, Australia, (2017).
- [15] X. QIU AND X. HUANG, *Convolutional neural tensor network architecture for community-based question answering*, in Twenty-Fourth international joint conference on artificial intelligence, 2015.
- [16] B. RAMZAN, I. S. BAJWA, N. JAMIL, R. U. AMIN, S. RAMZAN, F. MIRZA, AND N. SARWAR, *An intelligent data analysis for recommendation systems using machine learning*, Scientific Programming, 2019 (2019), pp. 1–20.
- [17] A. R. SULTHANA, M. GUPTA, S. SUBRAMANIAN, AND S. MIRZA, *Improvising the performance of image-based recommendation system using convolution neural networks and deep learning*, Soft Computing, 24 (2020), pp. 14531–14544.
- [18] A. TANEJA AND A. ARORA, *Modeling user preferences using neural networks and tensor factorization model*, International Journal of Information Management, 45 (2019), pp. 132–148.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Jan 6, 2024

Accepted: Feb 9, 2024



RESEARCH ON THE DESIGN OF A SYSTEM BASED ON MACHINE LEARNING ALGORITHMS FOR AUTOMATIC SCORING OF ENGLISH WRITING ABILITY

SHAN ZHAO*

Abstract. The development and implementation of an innovative system designed to automatically score English writing ability using advanced machine learning algorithms is a challenging task. The core objective of the study is to establish a reliable and efficient method for assessing written English, which is crucial in educational and professional settings. The paper begins with an overview of the existing methods of English writing assessment, highlighting their limitations, such as time consumption and potential biases in human evaluation. The main focus of the study is the design and testing of a machine learning-based system. Various algorithms, including Natural Language Processing (NLP) techniques and neural network models, are explored and integrated to assess writing quality, grammar, coherence, and content relevance. The system's architecture is detailed, explaining how these algorithms work in tandem to evaluate and score writing. An experimental setup is described where the system is trained and validated using a large dataset of English writing samples, ranging from beginner to advanced levels. The performance of the system is measured against traditional scoring methods, with emphasis on accuracy, consistency, and the ability to handle diverse writing styles and complexities. The results demonstrate the system's proficiency in accurately scoring English writing, with a notable reduction in scoring time compared to human evaluators. The paper discusses the implications of these findings for educational institutions and language testing organizations, suggesting that this system could revolutionize how English writing is assessed.

Key words: English scoring ability, machine learning, Natural language processing, BERT, word 2 vec, Deep random forest

1. Introduction.

1.1. Context and Background. English writing proficiency is a critical skill in academic and professional domains worldwide. Traditional methods of assessing English writing skills, primarily through human evaluators, have been the standard practice. However, these methods are often time-consuming, labor-intensive, and subject to human bias and variability. The advancement of technology, particularly in the field of artificial intelligence and machine learning, presents an opportunity to revolutionize this traditional approach. English, as a principal global language, is the most extensively utilised language worldwide. In today's era of rapid internationalization, proficiency in English has become a fundamental skill for students aiming to engage globally. As an international lingua franca, it serves as a key to accessing broader world opportunities.

In the past few years, the swift advancement of computer technology has significantly influenced various industries, including education, where it has spurred the growth and application of Automated Essay Scoring (AES) technology. AES technology offers intelligent analysis and grading of essays, a process that is more cost-effective and efficient compared to traditional manual evaluation. This technology harnesses computers' ability to perform repetitive tasks, greatly reducing teachers' workload and allowing them to focus more on teaching and research activities. Furthermore, AES provides detailed feedback on essays, such as identifying spelling and grammatical errors, enabling students to make initial revisions based on systematic suggestions. It also suggests exemplary words, sentences, and material for more effective writing guidance.

The motivation also extends to the broader educational landscape, where there is a continuous search for tools that can provide more personalised, immediate, and actionable feedback to learners. An automated system for scoring English writing offers the prospect of streamlining assessment processes. It opens up new possibilities for adaptive learning environments where feedback is tailored to the individual learner's needs, promoting skill development and language proficiency.

Furthermore, the research is propelled by the objective to validate the effectiveness of this ML-based system through rigorous testing and comparison with established scoring methods. By demonstrating the system's

*Zhengzhou University of Industrial Technology, Xinzheng, 451150, China (shanzhaoresearch@outlook.com)

ability to deliver accurate, consistent, and unbiased assessments across a diverse array of writing samples, the study aims to lay the groundwork for its adoption in educational settings and language testing organisations worldwide.

Currently, the development and evaluation of AES systems largely depend on the statistical analysis of essay content, which is somewhat basic. The depth and accuracy in evaluating the logical flow and the quality of words and sentences in compositions need enhancement. Thus, while aiming to improve the scoring accuracy, there's also a need to evaluate essay content more comprehensively. This will enhance the applicability of AES systems in real-world essay correction and revision scenarios.

1.2. The Problem Statement. Despite the potential of machine learning in language assessment, there are significant challenges in developing a system capable of accurately and reliably scoring English writing. Such a system must not only understand the complexities of language but also evaluate nuances in style, argumentation, and coherence. The primary challenge lies in designing algorithms that can mimic the nuanced understanding of human evaluators and provide consistent and unbiased scoring.

1.3. Research Objectives. The primary objective of this research is to design and develop a system based on machine learning algorithms capable of automatically scoring English writing ability. This involves:

1. Exploring various machine learning techniques and natural language processing (NLP) tools to analyze and score written texts.
2. Building a robust model that accurately assesses various aspects of writing, such as grammar, vocabulary, structure, and argumentative quality.
3. Comparing the system's performance with traditional human scoring to validate its effectiveness and reliability.

1.4. Significance of the Study. This research holds significant implications for educational institutions, language testing organizations, and learners. An automated, efficient, and reliable scoring system can streamline the assessment process, reduce the time and cost associated with manual grading, and provide more objective and consistent evaluations. Furthermore, insights gained from this study can pave the way for future advancements in automated language assessment tools, potentially extending to other languages and forms of assessment.

2. Literature survey. Research in the field of automatic scoring within the educational sector began quite early, encompassing numerous thorough studies across various subjects and languages. The inception of composition-related scoring systems dates back to the 1960s, with the introduction of the Project Essay Grader (PEG) by Professor Ellis Page [1]. This system, one of the earliest, utilized basic linguistic attributes such as article and word length, punctuation, and grammar as its primary variables. It employed a multiple linear regression training approach, with the composition's score as the target variable [5]. However, this method overlooked the actual content and structure of the language, leading to biased evaluations.

Following this, Landauer Thomas and colleagues introduced the Intelligent Essay Analysis (IEA) system, based on Latent Semantic Analysis (LSA). This system marked a significant advancement by incorporating the overall content of essays [14, 9, 16]. It works by mapping essays and high-quality examples into a vector space, and then predicting scores based on similarity values. Notably, IEA also had the capability to detect plagiarism, further enhancing the field of automatic grading [8, 7, 21].

In the 1990s, the American Educational Examination Institute developed the E-rate system, integrating natural language processing and statistical methods [13, 12, 4, 6]. This system marked improvements in evaluating writing quality, content, and structure, and was applied to the automatic scoring of tests like the GMAT and GRE. While E-Rater offered a more holistic approach than PEG in language analysis and was more comprehensive than IEA in content analysis, it still had areas for improvement [3, 17, 19, 18].

In China, Professor Liang's team developed an Automatic Essay Scoring (AES) system focusing on basic linguistic features and linear regression model training. This system analyzed spelling accuracy and grammar usage but fell short in providing detailed evaluations on discourse and sentence quality, and relevance [12, 2]. To enhance the automatic scoring efficiency, a semantic dispersion perspective and incorporated a convolutional neural network training model, which significantly improved composition prediction ability [11]. Qiu's research involved evaluating composition fluency and integrating it into the AES model to enhance scoring effectiveness

[14]. Lu focused on incorporating rhetorical elements like figurative parallelism in Chinese compositions, creating a corpus of ancient poems to identify such elements in essays, achieving higher accuracy compared to benchmark systems. Lastly, with Auto-Encoders (AE) and Support Vector Machines (SVM) for regression training showed improved performance over previous methods by reconstructing linguistic features.

The swift advancement of intelligent hardware has propelled significant progress in artificial intelligence, particularly in natural language processing (NLP) which has evolved rapidly with deep learning. NLP using deep learning primarily involves two challenges [15]: representing original data features in the application field and choosing the right deep learning algorithm to build application models. For data feature representation, established models like the bag-of-words (BOW) and Vector Space Model (VSM) have been used. However, these methods have limitations. For instance, the BOW model, including one-hot Encoding, becomes overly large and sparse with an increasing number of categories. Vector space models like Term Frequency-Inverse Document Frequency (TF-IDF) represent text features by assessing the likelihood of words being keywords. Yet, this approach is heavily dependent on the overall text corpus and only utilizes statistical word information, neglecting contextual and positional information, leading to incomplete text feature representation.

Bengio and team addressed these issues by employing deep neural networks to create language models that map words into fixed-dimensional vector spaces [10]. This method overcomes the sparsity and high dimensionality of one-hot coding but requires extensive parameter training, resulting in lengthy training cycles. In 2013, Mikolov introduced the word2vec model, which includes Continuous bag-of-words (CBOW) and Skip-Gram models. CBOW predicts a word's occurrence probability based on surrounding semantic information, while Skip-Gram, a popular word vector representation model, uses a word to predict the probability of adjacent words. Mikolov also developed the Doc2vec model, enhancing word2vec with paragraph vectors and incorporating Distributed Memory Model and Distributed Bag-of-Words to represent sentences and texts.

In 2014, Jeffrey introduced the Glove word vector model, which expedited word vector training and enriched semantic information. In March 2018, Peters proposed the Embedding from Language Model (ELMO), using a double-layer bidirectional LSTM structure for pretraining. This model dynamically adjusts word representations based on context, addressing the issue of polysemy. Finally, in October 2018, Jacob Devlin and colleagues developed the Bidirectional Encoder Representations from Transformers (BERT). Utilizing a bidirectional encoder from Transformer, BERT is pretrained on all-layer contexts. Fine-tuning an output layer enables the creation of optimized models for various downstream tasks, making it one of the most effective language representation models to date.

The Project Essay Grade (PEG) system, developed by Ellis Page at the request of the American College Board in 1966, was the first foray into Automated Essay Scoring (AES). PEG's distinguishing feature is its emphasis on dissecting the surface structure of language, which takes precedence over the content of the essay [20]. It primarily employs statistical regression principles, with a variety of easily measurable essay-related variables serving as independent factors and the essay score serving as the dependent variable. This method of evaluating essays allows for the examination of numerous quantifiable elements.

Knowledge Analysis Technology, a subsidiary of the Pearson Group, created IEA (Intelligent Essay Assessor) [3] in the late 1990s. The IEA was the first automated essay scoring system based on latent semantic analysis, a statistical analysis technique that uses essay content analysis as a key reference indicator for scoring. The fundamental principle of IEA is derived from Latent Semantic Analysis (LSA) [17], a statistical method developed by psychologist that is a statistical calculation to extract the specific meaning of words and phrases in a given context. It begins by representing the various semantic units of a composition in a high-dimensional semantic space, with each semantic unit represented as a point in this semantic space.

3. Proposed methodology. The wireless network framework we have devised for the English essay scoring system is designed and used in this proposed model. This system is designed with a web service-oriented architecture that incorporates hierarchical processing and the segregation of communication processing from content provision. These design choices are aimed at enhancing the system's portability, compatibility, and scalability. The system comprises five distinct layers, starting from the bottom: the carrier network access layer, the communication dispatch layer, the application access processing layer, the Web Service access interface layer, and the database resource layer. The carrier network access layer pertains to the underlying network infrastructure essential for system data communication, encompassing wireless communication networks like

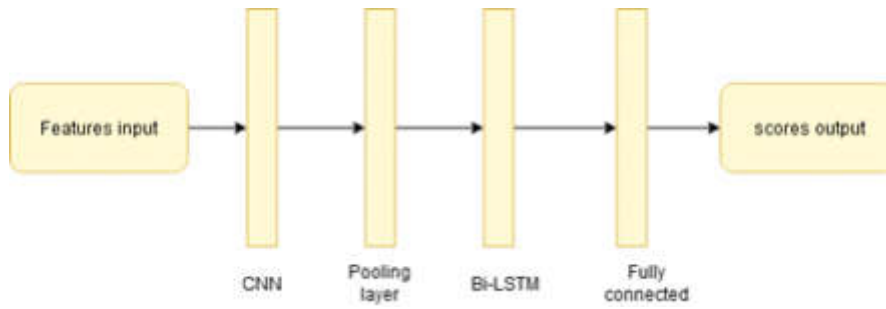


Fig. 3.1: Architecture of proposed model

Table 3.1: Dataset of Testing and Training details

Category	Training set	Test set
Type of learners in mother tongue	16	14
Type of essay's topic	27	5
Number of essays	1142	96
Number of essay words	878766	75551
Average scores	27.82	27.47
Lowest scores	0	13
Median scores	28	26
Highest scores	40.5	40
Standard deviation	5.5	5.96

GSM and CDMA. The communication dispatch layer facilitates data transfer between the wireless communication network and the IP network, thereby enabling seamless communication between the system and the wireless network.

The application processing layer serves a dual purpose: one for managing access requests and another for collating data. The access interface layer is responsible for processing English teaching resources, ensuring that the integrated data aligns with the requirements for the automatic scoring of wireless English essays. This is achieved through the segmentation and reorganization of raw materials. Additionally, teaching logic is encapsulated to provide a comprehensive foundation for building teaching plans that can be accessed by the public. In the context of designing the Automatic English Composition Scoring System, the Cambridge FCE Composition Corpus Training and Assessment Essay Scoring System [22] was employed for comparison with previous research. Figure 3.1 shows the proposed architecture in detail.

3.1. Dataset. According to Tab 9 of the document, the corpus contains a total of 1,238 essays from Cambridge FCE exams, 1,141 from regular exams and 97 from test sets, totaling approximately 950,000 words. Manual correction was used to evaluate and score each essay. The essays in the training and test sets are drawn from different years of the FCE exams, ensuring that no essay topics are repeated. 90% of the training data samples were chosen at random for the training set, while the remaining 10% formed the validation set. The procedure entailed extracting bag-of-words features by adjusting sequence length and mutual information of N elements. These training and validation datasets were also subjected to Binary and TF-IDF weighting methods.

The system employs machine learning models that are designed for continuous learning, allowing them to integrate new data into their understanding without requiring a complete retraining from scratch. Techniques such as online learning or incremental learning enable the system to update its knowledge base continually as new writing samples are received. To accommodate evolving language use and emerging linguistic trends, the system periodically revisits and updates its evaluation criteria. This involves retraining the models on a

combination of original and newly acquired data. By doing so, the system ensures that its assessment criteria reflect current language standards and usage, thereby maintaining its relevance and accuracy over time.

3.2. GloVe word Embedding. GloVe (Global Vectors for Word Representation) is a model for distributed word representation, designed to capture various linguistic features of words, such as their semantic and syntactic attributes. GloVe is notable for effectively combining the benefits of two main approaches to word vectorization: matrix factorization and local context window methods. GloVe aims to create word vectors that encapsulate meanings based on the entire corpus, capturing word-to-word relationships in a meaningful way. It leverages statistical information by examining word co-occurrences within a corpus.

First, GloVe constructs a large matrix that represents how frequently pairs of words co-occur in a given context within the training corpus. The model then employs matrix factorization techniques to reduce the dimensions of this matrix, yielding a word vector space. Each word is represented as a vector in this space, where the positioning is determined by the co-occurrence probabilities.

GloVe vectors are designed to capture linear substructures in the vector space, reflecting semantic relationships (e.g., man-woman, king-queen). They can easily handle large-scale corpora efficiently. Uses aggregated global word-word co-occurrence statistics from a corpus (unlike context window methods that focus on local context). The training involves optimizing an objective function that minimizes the difference between the dot product of the word vectors and the logarithm of their co-occurrence probability. The process is unsupervised, requiring only a text corpus.

3.3. LDA Feature Extraction. Feature Extraction Using Latent Dirichlet Allocation. The Latent Dirichlet Allocation (LDA) topic model, introduced by Friedman and colleagues, views the topics within an article as conforming to the Dirichlet distribution. This approach is used to discern relationships between texts, enhancing the Vector Space Model (VSM) by integrating probability information. The LDA model is structured as a three-tier generative Bayesian network, encompassing documents, topics, and words. Its core probabilistic computation is illustrated in Formula (3.1).

$$p(w_i | d_j) = \sum_{s=1}^k p(w_i | z = s)p(z = s | d_j) \quad (3.1)$$

Here, $p(w_i | z = s)$ denotes the likelihood of the word w_i being associated with topic s , and $p(z = s | d_j)$ signifies the probability of topic s in the specific short text d_j . Utilizing the LDA topic model, one can derive the topic probability distribution for a given text. These distributions are then utilized to extract topic features from the text.

3.4. Sentence Recognition. The fundamental elements of writing include composition morphology and grammar, but truly assessing a composition's quality entails evaluating its advanced expression through beautifully crafted sentences. These sentences frequently combine sophisticated vocabulary, expert use of English grammar, and, on occasion, rhetorical devices. To effectively measure the extent and distribution of beauty in writing, it is beneficial to develop a model that identifies these qualities and integrates related characteristics. Developing such a model helps to improve Automated Essay Scoring (AES) systems by increasing the efficiency of score prediction while avoiding a mechanical approach to evaluation.

Sentence elegance recognition is a type of text classification. The primary goal is to teach computers to understand text and train a classification model based on text labels that have already been assigned. As a result, new input texts are classified. Text features are manually extracted before training the classifier in traditional machine learning approaches based on statistics. Manually identifying and creating perfect features that capture the nuanced beauty of language, on the other hand, is difficult. Deep learning methods, on the other hand, excel at capturing text characteristics by automatically selecting and combining features. Traditional statistical and rule-based methods rely on manually created sentence features, which frequently fail to capture the essence of well-constructed sentences, particularly those containing advanced grammar or stylistic devices such as metaphors and personification. In comparison, neural network models can learn semantic vectors from large amounts of data on their own, effectively representing sentence features in binary classification tasks.

The Convolutional Neural Network (CNN) is a widely used type of artificial neural network. It employs convolutional kernels to capture local information, which is then synthesized into global information via the pooling layer. The core architecture of a CNN includes an input layer, convolutional layers, and pooling

layers. In automatic essay scoring tasks, the input layer typically consists of a text representation matrix formed by word vectors. The convolutional layer allows for the setting of kernels of various sizes, enabling the capture of certain contextual and sequential information. One of the key advantages of CNNs over traditional neural network models is the introduction of weight sharing, which simplifies the network's complexity and accelerates training. In the pooling layer, a segment-wise maximum pooling approach is used to preserve the relative location of multiple local maximum values. This method is also capable of detecting the intensity of features if strong characteristics are repeated. However, it's important to note that while this approach retains coarse position information, absolute position details are lost. After the convolution and pooling processes, a representation of the sentence level is achieved.

One of the most significant benefits is the system's ability to provide immediate, personalized feedback to students on their writing. This instant feedback loop can significantly enhance the learning process, allowing students to identify and correct errors, refine their writing style, and better understand the criteria for high-quality writing. Immediate feedback is particularly valuable in large classrooms or distance learning scenarios where individualized attention from instructors may be limited. With the system taking on the task of assessing basic grammar, spelling, and syntax, educators can devote more time and resources to teaching higher-level writing skills. These include argumentation, critical thinking, and creative expression. Teachers can focus on developing students' abilities to construct well-organized, coherent, and persuasive texts, rather than spending excessive time marking mechanical errors.

Proposed BiLSTM -CNN model. Creating a model that combines Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural Networks (CNN) entails creating a system that takes advantage of the advantages of both architectures. BiLSTM excels at understanding context and dependencies in sequential data, such as text, whereas CNNs excel at extracting features, in this case, sentence structures, from data. A step-by-step procedure for creating such a model:

Convolutional Layer. Firstly, Identify features in the word vector matrix. Then it generates three different types of convolution kernels (e.g., 3x128, 4x128, and 5x128), each with 50 kernels. These kernels will aid in the extraction of various local features from word vectors. Finally, use these kernels to extract features from the word vector matrix. To introduce nonlinearity and accelerate convergence, use a ReLU activation function for each neuron.

Pooling Layer. To distill the features extracted by the convolution layer and to reduce the dimensionality of the feature space. Divide each feature map into chunks (e.g., three parts), and apply max pooling to each chunk. This approach helps preserve the relative order information and capture the strongest features.

Bi-LSTM Layer. After the pooling layer, the output is fed into a BiLSTM layer. It analyzes the sequence data (features extracted and pooled from the CNN) in both forward and backward directions. This is crucial for understanding the context and dependencies in the text data. The BiLSTM processes the sequence of features, capturing information from both past and future contexts.

Fully Connected and Output Layers. The output is Flatten from the Bi-LSTM layer to create a one-dimensional vector. Add two dense layers to allow the model to learn non-linear combinations of features in fully connected layer. The sigmoid activation function used in the output layer for binary classification tasks (like sentiment analysis) or softmax for multi-class classification.

Model Training and Optimization. An appropriate loss function (like cross-entropy) is chosen and an optimizer stochastic gradient descent is used. Model is trained using backpropagation and adjust the weights iteratively. finally, implement dropout or L2 regularization to prevent overfitting.

The graph 2 representing the CNN model is the shortest, suggesting that it has the lowest accuracy among the three models presented. The accuracy percentage is approximately 85%, which indicates that while the CNN model is relatively accurate, there may be room for improvement in feature analysis tasks. The CNN-LSTM model, is significantly greater than the first, implying a noticeable increase in accuracy. The model's accuracy is around 90%, showing that combining CNN features with LSTM, which can capture sequential information, offers a substantial improvement over the plain CNN model. The Proposed CNN-BiLSTM model is the highest of all, indicating the highest accuracy among the three models on feature analysis. The accuracy is just under 91.6. The bidirectional LSTM allows the model to access information from both past and future states, potentially giving it an advantage in understanding the context within data, leading to higher accuracy

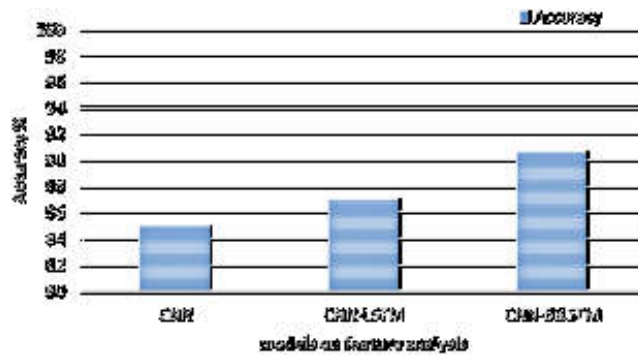


Fig. 3.2: The accuracy of feature analysis

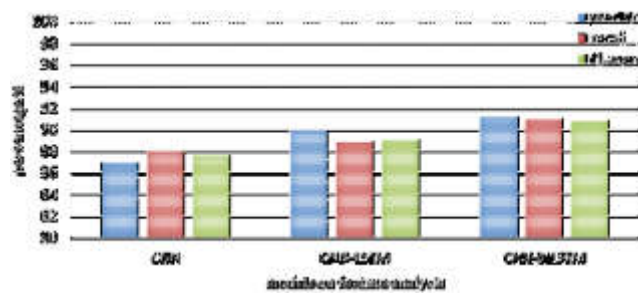


Fig. 3.3: The performance comparison of feature analysis model

Convolutional neural networks (CNNs) have a strong capability for extracting intricate features from sentences. To investigate this, experiments were conducted with three distinct approaches to feature extraction: manual feature engineering, CNN-based feature extraction, and a hybrid method combining manual and CNN. When examining the trends in accuracy and recall as illustrated in Figure 3.3, the performance metrics for all three methods were broadly comparable. However, it's notable that manual feature engineering CNN alone resulted in the lowest accuracy, specifically at 87%. Conversely, the CNN-based approach yielded the lowest recall rates in feature classification, and consistently, both accuracy and recall were lower for CNN and CNN-LSTM. This suggests an inherent challenge within the algorithm's ability to discern sentences, and highlights a discrepancy with human perception in real-world applications.

The data indicates that BiLSTM and CNN-based feature extraction methods significantly enhance the differentiation between the sentences. Additionally, it's observed that the performance metrics for sentences are consistently lower across all categories, underscoring the complexity of assessing the aesthetic quality of sentences—a factor that is also reflective of an individual's writing skill level.

In pursuit of refining the experiment, consideration was given not only to the blend of feature extraction techniques from machine learning but also to the analysis of different network classifications based on various feature combinations. This included looking at linguistic and semantic feature integration, as well as the incorporation of difficult features. The results, as evident in Figure 3.3, features outperformed using text-CNN and BiLSTM models. When comparing models with equivalent feature types, the LSTM model surpassed the text-CNN in performance, demonstrating its superior capability for memory learning in text-mining applications. Finally, an enhanced version of LSTM, known as Bi-LSTM, achieved the best results in the second set of experiments. This improvement is attributed to Bi-LSTM's adeptness in capturing temporal dependencies from different directions, thereby obtaining more temporally relevant sentence features.

Continuous evaluation and benchmarking against industry standards and datasets ensure that the systems performance meets the expected criteria for accuracy, fairness, and reliability. These evaluations guide further

refinements and adjustments to the system.

4. Conclusion. The research highlights an approach to automating the assessment of English writing proficiency using cutting-edge machine learning algorithms. Addressing the inefficiencies and biases inherent in traditional evaluation methods, this research outlines the development of an intelligent system that employs Natural Language Processing and neural network models to deliver swift, consistent, and objective analysis of written English. The system's architecture, which harnesses a synergy of algorithms to evaluate various aspects of writing, is rigorously tested against a vast corpus of English samples. The findings are clear: the proposed machine learning-based system not only rivals but also potentially surpasses human raters in terms of accuracy and speed, marking a significant advancement in the field of language assessment. This breakthrough holds considerable promise for educational and professional domains, offering a scalable, reliable alternative that could fundamentally transform the assessment landscape of English writing ability.

REFERENCES

- [1] S. BONTHU, S. RAMA SREE, AND M. KRISHNA PRASAD, *Automated short answer grading using deep learning: A survey*, in Machine Learning and Knowledge Extraction: 5th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2021, Virtual Event, August 17–20, 2021, Proceedings 5, Springer, 2021, pp. 61–78.
- [2] R. G. GARROPO, M. G. SCUTELLÀ, AND F. D'ANDREAGIOVANNI, *Robust green wireless local area networks: A matheuristic approach*, Journal of Network and Computer Applications, 163 (2020), p. 102657.
- [3] Y. GUO, *A study of english informative teaching strategies based on deep learning*, Journal of Mathematics, 2021 (2021), pp. 1–8.
- [4] V. KUMAR AND D. BOULANGER, *Explainable automated essay scoring: Deep learning really has pedagogical value*, in Frontiers in education, vol. 5, Frontiers Media SA, 2020, p. 572367.
- [5] V. S. KUMAR AND D. BOULANGER, *Automated essay scoring and the deep learning black box: How are rubric scores determined?*, International Journal of Artificial Intelligence in Education, 31 (2021), pp. 538–584.
- [6] K. KYRIAKOPOULOS, K. M. KNILL, AND M. J. GALES, *A deep learning approach to assessing non-native pronunciation of english using phone distances*, ISCA, 2018.
- [7] Y. LI, *Deep learning-based correlation analysis between the evaluation score of english teaching quality and the knowledge points*, Computational Intelligence and Neuroscience, 2022 (2022).
- [8] Y. LIU AND R. LI, *Deep learning scoring model in the evaluation of oral english teaching*, Computational Intelligence and Neuroscience, 2022 (2022).
- [9] C. LU AND M. CUTUMISU, *Integrating deep learning into an automated feedback generation system for automated essay scoring.*, International Educational Data Mining Society, (2021).
- [10] X. LU AND R. HU, *Sense-aware lexical sophistication indices and their relationship to second language writing quality*, Behavior research methods, 54 (2022), pp. 1444–1460.
- [11] O. LYASHEVSKAYA, I. PANTELEEVA, AND O. VINOGRADOVA, *Automated assessment of learner text complexity*, Assessing writing, 49 (2021), p. 100529.
- [12] H. MEISHERI, R. SAHA, P. SINHA, AND L. DEY, *Textmining at emoint-2017: A deep learning approach to sentiment intensity scoring of english tweets*, in Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2017, pp. 193–199.
- [13] F. QIN, *College english intelligent writing score system based on big data analysis and deep learning algorithm*, Journal of Database Management (JDM), 33 (2022), pp. 1–26.
- [14] D. RAMESH AND S. K. SANAMPUDI, *An automated essay scoring systems: a systematic literature review*, Artificial Intelligence Review, 55 (2022), pp. 2495–2527.
- [15] R. RIDLEY, L. HE, X. DAI, S. HUANG, AND J. CHEN, *Prompt agnostic essay scorer: a domain generalization approach to cross-prompt automated essay scoring*, arXiv preprint arXiv:2008.01441, (2020).
- [16] J. SAWATZKI, T. SCHLIPPE, AND M. BENNER-WICKNER, *Deep learning techniques for automatic short answer grading: Predicting scores for english and german answers*, in International Conference on Artificial Intelligence in Education Technology, Springer, 2021, pp. 65–75.
- [17] S. THARA AND P. POORNACHANDRAN, *Social media text analytics of malayalam–english code-mixed using deep learning*, Journal of big Data, 9 (2022), p. 45.
- [18] M. UTO, *A review of deep-neural automated essay scoring models*, Behaviormetrika, 48 (2021), pp. 459–484.
- [19] Z. WANG, H. HUANG, L. CUI, J. CHEN, J. AN, H. DUAN, H. GE, N. DENG, ET AL., *Using natural language processing techniques to provide personalized educational materials for chronic disease patients in china: development and assessment of a knowledge-based health recommender system*, JMIR medical informatics, 8 (2020), p. e17642.
- [20] T. XIA AND X. CHEN, *A weighted feature enhanced hidden markov model for spam sms filtering*, Neurocomputing, 444 (2021), pp. 48–58.
- [21] S. YUAN, T. HE, H. HUANG, R. HOU, AND M. WANG, *Automated chinese essay scoring based on deep learning*, CMC-Computers Materials & Continua, 65 (2020), pp. 817–833.
- [22] Z. YUAN, *Interactive intelligent teaching and automatic composition scoring system based on linear regression machine learning algorithm.*(retraction of vol 40, pg 2069, 2020), 2021.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Jan 6, 2024

Accepted: Feb 9, 2024



RESEARCH ON CRYPTOGRAPHY-BASED DATA SECURITY AND TRUSTWORTHINESS IN DIGITAL CONSTRUCTION OF WATER RESOURCES AND HYDROPOWER

CHAO YUE*, WEI LIU†, LICHENG CHEN‡ AND CHONG ZUO§

Abstract. This study aims to strengthen data security and establish credibility using novel cryptography-based techniques in the context of the digital revolution in the water resource and hydropower development industry. Protecting sensitive data and guaranteeing the confidentiality of digital assets becomes crucial as the sector depends more and more on digital technology for communication, monitoring, and project management. The goal of this work is to create developed cryptographic protocols and structures that have been tailored to the needs of the water resources and hydropower industry. This study offers a thorough investigation into the use of cryptographic methods to tackle the difficulties presented by the digital construction surroundings in these projects. The research process combines algorithm creation, theoretical advancements, and real-world application. The efficiency and viability of the suggested cryptographic approaches in resolving trust and security issues intrinsic in digital building environments will be evaluated using actual-life scenarios and simulations. The results of this study should offer a strong basis for improving data security, reliability, and integrity in digital construction projects related to water resources and hydropower. Through the development of cryptography techniques specifically suited to this vital infrastructure industry, the research helps to build digital ecosystems that are resilient and secure, which is important for the sustainable growth of hydropower and water resources.

Key words: Cryptography, Data Security, Trustworthiness, Digital Construction, Water Resources and Hydropower

1. Introduction. Energy is crucial to the economy’s ability to grow sustainably [10]. Although nuclear energy and fossil fuels are frequently used to generate electricity, they frequently cause some environmental harm due to the emissions of CO₂ as well as other radioactive substances. Sustainable and alternative sources of energy have been heavily pushed in such a situation. One option for efficiently preserving the natural world is hydropower. It is seen as a component of a low-carbon economic system’s energy combine, particularly for nations that are developing [20]. Hydropower, the world’s most productive renewable energy resource for electricity generation, produces 71% more electrical power than other energy sources including coal, gas, and oil [13].

There are certain benefits to using hydropower to generate electricity. It is less expensive, more sustainable, and more dependable than coal, gas, or oil. There are less environmental restrictions on hydropower than on solar and wind power [23]. Hydropower now plays a larger role than it did previously since it is acknowledged as being fundamental to the production of energy [16]. Since concealed or in-conduit hydropower systems are completely incorporated into the current infrastructure, they have less of an environmental impact than conventional hydropower plants that operate in rivers [6]. Particularly, these hydropower systems capture the extra energy of water that is being utilized for purposes other than electricity production.

A geodatabase of unexplored prospective places for energy recovery at current hydro facilities in particular European towns and nations has been produced by an ongoing study [24, 5]. Building an in-conduit hydropower system could be advantageous for governmental treatment works, including wastewater amenities, and public water systems, considering their respective consequences. The author [26, 27] provided a thorough summary

*Hydropower and Water Conservancy Engineering Institute POWERCHINA HUADONG Engineering Corporation Limited, Hangzhou, Zhejiang, 311122, China (chaoyuedigital12@outlook.com)

†Hydropower and Water Conservancy Engineering Institute POWERCHINA HUADONG Engineering Corporation Limited, Hangzhou, Zhejiang, 311122, China

‡Dagu Hydropower Branch of Huadian Xizang Energy Co., Ltd., Shannan, Xizang, 856000, China

§Hydropower and Water Conservancy Engineering Institute POWERCHINA HUADONG Engineering Corporation Limited, Hangzhou, Zhejiang, 311122, China

of the advancements in in-conduit hydropower technology and their uses. In [22], actual case studies of small hydro turbines incorporated into drinking water and wastewater networks were provided together with a succinct technical explanation. Numerous nations have evaluated the potential for hydropower, including the feasibility of putting turbines in water and wastewater infrastructure from a technical and financial standpoint.

Turbine installation is typically ideal near wastewater treatment outlets because of the steady and enough water flow. The WWTP process involves constant monitoring of the variables needed to choose hydro turbines, like head and flow. As such, monitoring the turbine's functioning can be rather simple [4]. On the other hand, low- or ultralow-head plants may face a problem if the tailwater effect is overlooked. The head is reduced at most sites during a flood period because the tailwater level at the outfall rises greater than the level upstream of the intake, depending on the receiving water body (such as a river). Nevertheless, the literature hardly ever discusses these situations.

The motivation for this research emerges from the critical need to enhance data security and establish a foundation of trust within the rapidly digitising landscape of the water resources and hydropower development industry. As this sector increasingly relies on digital technologies for essential operations such as communication, monitoring, and project management, the protection of sensitive data and the confidentiality of digital assets become paramount. The advent of the digital revolution in this field presents immense opportunities and significant challenges, particularly regarding safeguarding against cyber threats and ensuring the integrity of digital construction environments.

This study is driven by the recognition that conventional cryptographic protocols and security measures may not fully address the unique complexities and requirements of the water resources and hydropower industry. These projects are characterized by their extensive scale, long duration, and the critical nature of their infrastructure, which necessitates a bespoke approach to data security. The research aims to develop and refine cryptographic techniques specifically tailored to meet these industry-specific needs, providing robust protection for digital assets and sensitive information.

The main contribution of the proposed method is given below:

1. Creation of customized cryptography protocols intended to handle the trust and security issues that arise during the digital construction lifespan of hydropower and water resource projects.
2. The security and reliability of vital project information are guaranteed by these protocols, which offer a framework for protecting sensitive data throughout transfer, storage spaces. and retrieval.
3. Scaling and efficiency improvements for cryptographic solutions while considering the special requirements of large-scale water resource and hydropower projects involving a variety of stakeholders.
4. By balancing strong security measures with effective data processing, the research helps build cryptographic algorithms that are useful in real-world construction circumstances.

The rest of our research article is written as follows: Section 2 discusses the related work on various classification of brain image processing and methodData Security and Trustworthiness in Digital Construction of Water Resources and Hydropower s. Section 3 shows the algorithm process and general working methodology of proposed work. Section 4 evaluates the implementation and results of the proposed method. Section 5 concludes the work and discusses the result evaluation.

2. Related Works. Cyberattacks are online activities that try to break into the computer networks of people or organizations with the intention of causing damage or interrupting operations. These assaults may target various objectives, such as stealing sensitive data or jeopardizing data integrity [2]. For energy and electricity systems to operate securely and dependably, sufficient protection layers must be developed for a CPS. Nonetheless, the electricity sector has seen a rise in cyberattack efforts in recent years. Approximately 800 cyberattacks have been reported in the energy sector since the 1980s [3].

A thorough review of turbines suitable for concealed hydro and in-conduit hydropower was provided in [21, 17], with a focus on current developments in the field of turbine technology. Novel technological approaches have been put forth that enhance traditional turbines with stronger designs, increased efficiency, and potentially cheaper costs [14]. However, while more recent or developing technologies present creative methods for in-conduit hydro generation, they may not necessarily be the most economical option [19]. The comparison of equipment costs is complex because of the different sites and turbines. However, modular structures may have higher hydromechanical and electric running costs than traditional turbines, even though their building and

installation expenses may be lower [12, 11].

Low-cost engines, such as pumps as turbines (PaTs), are recommended because traditional hydro turbine technologies aren't always competitive in the market. These are regular pumps that have had their flow direction reversed so that they can be used as turbines. PaTs have been the subject of research for almost a century, and their use in small- and micro-hydropower is still significant today [7, 8, 25]. PaTs are typically employed at locations with greater head counts; the literature hardly ever discusses low-head application experience.

Most importantly, tools that assist water and wastewater providers in determining whether establishing hydropower facilities is both technically and financially feasible should be developed [1, 28]. Evaluation instruments must be as simple to use and economical as feasible because the majority of in-conduit or hidden hydro systems have comparatively limited capacities and, as a result, require a highly expensive feasibility study. For small developers, these needs are not met by the tools that are now available [15]. It has been suggested that conduit projects be evaluated using a few engineering design tools. However, these have not yet been used in more comprehensive analyses. To be sure, the US-developed tools are partly to blame for some exceptions [18, 9]. These are free-to-use tools that work with widely accessible spreadsheet software.

Given the extensive geographical spread of water resources and hydropower infrastructure, cryptographic protocols must be scalable across large distributed networks. This might involve optimizing encryption algorithms for low-latency operations and ensuring they can handle the high volume of data generated by IoT devices and sensors without compromising performance. Many operations within the sector rely on real-time data for monitoring and control. Cryptographic protocols can be modified to support efficient real-time encryption and decryption processes, enabling secure yet timely data transmission crucial for operational decision-making. Adapting cryptographic protocols to be compatible with existing industrial control systems (ICS) and operational technology (OT) used in the sector. This may require developing lightweight cryptographic solutions that can be implemented on legacy systems without significant hardware upgrades. With the increase in remote monitoring and management of hydropower plants, cryptographic protocols need to ensure secure remote access. This could involve adapting protocols to provide robust authentication and secure communication channels for remote users, preventing unauthorized access and ensuring data integrity.

3. Proposed Methodology. The proposed method uses Cryptographic techniques for Data Security and Trustworthiness in the Digital Construction of Water Resources and Hydropower. Create protocols for encryption with data transfer, storage, and access control specifically suited to the digital building lifecycle. Create protocols that handle issues including permission procedures, secure interaction with stakeholders, and data tampering prevention. The purpose of this proposed technique is to improve data security and reliability in the digital design of hydropower and water resource projects by exploring and carrying out cryptography-based technologies in an organised and rigorous manner. In figure 3.1 shows the architecture of the proposed method.

The water resource and hydropower industry faces unique cryptographic needs and challenges stemming from its critical infrastructure status, the complexity of its operational environments, and the increasing digitization of its processes. Addressing these challenges is crucial for ensuring the security, reliability, and resilience of these essential services. Here are some of the specific cryptographic needs and challenges in this industry: 1. The industry relies heavily on real-time data for monitoring water levels, flow rates, and power generation metrics. Cryptographic solutions must provide real-time encryption and decryption of data streams without introducing significant latency, which could impact operational efficiency and safety. 2. Water resource and hydropower systems often encompass extensive geographical areas with multiple sites and installations connected via distributed networks. Cryptographic protocols must be scalable and flexible enough to secure communications across these vast and varied landscapes. 3. Given its importance to national security and the economy, the industry is a potential target for state-sponsored and sophisticated cyber-attacks, including Advanced Persistent Threats (APTs). Cryptographic measures must be robust enough to protect against such threats, ensuring the integrity and availability of control systems.

3.1. Homomorphic Encryption Techniques. These allow calculations to be performed on encrypted data without requiring decryption. Homomorphic encryption might be useful for secure computing on sensitive data without exposing it during processing in water resources and hydropower.

HE is a type of encryption that enables computation between plaintexts that are concealed in ciphertext. Another cipher-text with the correct plaintext-to-plaintext calculation output can be the result of the

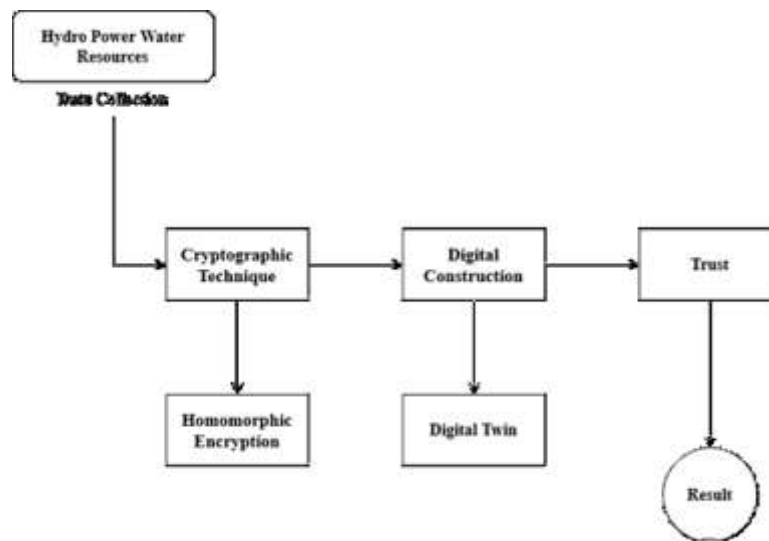


Fig. 3.1: Architecture of Proposed method

computation in HE. Since the ciphertext completely encloses the plaintext and data encryption only permits the decryption and encryption procedures, the hidden plaintext within the ciphertext cannot be altered using conventional encryption techniques. Therefore, to obtain the plaintexts, which can be utilized to perform operations on the original message contained in the cipher-texts, the cipher-texts must be decoded.

By enabling secure computations on encrypted data, HE reduces the need for complex data protection measures that might otherwise slow down processing or increase operational costs. This can lead to more efficient system operations and reduced overhead for security. As cyber threats evolve, the ability to compute on encrypted data provides a forward-looking approach to data security, ensuring that the sector is prepared for emerging challenges and can safeguard sensitive information against future vulnerabilities.

3.2. Digital Twin-based Digital Construction of Water Resources and Hydropower. The preparation, design, building, and administration of infrastructure connected to water resources and hydropower generation are improved using cutting-edge digital technologies and data-driven approaches in digital construction projects. With the creation and management of water-related projects, this innovative strategy makes use of digital tools to streamline procedures, boost productivity, and guarantee sustainability.

A virtual copy of a real object, system, or procedure is called a digital twin. Digital twins can simulate a project's whole lifecycle in the context of hydropower and water resources, offering real-time insights and assisting in improved decision-making. The use of digital twins makes it possible to simulate, analyse, and monitor hydropower facilities, dams, and water systems. They support the long-term resilience of infrastructure, performance optimization, and maintenance demand prediction.

Making a thorough 3D model or depiction of the real object or system is the first step in creating a digital twin. The digital twin is built on top of this model. When it comes to water resources and hydropower, the physical assets—like dams, water treatment facilities, or hydropower plants—as well as their components, dimensions, and functional features are digitally modelled.

3.3. Trustworthiness for water resources and Hydropower. The reliability, integrity, and security of the systems, procedures, and data involved in controlling and producing electricity from water resources are referred to as trustworthy in the context of hydropower plants and water resources. Establishing credibility is essential to guarantee the security, longevity, and effective functioning of water-related infrastructure.

Guaranteeing the precision and dependability of information gathered from sensors, surveillance tools, and additional sources in water infrastructure. To make well-informed decisions about hydropower generation, dam safety, and water flow, one must have faith in the accuracy of data. To keep data accurate, regular validation and

Algorithm 11 Homomorphic Encryption

```

1: Input: Public key, KW
2: Output: verifying the result
3: initialize keywords KW into T0;
4: select key  $K_{se}$  for  $P_{R_f}$  //  $K_{se}$  is Key search
5: select  $K_x, K_i, K_z$  for  $P_{R_f} F_p$ 
6:  $KeF(K_{se}, W)$ 
7: for  $i \in DB(W)$  do
8:   counter C1
9:   evaluate  $X_{i_d} \leftarrow F_p(K_i, i_d), Z \leftarrow F_p(K_z, w||C)$ ;


$$Y \leftarrow X_{i_dz} - 1e \leftarrow E_{n_c}(K_e, i_d);$$


$$X_{tag} \leftarrow gF_p(K_x, w) X_{i_d} \text{ and } X_{set} \leftarrow X_{set} U X_{tag};$$

10:   append (y,e) to t and  $C \leftarrow C + 1$ ;
11:    $T[w] \leftarrow t$ ;
12: end for
13: return  $E_{DB}, K = (K_{se}, K_x, K_i, K_z, K_t)$ ;
14: generating a token ( $q(w), K$ );
15: evaluate  $stag \leftarrow T_{set}.GetTag(K_t, w_1)$ ;
16: The server receives data from the user.
17: for  $C = 1, 2, \dots$  Until the server halts do
18:   for  $i = 2, \dots, n$  do
19:      $x_{token|C|} \leftarrow gF_p(K_z, w_1||C) F_p(K_x, w_i)$ ;
20:   end for
21:    $x_{token|C|} \leftarrow (x_{token|C|,2}, \dots, x_{token|C|,n})$ ;
22: end for
23:  $Tokq \leftarrow (stag, x_{token})$ ;
24: return T okq;
25: end

```

verification procedures are necessary. Building and upholding technology that is resilient to calamities, severe weather, and other possible disruptions. The incorporation of resilience into water infrastructure guarantees its capacity to operate in challenging circumstances, mitigating the likelihood of malfunctions, and enhancing its enduring reliability.

Keeping lines of communication open and honest with all parties involved, such as the public, neighbours, and regulatory bodies. By giving accurate information about the workings, safety precautions, and possible effects of water infrastructure projects, open communication promotes healthy relationships with neighbours and fosters trust.

4. Result Analysis. Numerous writers have created geodatabases and utilized geographic information (GIS data) to find possible hydro sites in water distribution systems. Spatial databases, or high-resolution digital elevation or terrain models (DEMs), are available in many nations. Along with the Shuttle Radar Topography Mission (SRTM) DEMs from the United States Geological Survey, global terrain data from Google Earth or other platforms can also be used, but they should be used cautiously—that is, only for the initial assessment of SHP locations and not for flat terrains or low-lying countries and areas with a low vertical resolution in geography.

The proposed method uses parameter metrics such as accuracy, precision, recall and f1-score for hydropower water resources.

Accuracy is a crucial criterion that relates to the precision and correctness of numerous processes, data, and outcomes in the context of digital creation of hydropower and water resources. the accuracy of spatial data utilized in the planning and design stages, such as maps, surveys, and geographical information systems (GIS). Precise geographic information guarantees that the project site's physical attributes are accurately depicted,

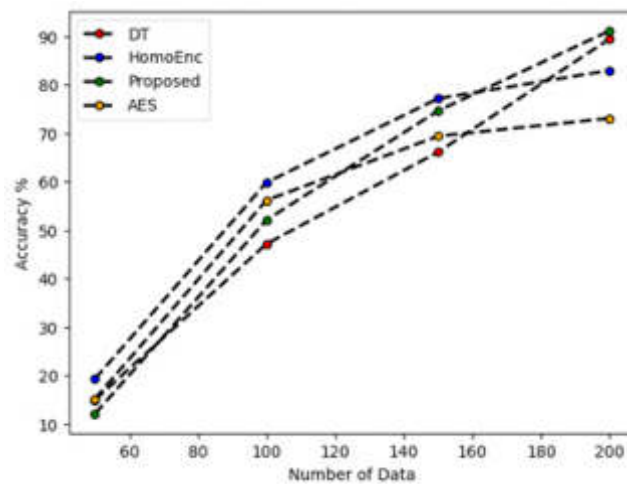


Fig. 4.1: Accuracy

reducing mistakes in both design and construction.

The accuracy of quality assurance procedures and examinations conducted both during and following construction. Precise quality control and inspections guarantee that constructed pieces fulfil the intended quality standards and help to ensure compliance with industry standards, legal requirements, and project specifications. the accuracy of the information kept in asset management systems, which are used to keep an eye on and repair water infrastructure. For efficient maintenance scheduling, preparation, and management throughout their lifecycle to ensure the durability and dependability of water-related resources, reliable asset data is crucial. In figure 4.1 shows the accuracy of proposed method.

When discussing digital construction for hydropower and water resource projects, precision pertains to the precision and dependability of the algorithm or system in recognizing and detecting elements or features within digital data. When it comes to activities such as object detection, where the objective is to reduce false positives and make sure that features recognized are relevant to the construction process, precision is an important parameter.

The ratio of true positives to the total of true positives and false positives is used to compute precision. This refers to precisely recognizing and finding pertinent objects or elements inside the digital model of the infrastructure or building site in the context of digital construction. To reduce false positives, which might influence the construction process and result in wrong judgments, high precision is necessary. It guarantees the reliability and applicability of the features found. In figure 4.2 shows the evaluation of Precision.

"Recall" generally refers to a measurement of performance used to assess the efficacy of methods or systems in the context of digital construction of hydropower and water resource projects, particularly in activities involving object detection or recognition. Recall, which is sometimes referred to as sensitivity or true positive rate, quantifies a system's capacity to accurately identify every pertinent case among all actual occurrences. In digital construction, recall evaluation is an element of an iterative process. Based on the feedback from memory evaluations, the system can be modified and fine-tuned to increase its capacity to recognize and recall pertinent aspects or abnormalities in the building procedure. In figure 4.3 shows the evaluation of Recall.

A metric called the F1-score, sometimes referred to as the F1 measure or F1-value, combines recall and precision into a single number. It is especially helpful in situations when there is an unequal distribution of classes, and it is important to consider both false positives and false negatives. The F1-score can be utilized in the digital design of hydropower and water resource projects to assess how well models or algorithms perform in tasks like object detection, image categorization, or predictive maintenance.

Evaluating the precision of algorithms used to detect things in photos, such as tracking infrastructure elements or spotting anomalies. Assessing the effectiveness of models that forecast equipment breakdowns

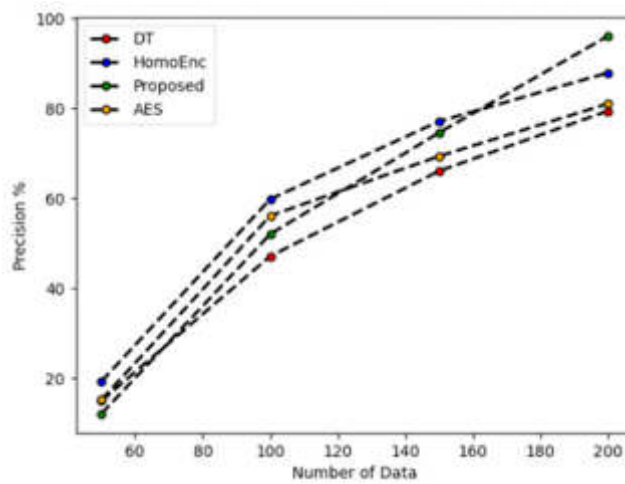


Fig. 4.2: Precision

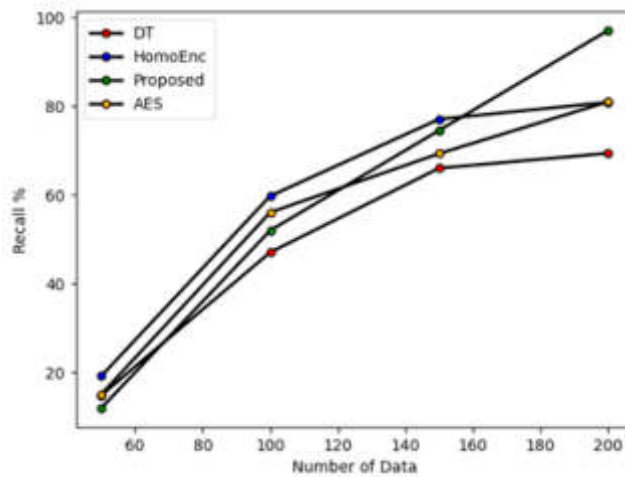


Fig. 4.3: Recall

or the need for maintenance to guarantee the dependability of hydropower facilities. Evaluating how well models categorize photos of building sites or the state of infrastructure connected to water. It is imperative to consider the goals of the work and the relative significance of recall and precision considering the application requirements when interpreting the F1-score. A balance between recall and precision may be more acceptable in certain situations, while a greater emphasis on precision may be preferred in others. In figure 4.4 shows the evaluation of F1-score.

5. Conclusion. The objective of this research is to enhance data security and establish credibility in the context of the digital revolution in the water resource and hydropower development industry by utilizing innovative cryptography-based techniques. As the industry grows more and more reliant on digital technology for project management, monitoring, and communication, safeguarding sensitive data and ensuring the confidentiality of digital assets becomes essential. This initiative aims to design cryptographic structures and protocols specifically suited to the demands of the hydropower and water resources sectors. This paper provides a com-

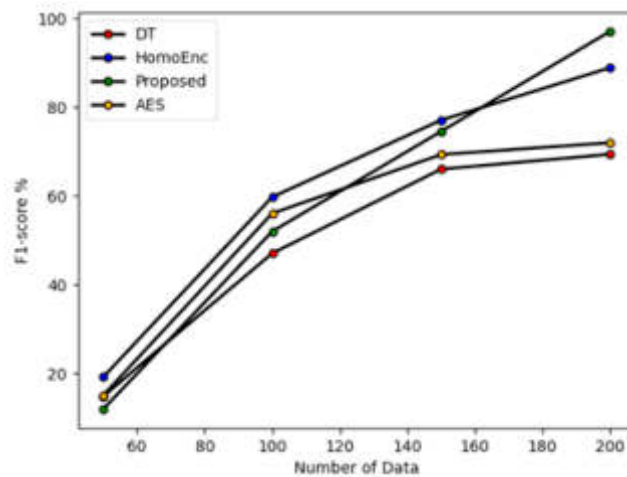


Fig. 4.4: F1-score

prehensive analysis of the application of cryptographic techniques to address the challenges posed by the digital building environment in these projects. The development of algorithms, theoretical breakthroughs, and practical implementation are all combined in the research process. Using real-world examples and simulations, the effectiveness and practicality of the proposed cryptographic techniques in addressing trust and security concerns inherent in digital building environments will be assessed. The findings of this research should provide a solid foundation for enhancing data security, dependability, and integrity in digital construction projects involving hydropower and water resources. By creating cryptographic methods tailored to this essential infrastructure sector, the research contributes to the construction of robust and secure digital ecosystems, which is necessary for the long-term development of hydropower and water resources.

Acknowledgement. This work was sponsored in part by National Natural Science Foundation of China (2345678)

REFERENCES

- [1] M. ABDELMALAK, V. VENKATARAMANAN, AND R. MACWAN, *A survey of cyber-physical power system modeling methods for future energy systems*, IEEE Access, (2022).
- [2] E. AHMADIAN, C. BINGHAM, A. ELNOKALY, B. SODAGAR, AND I. VERHAERT, *Impact of climate change and technological innovation on the energy performance and built form of future cities*, Energies, 15 (2022), p. 8592.
- [3] E. AHMADIAN, H. BYRD, B. SODAGAR, S. MATTHEWMAN, C. KENNEY, AND G. MILLS, *Energy and the form of cities: the counterintuitive impact of disruptive technologies*, Architectural science review, 62 (2019), pp. 145–151.
- [4] E. AHMADIAN, B. SODAGAR, C. BINGHAM, A. ELNOKALY, AND G. MILLS, *Effect of urban built form and density on building energy performance in temperate climates*, Energy and Buildings, 236 (2021), p. 110762.
- [5] W. ASCHER, *Rescuing responsible hydropower projects*, Energy Policy, 150 (2021), p. 112092.
- [6] N. DIAZ-ELSAIED, N. REZAEI, A. NDIAYE, AND Q. ZHANG, *Trends in the environmental and economic sustainability of wastewater-based resource recovery: A review*, Journal of Cleaner Production, 265 (2020), p. 121598.
- [7] D. DU, M. ZHU, X. LI, M. FEI, S. BU, L. WU, AND K. LI, *A review on cybersecurity analysis, attack detection, and attack defense methods in cyber-physical power systems*, Journal of Modern Power Systems and Clean Energy, (2022).
- [8] W. DUO, M. ZHOU, AND A. ABUSORRAH, *A survey of cyber attacks on cyber physical systems: Recent advances and challenges*, IEEE/CAA Journal of Automatica Sinica, 9 (2022), pp. 784–800.
- [9] A. FAUSTO, G. B. GAGGERO, F. PATRONE, P. GIRDINIO, AND M. MARCHESI, *Toward the integration of cyber and physical security monitoring systems for critical infrastructures*, Sensors, 21 (2021), p. 6970.
- [10] Y. Y. GHADI, D. B. TALPUR, T. MAZHAR, H. M. IRFAN, U. A. SALARIA, S. HANIF, T. SHAHZAD, AND H. HAMAM, *Enhancing smart grid cybersecurity: A comprehensive analysis of attacks, defenses, and innovative ai-blockchain solutions*, (2023).
- [11] S. HE, Y. ZHOU, X. LV, AND W. CHEN, *Detection method for tolerable false data injection attack based on deep learning framework*, in 2020 Chinese Automation Congress (CAC), IEEE, 2020, pp. 6717–6721.

- [12] S. KARAMDEL, X. LIANG, S. O. FARIED, AND M. MITOLO, *Optimization models in cyber-physical power systems: A review*, IEEE Access, (2022).
- [13] X. LEI, *Research on development and utilization of hydropower in myanmar*, Energy Reports, 8 (2022), pp. 16–21.
- [14] M. LEZZI, M. LAZOL, AND A. CORALLO, *Cybersecurity for industry 4.0 in the current literature: A reference framework*, Computers in Industry, 103 (2018), pp. 97–110.
- [15] J. LIU, W. ZHANG, T. MA, Z. TANG, Y. XIE, W. GUI, AND J. P. NIYOYITA, *Toward security monitoring of industrial cyber-physical systems via hierarchically distributed intrusion detection*, Expert Systems with Applications, 158 (2020), p. 113578.
- [16] R. LLÁCER-IGLESIAS, J. M. PÉREZ, J. R. SATORRE-AZNAZ, P. A. LÓPEZ-JIMÉNEZ, AND M. PÉREZ-SÁNCHEZ, *Energy recovery in wastewater treatment systems through hydraulic micro-machinery. case study*, Journal of Applied Research in Technology & Engineering, 1 (2020), pp. 15–21.
- [17] F. LONGO, A. PADOVANO, G. AIELLO, C. FUSTO, AND A. CERTA, *How 5g-based industrial iot is transforming human-centered smart factories: a quality of experience model for operator 4.0 applications*, IFAC-PapersOnLine, 54 (2021), pp. 255–262.
- [18] D. L. MARINO, C. S. WICKRAMASINGHE, B. TSOVALAS, C. RIEGER, AND M. MANIC, *Data-driven correlation of cyber and physical anomalies for holistic system health monitoring*, IEEE Access, 9 (2021), pp. 163138–163150.
- [19] E. M. NAVARRO, A. N. R. ÁLVAREZ, AND F. I. S. ANGUIANO, *A new telesurgery generation supported by 5g technology: benefits and future trends*, Procedia Computer Science, 200 (2022), pp. 31–38.
- [20] P. PUNYS AND L. JUREVIČIUS, *Assessment of hydropower potential in wastewater systems and application in a lowland country, lithuania*, Energies, 15 (2022), p. 5173.
- [21] D. A. PUSTOKHIN, I. V. PUSTOKHINA, P. RANI, V. KANSAL, M. ELHOSENY, G. P. JOSHI, AND K. SHANKAR, *Optimal deep learning approaches and healthcare big data analytics for mobile networks toward 5g*, Computers and Electrical Engineering, 95 (2021), p. 107376.
- [22] A. RAYMAKERS, C. SUE-CHUE-LAM, V. HALDANE, A. COOPER-REED, AND D. TOCCALINO, *Climate change, sustainability, and health services research*, Health Policy and Technology, 12 (2023), p. 100694.
- [23] L. F. RIBAS MONTEIRO, Y. R. RODRIGUES, AND A. ZAMBRONI DE SOUZA, *Cybersecurity in cyber-physical power systems*, Energies, 16 (2023), p. 4556.
- [24] M. M. M. SAW AND L. JI-QING, *Review on hydropower in myanmar*, Applied Water Science, 9 (2019), pp. 1–7.
- [25] S. SURYA, M. K. SRINIVASAN, AND S. WILLIAMSON, *Technological perspective of cyber secure smart inverters used in power distribution system: State of the art review*, Applied Sciences, 11 (2021), p. 8780.
- [26] S. TANG, J. CHEN, P. SUN, Y. LI, P. YU, AND E. CHEN, *Current and future hydropower development in southeast asia countries (malaysia, indonesia, thailand and myanmar)*, Energy Policy, 129 (2019), pp. 239–249.
- [27] Y. TIAN, F. ZHANG, Z. YUAN, Z. CHE, AND N. ZAFETTI, *Assessment power generation potential of small hydropower plants using gis software*, Energy Reports, 6 (2020), pp. 1393–1404.
- [28] J.-P. A. YAACOUB, O. SALMAN, H. N. NOURA, N. KAA NICHE, A. CHEHAB, AND M. MALLI, *Cyber-physical systems security: Limitations, issues and future trends*, Microprocessors and microsystems, 77 (2020), p. 103201.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Jan 6, 2024

Accepted: Feb 9, 2024



RESEARCH ON DEEP LEARNING-BASED ALGORITHM FOR DIGITAL IMAGE COMBINATION AND TARGET DETECTION

SHANLU HUANG* AND JIALIN LAI†

Abstract. This study uses deep learning techniques to improve target recognition and digital picture processing, combining efficiency and accuracy in the fields of computer vision and image processing. Different situations, heterogeneous circumstances in the environment, and a wide range of image properties present obstacles for the traditional approaches of combining images and target recognition. To address these issues, our research suggests a novel method that makes use of deep learning methods to identify relevant characteristics and trends from a variety of sources that provide diverse pictures. As part of the research process, a complex deep learning system that can recognize ordered representations of input photos is developed and trained. We will investigate whether faster RCNN are suitable for capturing temporal and spatial relationships in the image data. To maximize the model's performance, deep learning techniques will be used to make use of pre-trained networks on sizable datasets. Benchmark datasets will be used in the method's assessment, and it will be pitted with conventional image processing techniques. The accuracy and dependability of the algorithm's performance will be evaluated using quantitative metrics including precision, recall, and F1-score. Furthermore, qualitative evaluations will be conducted to determine the visual appeal and interpretive capacity of the created composite images.

Key words: deep learning, faster RCNN, digital image combination, target detection, satellite images

1. Introduction. The branch of computer science known as artificial intelligence (AI) studies how to make machines intelligent. In a perfect world, these devices would react similarly to humans in terms of perception, comprehension, and problem-solving decision-making [21, 5, 26]. Artificial Intelligence (AI) encompasses a broad range of fields, most of which are related to the senses that humans have, including computer vision (CV), the processing of natural languages (NLP), oversight, and robots. Through its ability to comprehend digital images and movies, computer vision is a branch of computer science that strives to emulate human vision [16, 2, 27, 15]. It analyses photos using a variety of optimization approaches and techniques. CV is a multidisciplinary field that includes automation, math, probability, artificial intelligence, and recognition of patterns. The branch of artificial intelligence called machine learning (ML) uses data to learn instead of explicit programming [8].

A more intricate and sophisticated model is needed to comprehend pictures and videos. Neural networks (NNs) are remarkably adept at processing vast volumes of data (such as photos and videos) and deciphering it, according to research findings. Scientists were able to resolve challenging issues such picture categorization, recognizing objects, recognition of objects, and segmentation of instances recognition of optical characters by utilizing neural networks in CV. Using deep learning methods, computer vision additionally plays a role in the analysis and detection of objects in images. By resolving the issues, CV has influenced several industries, including the analysis of documents, self-driving cars, medical imagery analysis, and satellite picture research. [14].

For several years, one of the main goals of computer vision research has been to identify objects. The primary objective of recognizing objects is to identify an instance in pictures and videos [28]. Using a bounding box, object identification in CV refers to identifying things of interest (such as people, pets, dogs, cycles, etc.) at a given spot in an image [1]. In the fields of artificial information, machine vision, and robotic seeing, object detection finds numerous uses, such as in augmented reality, security, and surveillance. Two types exist for

*School of Public Administration, Guangxi Technological College of Machinery and Electricity, Nanning 530007, China

†School of Public Administration, Guangxi Technological College of Machinery and Electricity, Nanning 530007, China (jialinlaidigita@outlook.com)

object detection. Finding general categories (person, cat, etc.) is the first form of detection; the second type targets particular examples, like the president's face.

Identifying objects in satellite images is a crucial, essential, and difficult task since objects are small, multi-oriented, and densely grouped. Thus, the main challenge is to identify and locate small objects in satellite images. Because the low-resolution image dataset shortens the training period, we have created a custom dataset with low-resolution images of objects (like small-sized aircraft) to achieve good accuracy with a minimal amount of computational power. Using a custom dataset, we have analysed the speed and accuracy of various object detection pipelines.

The main contribution of the proposed method is given below:

1. We assembled a dataset of satellite photos of airplanes and pre-processed it for training and testing purposes.
2. To speed up the target identification process, the suggested algorithm makes use of the Faster R-CNN architecture, which is well-known for its effectiveness in object detection tasks.
3. The model overcomes the computational performance constraints of standard methods by effectively localizing and classifying targets inside the images using region-based convolutional neural networks.
4. Using a bespoke dataset, the effectiveness of the main algorithms for the identification and categorization of aircraft in satellite imagery was compared in terms of execution speed and accuracy.

The remaining sections of this paper are structured as follows: Section 2 discusses the related research works, Section 3 describes the digital image combination and target detection, Section 4 presents the methods used to adopt the proposed model, Section 5 discusses the experimented results and Section 6 concludes the proposed system with future work.

2. Related Works. In the past few years, deep learning and machine learning approaches have been used to overcome many issues related to object identification in satellite imagery. There are three pipelines that offer real-time remedies: YOLO, SSD, and Faster-RCNN. The cutting-edge real-time object recognition framework YOLO (you only look once) uses a 416×416 resolution image and is based on a CNN (convolutional neural network) algorithm [23, 19]. The state-of-the-art framework Faster RCNN uses a 1000×600 resolution image and is based on the region suggestion method. The SSD (single shot detector) architecture operates on either 300×300 or 512×512 pixels per image, extracting feature maps across various layers and applying CNN filters to recognize an item.

The International Society for Photogrammetry and Remote Sensing (ISPRS) Vaihingen and dam benchmark datasets [6], which include high-resolution photos followed by CNN for fine-tuning and hitting state-of-the-art accuracy, were subjected to dense labelling by the author [9, 4, 29]. By using region-based approaches and classification algorithms, the researcher in [20] focused on remote sensing and localization and produced improved object localization outcomes. Nevertheless, the region-based method's significant latency prevented the huge area from being covered (40 s covered area of 1280×1280 pixels). Although it was shown to be slower for segmentation, the author's work [12] used separation and additional processing approaches and produced trustworthy findings regarding automated road detection in satellite data.

Sparse and collaborative representation, as well as kernel-based machine learning, have seen the effective application of machine learning in HTD. By expanding traditional statistical techniques, several kernel-based target detectors have been presented, such as kernel target-constrained interference-minimized filter (KTCIMF) [25], kernel adaptive subspace detector (KASD) [10], and kernel orthogonal sub-space projection (KOSP) [11]. However, a lot of presumptions are also extensively relied upon by these procedures. Regarding limited and cooperative depictions, since the author developed a sparsity-based target detector (STD) [17], several other useful works have been presented. These include the hybrid sparsity and statistics detector (HSSD) [24], the combination of sparse and collaborative representation (CSCR) [13], and the sparse representation-based binary hypothesis-based detector (SRBBHD) [19].

Only a few techniques have been developed for the relatively new use of deep learning to HTD. Guided detectors use synthesizing to primarily increase target data. They then build an end-to-end detector through extensive pixel-pair training. Among the well-liked techniques are two-stream convolutional network-based target detector (TSCNTD) [17], deep network-based HTD (referred to as HTD-Net) [18], and convolutional neural network target detector (CNNTD) [7]. Furthermore, to transfer information from a large-sample domain

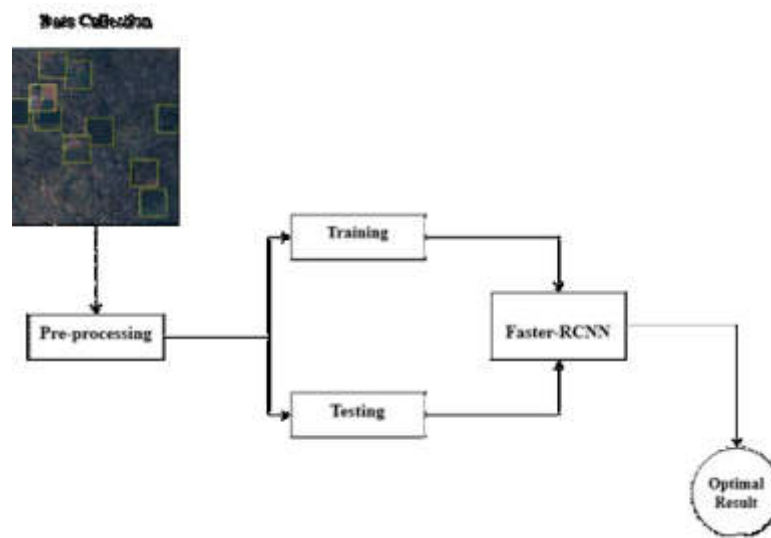


Fig. 3.1: Architecture Diagram of Proposed Method

of origin to a small-sample target domain, a domain adaptable learning model has been developed [22]. Under specific restrictions, unsupervised methods often improve the discriminating ability with unsupervised networks and then use a straightforward matching strategy to detect targets [3].

3. Proposed Methodology. In this work, Faster RCNN method is used for digital image combination and target detection. Initially the dataset is collected, in this work the aircraft satellite images are used as a dataset. Next the collected dataset is pre-processed and then the data is divided for training and testing. The training and testing is carried out by deep learning method Faster-RCNN. In figure 3.1 shows the architecture diagram of proposed method.

3.1. Dataset Collection. One of the most important phases in the entire object identification process is creating the dataset since the dataset has a significant impact on the model's accuracy and performance. It is the most crucial factor to consider while evaluating and examining the efficacy of different algorithms. The internet makes it possible to use larger images in a multitude of categories to accurately depict the intricacy and diversity of objects. The emergence of extensive datasets such as millions of photos has been crucial in facilitating exceptional object detection capabilities.

We obtained satellite images with a 1920×1080 -pixel resolution using Google Earth. Since they are typically classified, real-time satellite surveillance photographs are extremely difficult to find. For this reason, Google Earth is your best bet when looking for satellite photos of aircraft. Consequently, we tried to locate as many pictures of aircraft as we could. The dataset ought to be larger, but our options are limited. To cut down on training time, we separated the collected photos into 550×350 resolution after collecting. Next, we manually eliminated every picture that didn't include any items. In our dataset, there are 442 photos including 2213 aircraft objects. Next, we tagged photos using the labelling tool.

Pre-trained models are adept at extracting complex features from images, thanks to their exposure to diverse datasets. These features can range from basic textures and shapes to more intricate patterns, providing a rich set of characteristics for the system to use in target recognition tasks. Leveraging pre-trained models can drastically reduce the time and resources needed to train deep learning systems from scratch. Since these models have already learned a broad set of features, they require less data and fewer iterations to adapt to the specific nuances of a new task.

3.2. Pre-processing. An essential first step in getting the data ready for further examination or use is pre-processing satellite photos. To improve the quality, fix distortions, and retrieve pertinent information,

a sequence of actions must be followed. Adjust photometric aberrations unique to each sensor to guarantee uniformity in color and brightness throughout the image. Adjust for distortions in geometry brought on by differences in geography, Earth's curvature, and viewing angles of satellite sensors. To create an accurate planimetric description of the image, this stage entails orthorectification.

Adjust for environmental variables including skies, haze, and particles to enhance the image's quality. For applications involving remote sensing, this is especially crucial. If required, adjust the image's spatial resolution to conform to the analysis's specifications or to match other datasets. This frequently entails resampling methods such as cubic or bilinear convolution.

3.3. Training and Testing the data using Faster-RCNN. The Faster RCNN is a two-stage detecting architecture that involves the categorization and localisation of objects in the second phase and the creation of areas in the first. Fast RCNN has a quick detection process and is dependent on external region recommendations. According to recent research, CNN can localize items in CONV (convolutional) the layers, but its performance is less in fully linked layers. Consequently, a targeted quest for generating regional proposals took the place of CNN. They suggested replacing selective search with a precise and effective region proposal network (RPN) to generate region proposals. They split the structure into two components: fast RCNN for object categorization and the localization of operations and RPN for region proposal creation.

Design or utilize existing APIs (Application Programming Interfaces) that allow for smooth data exchange and communication between the deep learning system and existing software. This may involve developing custom middleware or adapters. Ensure that data formats, including input images and output recognition results, are standardized across systems to facilitate easy sharing and processing.

The categorization and placement of objects using bounding boxes is carried out by an extensive number of convolutional layers used in RPN and the last convolutional layers in the faster RCNN. Figure 4.2 shows the network topology of the faster RCNN. When features are retrieved by CONV layers, RPN creates $k \times n \times n$ anchor boxes with varying aspect ratios and sizes. Every $n \times n$ anchor is transformed into a low dimensions vector, like 512 for the group known as Visual Geometry Group (VGG) and 256 for ZF. These vectors are then fed into two fully linked layers, which comprise layers for object categorization and bounding box regressors.

Since RPN is a sort of fully convolutional network, it shares features with the rapid RCNN and facilitates the computing of region suggestions efficiently. Instead of using manually created features, CNN uses faster RCNN only for feature extraction (Figure 3.2). Using three hundred suggestions per image, the faster RCNN with the VGG16 model reaches object detection accuracy on the PASCAL VOC dataset at 5 frames per second on the GPU. The author investigated the role of region suggestion the generation through selective search and region proposal generation through CNN considering the quicker RCNN growth. They concluded that CNN-based RPN includes less geometric data for identifying objects in the CONV Layers as opposed to FC layers.

$$(himg, wimg, x, y, w, h, objectives)$$

K is generated at each sliding window location when training a faster RCNN with anchors and various proposals. A class probability of object or not object is represented by a 2K score in the CLS layer, whereas the 4K boxes with coordinates in the Reg layer. K anchors, sometimes known as boxes, are the subject of the K parameter. They produce nearly WHK anchors at the convolutional feature map $W \times H$ by using $k = 9$ with three scales and three aspect ratios at each sliding window. To solve the multiscale problem based on anchors, Faster RCNN employs CNN for features computed on a single scale image. Sharing features and addressing multiscale at a lower cost are two advantages of this.

They give each object a binary label that indicates whether the object is present or absent for training purposes. They give anchors a favourable label. Anchors can be computed in two different methods. Ground truth boxes assign labels to numerous anchors. Initially, one picks those anchors whose crossover over union is high with a ground truth box. Secondly, one selects those anchors whose intersection over union is bigger than 0.7 with a ground truth box. As a result, the second requirement is inappropriate for accurate anchor prediction. As a result, they apply the initial criterion, which gives anchors positive labels and has the highest IOU with the ground truth box.

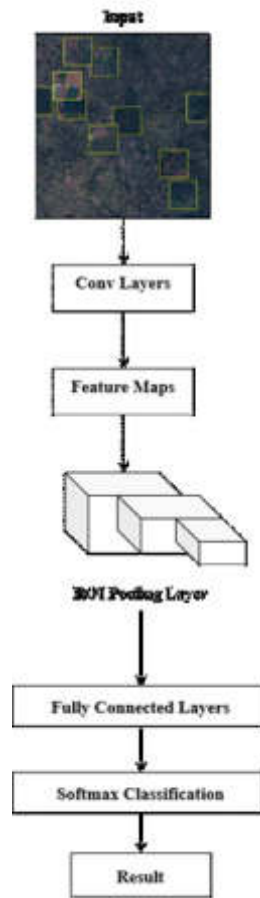


Fig. 3.2: Overall Process of Proposed Method

4. Result Analysis. Our work concentrated on creating and refining a Faster R-CNN-based algorithm for target detection in satellite photos of aircraft and digital image combining. When compared to traditional methods, the suggested algorithm showed notable improvements in terms of speed, accuracy, precision, recall and F1-score.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \quad (4.1)$$

$$precision = \frac{TP}{TP + FP} \times 100 \quad (4.2)$$

$$recall = \frac{TP}{TP + FN} \quad (4.3)$$

When evaluating the accuracy of Faster R-CNN target recognition on satellite images, the model's predictions are usually compared with ground truth annotation. Determine how many times the model's predictions coincide with the actual data (accurately predicted targets). Find out how many targets there are in the dataset overall. Utilize the formula to determine the accuracy. Remember that although while accuracy is a widely used metric, it might not be enough in all situations, particularly when working with datasets that are unbalanced

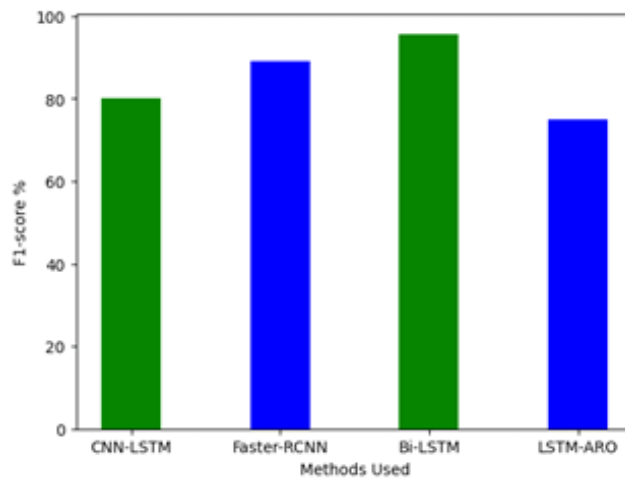


Fig. 4.2: F1-Score

or in situations where false positives or false negatives have distinct outcomes. In these circumstances, further measures such as recall, precision, and F1-score may be considered to offer a more thorough assessment of the model's effectiveness in target detection on satellite photos. In figure ??shows the evaluation of Accuracy.

A high F1-score suggests a good trade-off between precision and recall when evaluating a Faster R-CNN model for satellite image target recognition, indicating that the algorithm is successfully locating and recognizing targets in the images. When analysing F1-score results, it's crucial to consider the requirements of the application as well as the implications of false positives and false negatives. In figure 4.2 shows the evaluation of Precision. When employing Faster R-CNN for object detection tasks in satellite pictures, precision is an essential evaluation criterion. By quantifying the precision of the model's positive predictions, one can ascertain the proportion of projected positive instances that turn out to be true positives. figure 4.3 shows the evaluation.

Recall is a crucial parameter in satellite image analysis that assesses the model's accuracy in identifying and capturing all pertinent instances of the target class in the dataset when employing a Faster R-CNN-based algorithm. Recall is critical in the context of satellite photography since it offers insights into the algorithm's

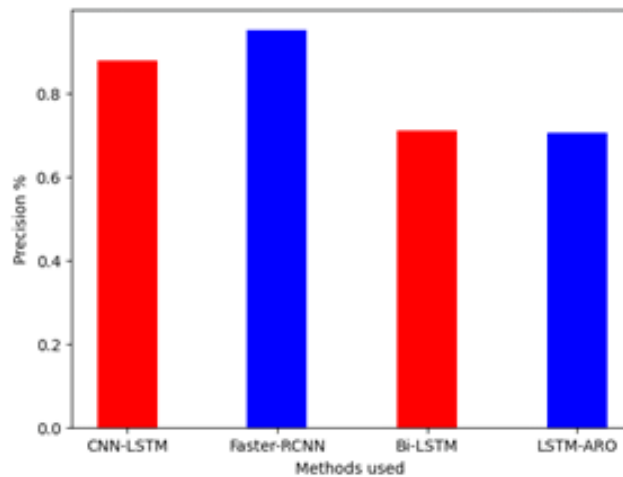


Fig. 4.3: Precision

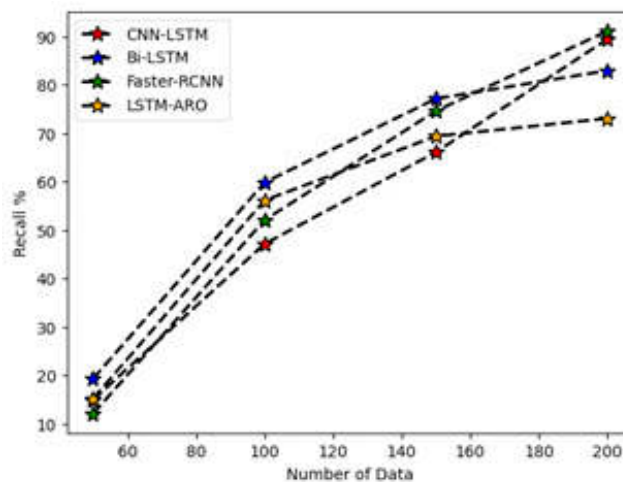


Fig. 4.4: Recall

capacity to identify every instance of the target class, guaranteeing that no significant data is overlooked. A high recall score means that the model can identify most real instances of the target class and effectively reduce false negatives. In figure 4.4 shows the evaluation of Recall.

5. Conclusion. To combine efficiency and accuracy in the domains of computer vision and image processing, this work employs deep learning approaches to enhance target detection and digital picture processing. Traditional ways of merging images and target recognition face challenges from varying scenarios, variable environmental conditions, and a broad range of image attributes. The study we conducted proposes a novel approach to address these problems by using deep learning techniques to extract significant features and patterns from several sources that offer a range of images. A sophisticated deep learning system that can identify ordered presentations of input photographs is created and trained as part of the study process. We will examine if temporal and geographical links in the visual data can be captured by quicker RCNN. Deep Learning techniques will be applied to utilize pre-trained networks on large datasets to optimize the model's performance.

The method will be evaluated using benchmark datasets and compared to traditional image processing techniques. Quantitative measurements like precision, recall, and F1-score will be used to assess the algorithm's correctness and reliability. Additionally, qualitative assessments will be carried out to ascertain the composite images' visual appeal and interpretive potential.

Acknowledgement. This work was sponsored in part by the Basic Ability Improvement Project of Young and middle-aged people in Guangxi - Research on the integrated development path of "Intangible cultural heritage + Cultural Innovation" from the perspective of Rural Revitalization - taking Bobai miscanthus weaving as an example (2023KY1102).

REFERENCES

- [1] Q. ALI, M. J. THAHEEM, F. ULLAH, AND S. M. SEPASGOZAR, *The performance gap in energy-efficient office buildings: how the occupants can help?*, *Energies*, 13 (2020), p. 1480.
- [2] G. CALLEBAUT, G. LEENDERS, J. VAN MULDER, G. OTTOY, L. DE STRYCKER, AND L. VAN DER PERRE, *The art of designing remote iot devicetechnologies and strategies for a long battery life*, *Sensors*, 21 (2021), p. 913.
- [3] S. CHAKRABORTY, J. PHUKAN, M. ROY, AND B. B. CHAUDHURI, *Handling the class imbalance in land-cover classification using bagging-based semisupervised neural approach*, *IEEE Geoscience and Remote Sensing Letters*, 17 (2019), pp. 1493–1497.
- [4] S. I. KHAN, Z. QADIR, H. S. MUNAWAR, S. R. NAYAK, A. K. BUDATI, K. D. VERMA, AND D. PRAKASH, *Uavs path planning architecture for effective medical emergency response in future networks*, *Physical Communication*, 47 (2021), p. 101337.
- [5] Y. LI, Y. SHI, K. WANG, B. XI, J. LI, AND P. GAMBA, *Target detection with unconstrained linear mixture model and hierarchical denoising autoencoder in hyperspectral imagery*, *IEEE Transactions on Image Processing*, 31 (2022), pp. 1418–1432.
- [6] M. U. LIAQUAT, H. S. MUNAWAR, A. RAHMAN, Z. QADIR, A. Z. KOUZANI, AND M. P. MAHMUD, *Sound localization for ad-hoc microphone arrays*, *Energies*, 14 (2021), p. 3446.
- [7] S. LOW, F. ULLAH, S. SHIROWZHAN, S. M. SEPASGOZAR, AND C. LIN LEE, *Smart digital marketing capabilities for sustainable property development: A case of malaysia*, *Sustainability*, 12 (2020), p. 5402.
- [8] MANJU, P. BHAMBU, AND S. KUMAR, *Target k-coverage problem in wireless sensor networks*, *Journal of Discrete Mathematical Sciences and Cryptography*, 23 (2020), pp. 651–659.
- [9] A. MAQSOOM, B. ASLAM, M. E. GUL, F. ULLAH, A. Z. KOUZANI, M. P. MAHMUD, AND A. NAWAZ, *Using multivariate regression and ann models to predict properties of concrete cured under hot weather*, *Sustainability*, 13 (2021), p. 10164.
- [10] H. S. MUNAWAR, *Flood disaster management: Risks, technologies, and future directions*, *Machine Vision Inspection Systems: Image Processing, Concepts, Methodologies and Applications*, 1 (2020), pp. 115–146.
- [11] H. S. MUNAWAR, A. W. HAMMAD, S. T. WALLER, M. J. THAHEEM, AND A. SHRESTHA, *An integrated approach for post-disaster flood management via the use of cutting-edge technologies and uavs: A review*, *Sustainability*, 13 (2021), p. 7925.
- [12] H. S. MUNAWAR, H. INAM, F. ULLAH, S. QAYYUM, A. Z. KOUZANI, AND M. P. MAHMUD, *Towards smart healthcare: Uav-based optimized path planning for delivering covid-19 self-testing kits using cutting edge technologies*, *Sustainability*, 13 (2021), p. 10426.
- [13] H. S. MUNAWAR, S. I. KHAN, Z. QADIR, A. Z. KOUZANI, AND M. P. MAHMUD, *Insight into the impact of covid-19 on australian transportation sector: An economic and community-based perspective*, *Sustainability*, 13 (2021), p. 1276.
- [14] H. S. MUNAWAR, M. MOJTAHEDI, A. W. HAMMAD, M. J. OSTWALD, AND S. T. WALLER, *An ai/ml-based strategy for disaster response and evacuation of victims in aged care facilities in the hawkesbury-nepean valley: A perspective*, *Buildings*, 12 (2022), p. 80.
- [15] H. S. MUNAWAR, S. QAYYUM, F. ULLAH, AND S. SEPASGOZAR, *Big data and its applications in smart real estate and the disaster management life cycle: A systematic analysis*, *Big Data and Cognitive Computing*, 4 (2020), p. 4.
- [16] H. S. MUNAWAR, F. ULLAH, S. I. KHAN, Z. QADIR, AND S. QAYYUM, *Uav assisted spatiotemporal analysis and management of bushfires: A case study of the 2020 victorian bushfires*, *Fire*, 4 (2021), p. 40.
- [17] H. S. MUNAWAR, F. ULLAH, S. QAYYUM, S. I. KHAN, AND M. MOJTAHEDI, *Uavs in disaster management: Application of integrated aerial imagery and convolutional neural network for flood detection*, *Sustainability*, 13 (2021), p. 7547.
- [18] Z. QADIR, S. I. KHAN, E. KHALAJI, H. S. MUNAWAR, F. AL-TURJMAN, M. P. MAHMUD, A. Z. KOUZANI, AND K. LE, *Predicting the energy output of hybrid pv-wind renewable energy system using feature selection technique for smart grids*, *Energy Reports*, 7 (2021), pp. 8465–8475.
- [19] Z. QADIR, A. MUNIR, T. ASHFAQ, H. S. MUNAWAR, M. A. KHAN, AND K. LE, *A prototype of an energy-efficient maglev train: A step towards cleaner train transport*, *Cleaner Engineering and Technology*, 4 (2021), p. 100217.
- [20] M. A. SHAUKAT, H. R. SHAUKAT, Z. QADIR, H. S. MUNAWAR, A. Z. KOUZANI, AND M. P. MAHMUD, *Cluster analysis and model comparison using smart meter data*, *Sensors*, 21 (2021), p. 3157.
- [21] A. TAHIR, H. S. MUNAWAR, J. AKRAM, M. ADIL, S. ALI, A. Z. KOUZANI, AND M. P. MAHMUD, *Automatic target detection from satellite imagery using machine learning*, *Sensors*, 22 (2022), p. 1147.
- [22] J. THEILER, A. ZIEMANN, S. MATTEOLI, AND M. DIANI, *Spectral variability of remotely sensed target materials: Causes, models, and strategies for mitigation and robust exploitation*, *IEEE Geoscience and Remote Sensing Magazine*, 7 (2019), pp. 8–30.

- [23] F. ULLAH, *A beginners guide to developing review-based conceptual frameworks in the built environment*, Architecture, 1 (2021), pp. 5–24.
- [24] F. ULLAH AND F. AL-TURJMAN, *A conceptual framework for blockchain smart contract adoption to manage real estate deals in smart cities*, Neural Computing and Applications, 35 (2023), pp. 5033–5054.
- [25] F. ULLAH, S. KHAN, H. MUNAWAR, Z. QADIR, AND S. QAYYUM, *Uav based spatiotemporal analysis of the 2019–2020 new south wales bushfires*, sustainability 2021, 13, 10207, 2021.
- [26] F. ULLAH, S. QAYYUM, M. J. THAHEEM, F. AL-TURJMAN, AND S. M. SEPASGOZAR, *Risk management in sustainable smart cities governance: A toe framework*, Technological Forecasting and Social Change, 167 (2021), p. 120743.
- [27] F. ULLAH, S. SEPASGOZAR, F. TAHMASEBINIA, S. M. E. SEPASGOZAR, AND S. DAVIS, *Examining the impact of students' attendance, sketching, visualization, and tutors experience on students' performance: A case of building structures course in construction management*, Construction Economics and Building, 20 (2020), pp. 78–102.
- [28] F. ULLAH AND S. M. SEPASGOZAR, *Key factors influencing purchase or rent decisions in smart real estate investments: A system dynamics approach using online forum thread data*, Sustainability, 12 (2020), p. 4382.
- [29] F. ULLAH, S. M. SEPASGOZAR, M. J. THAHEEM, C. C. WANG, AND M. IMRAN, *Its all about perceptions: A dematel approach to exploring user perceptions of real estate online platforms*, Ain Shams Engineering Journal, 12 (2021), pp. 4297–4317.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Jan 6, 2024

Accepted: Feb 9, 2024



RESEARCH ON LEARNING EFFICIENCY IMPROVEMENT STRATEGIES OF PUBLIC ENGLISH PERSPECTIVE BASED ON ANT COLONY ALGORITHM

QINGZHU LI*

Abstract. A ground-breaking smartphone app called EngageLearnPro was created to improve learning efficiency improvement strategies in the context of public English education. With the use of cutting-edge technologies like Ant Colony Optimization (ACO) and Long Short-Term Memory (LSTM), this app creates a dynamic and captivating method of language learning. Intelligent sequence modeling is made possible by the combination of LSTM and allows for customized learning paths that adjust based on the progress of each individual user. On the other hand, ACO maximizes the app’s decision-making processes, improving the overall effectiveness of language learning techniques. The decision to use a mobile app environment for this initiative was made in light of the fact that smartphones are widely used and can provide education to a wider range of people. By utilizing the interactive and user-centric qualities of mobile devices, EngageLearnPro makes sure that learning happens naturally in users’ everyday lives. By combining LSTM and ACO technologies, a customized and adaptive learning experience is provided, accommodating a wide range of learning styles. EngageLearnPro offers an inclusive, cutting-edge, and effective platform with the goal of closing the gap in public English education. We hope to transform language learning by combining the best features of LSTM and ACO into a mobile application that is not only efficient but also fun and available to students of all backgrounds and ability levels.

Key words: Learning efficiency improvement, public English education, LSTM, ACO, mobile application

1. Introduction. Public English instruction stands out as a crucial catalyst for promoting cultural integration and improving language proficiency in diverse communities [1, 8]. In today’s globally interconnected world, where effective communication cuts across linguistic barriers and becomes a basic prerequisite for both personal and professional growth, the importance of English proficiency is highlighted. Because public education systems are made to serve people from a variety of socioeconomic backgrounds, they are essential to democratizing language opportunities by providing equal access to English language instruction [7]. Nevertheless, there are frequently issues with the effectiveness of language education methods in these systems. Problems like outmoded techniques, unequal access to resources, and the requirement for customized methods continue to impede the best possible achievement of language learning objectives [6]. This emphasizes the urgent need for creative fixes and cutting-edge approaches that can handle the particular difficulties in public English education and guarantee a more welcoming, flexible, and productive language learning environment for people from all backgrounds [16, 4].

Machine learning techniques have become powerful tools for improving learning efficiency in the modern era [2]. These approaches, which include deep learning and reinforcement learning, add a new level of customization to individualized learning by utilizing large datasets and complex algorithms [3]. Neural networks in particular, which are deep learning models, show exceptional capacity for pattern recognition, allowing for personalized learning pathways. By offering continuous feedback mechanisms, personalized content recommendations, and real-time adaptability, these techniques improve language learning efficiency [14]. Additionally, learner behavior can be analyzed using machine learning algorithms, which can be used to pinpoint the advantages and disadvantages of educational interventions. In addition to enhancing learning outcomes, the combination of machine learning and language learning technologies advances educational methodologies, resulting in a dynamic environment tailored to each student’s individual needs [4]. As machine learning develops further, it will play a more and more important role in improving learning efficiency by providing creative answers to complex issues in the field of education.

In this proposed study, we introduce a novel mobile application design called “EngageLearnPro”. A strategic choice based on accessibility, engagement, and inclusivity led to promoting learning efficiency improvement

*Puyang Vocational and Technical College Puyang ,457000,China (qingzhulitarget1@outlook.com)

strategies in the mobile app market within the Public English perspective. Because they are so common and easily incorporated into daily life, mobile apps provide public education systems with a platform to reach a wide range of users. Due to the widespread use of smartphones, these applications offer a portable and practical medium that guarantees users' lives are seamlessly integrated with language learning. Mobile apps' interactive and user-centered design encourages interaction, which makes learning more dynamic and engaging.

The motivation behind the development of EngageLearnPro, a pioneering smartphone application, stems from a commitment to revolutionize public English education through the integration of advanced technologies. In an era where the accessibility and efficiency of educational tools are paramount, EngageLearnPro emerges as a beacon of innovation, designed to bridge the educational divide and foster an inclusive environment for language learners worldwide. At the core of EngageLearnPro's design philosophy is the utilization of Ant Colony Optimization (ACO) and Long Short-Term Memory (LSTM) algorithms. These cutting-edge technologies synergize to create a dynamic, engaging, and personalized language learning experience. LSTM's intelligent sequence modelling capabilities enable the app to offer customized learning paths that evolve in real-time, adapting to the unique pace and progress of each user. This personalization ensures that learners are not just passive recipients of information but active participants in their educational journey.

With the use of cutting-edge technologies, EngageLearnPro is a novel and creative app that makes learning English more effective and pleasurable. Through Adaptive Learning Paths, the app creates dynamic and personalized learning journeys based on each user's progress. To further increase the effectiveness of language learning, it also makes use of intelligent technologies like Ant Colony Optimization (ACO) for better decision-making and Long Short-Term Memory (LSTM) for intelligent sequence modeling [19, 15, 21]. Real-time feedback mechanisms reinforce language usage, quickly correct errors, and give instant insights into progress [20, 17]. Learning is made interesting and enjoyable by the app's interactive, user-centered design, which includes multimedia and gamification features. Because it makes use of smartphones, EngageLearnPro is widely accessible and aims to make English learning possible for people from a variety of backgrounds. Its inclusive design takes into account different learning styles and skill levels. Through the use of cutting-edge technology, seamless integration into daily life, individualized learning experiences, and a commitment to making learning fun, EngageLearnPro makes learning a language an exciting and positive experience.

The contribution of the paper as follows

1. Proposed the mobile app design of EngageLearnPro, for the promotion of learning efficiency improvement strategies in the mobile app market within the Public English perspective.
2. Novel EngageLearnPro which leverages the techniques of LSTM and ACO, where ACO for better decision-making and LSTM for intelligent sequence modelling.
3. In the context of English learning, a thorough analysis and assessment of the proposed EngageLearnPro are conducted using Chinese colleges.
4. Proposed efficacy was demonstrated with valid experiments

2. Research Analysis. [12] The authors of this study discuss the difficulties brought about by the growing size of colleges and universities as well as the growing complexity of teaching duties as a result of an increase in student body and a diversity of course offerings. They suggest a new college scheduling algorithm built on top of an enhanced hybrid optimization strategy based on genetic ant colonies. Improvements like gene infection crossover, fitness-enhanced elimination, and parallel fuzzy adaptive mechanisms are incorporated into the algorithm, which improves convergence, stability, and operating speed. [10] The study investigates the use of an ant colony algorithm-based College English microcurriculum model, utilizing the effectiveness of ant behavior in learning difficult tasks. Teachers have initially resisted using this ant colony algorithm in microcourse design, even though researchers and educators have long used it. This is in spite of the fact that the State supports the use of digital teaching resources. The model's goal is to simplify college students' learning procedures in the context of a foundational course like College English by taking inspiration from the cooperative and algorithmic behavior of ants. [13] Through the use of a bipartite graph model derived from graph theory and clustering analysis on student performance data, this study presents a comprehensive approach to teaching quality evaluation in performing arts courses. The evaluation process is made more robust by applying an ant colony algorithm that takes memory capacity and prior knowledge mastery into account. The results provide theoretical and practical insights for the development of performing arts courses, highlighting the need for

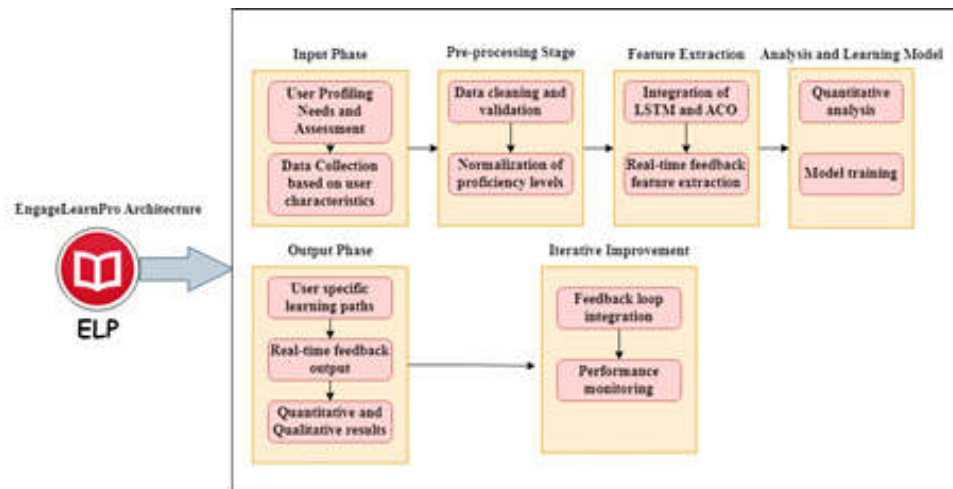


Fig. 3.1: Proposed EngageLearnPro App Design

specialized courses in art performance majors and suggesting targeted training for unqualified teachers based on the PDCA principle. [18] The study offers a multimedia comprehensive framework that integrates big data technology and multimedia teaching modes to address the problem of diverse materials in college English translation. An ant colony optimization algorithm serves as the foundation for the recursive neural network algorithm, which is tested and shown to significantly increase accuracy and retention rate. The suggested method offers a promising way to improve college English translation teaching models by skillfully integrating big data and multimedia into the English teaching process.

3. Proposed design of EngageLearnPro. EngageLearnPro's methodology is divided into discrete stages to guarantee a thorough and data-driven approach to individualized language learning. To understand the diverse needs of English language learners, a comprehensive needs assessment and user profiling are carried out during the input phase. This entails gathering necessary input data, such as user demographics, skill levels, and preferred methods of learning. Concurrently, information is acquired regarding language ability, past educational experiences, and personal preferences in order to further customize learning pathways. Pre-processing involves standardizing proficiency level scales to create a uniform scale for consistent analysis, as well as cleaning and verifying the gathered data to assure accuracy. Advanced algorithms for sequence modeling and effective learning strategies, such as LSTM and ACO, are integrated during the feature extraction phase. Features are taken out for real-time feedback systems and adaptive learning pathways, giving useful information about user progress, mistakes, and proper language usage. Quantitative analysis is used to assess EngageLearnPro's efficacy as we move on to the analysis and learning model phase. This entails analyzing the overall impact on language learning and using pre- and post-assessment scores for proficiency level analysis. Then, using the features that were extracted, machine learning models are trained to find patterns in user behavior and results, which helps to improve the learning model. Personalized learning paths with adaptive features based on the unique characteristics and progress of each user are generated as outputs during the output phase. Users receive real-time feedback that emphasizes areas for development and reinforces the proper use of language. To gauge EngageLearnPro's overall impact, both quantitative and qualitative results such as user feedback and proficiency level improvements are provided. The iterative improvement phase, which brings the methodology to a close, establishes a continuous feedback loop to incorporate user feedback into the learning model. By using this data, features are improved iteratively, learning paths are adjusted, and the user experience is enhanced overall. EngageLearnPro is continuously improved to meet changing user needs through data-driven enhancements through ongoing performance monitoring that makes use of usage analytics and user feedback.

3.1. Proposed EngageLearnPro feature Extraction. Recent research of [9, 7, 11, 5] examines the state-of-the-art mobile app technology utilized to increase the effectiveness of English language learning. Under the studies, there are lucid discussions. Based on the discussions, we are now going to carry out the suggested EngageLearnPro feature extraction procedure.

EngageLearnPro utilizes cross-platform development frameworks to ensure consistent functionality and user experience across iOS and Android devices. This approach allows the app to maintain high performance and adapt to the specific hardware and software configurations of each platform. The app features an adaptive design that adjusts to different screen sizes and resolutions, ensuring that the learning experience is seamless on a wide range of devices, from high-end smartphones to more basic models with limited processing power. EngageLearnPro incorporates performance optimization techniques such as efficient memory management, data compression, and lazy loading of resources to minimize the app's footprint and ensure smooth operation even on devices with lower hardware capabilities.

3.1.1. LSTM for intelligent sequence modeling. Long Short-Term Memory (LSTM) is essential for developing a dynamic and adaptable learning environment when using EngageLearnPro to improve English learning efficiency. Neural network architectures known as long short-term memory (LSTMs) are excellent at understanding and simulating sequential patterns, which makes them especially useful for language learning tasks. By using LSTMs to intelligently model sequences, EngageLearnPro is able to gradually understand and retain the subtleties of English language learning. This implies that the system can make dynamic adjustments to its learning approach based on a user's progress data, guaranteeing an effective and personalized learning trajectory. To maximize each learner's experience, LSTMs, for instance, can dynamically adjust the course material based on an analysis of how the learner interacts with the content and identify difficult areas.

Algorithm 12 Intelligence sequence modeling

- 1: **Input:** E : Sequence of input, where $E = \{E_1, E_2, \dots, E_t\}$, H_{t-1} -previous hidden state, C_{t-1} -Previous Cell state, Weight matrices $w_{Fo}, w_{In}, w_{Ou}, w_c$; Bias Terms $b_{Fo}, b_{In}, b_{Ou}, b_c$.
 - 2: **Output:** h_t -current hidden state, c_t - current cell state
- Initialization**
- 3: Initialize h_o, c_o as the initial hidden and cell states.
 - 4: Define weight matrices $w_{Fo}, w_{In}, w_{Ou}, w_c$
 - 5: Define Bias term $b_{Fo}, b_{In}, b_{Ou}, b_c$.
 - 6: for each time step t
 - 7: calculate forget gate $Fo_t = \sigma(w_{Fo} \cdot [h_{t-1}, E_t] + b_{Fo})$
 - 8: Calculate the input gate $In_t = \sigma(w_{In} \cdot [h_{t-1}, E_t] + b_{In})$
 - 9: Calculate Candidate cell state $\bar{c}_t = \tanh(w_c \cdot [h_{t-1}, E_t] + b_c)$
 - 10: Update cell state $c_t = Fo_t * c_{t-1} + In_t * \bar{c}_t$
 - 11: Calculate output gate $Ou_t = \sigma(w_{Ou} \cdot [h_{t-1}, E_t] + b_{Ou})$
 - 12: Calculate hidden state $h_t = Ou_t * \tanh(c_t)$
 - 13: Output the current hidden state h_t and cell state c_t at each time step t
-

Intelligent sequence modeling is achieved through the use of the (LSTM) process within the algorithmic structure of EngageLearnPro. The algorithm reflects the essential steps in LSTM computation and is expressed as a set of equations. Establishing the initial hidden state H_o and cell state C_o as well as defining weight matrices $w_{Fo}, w_{In}, w_{Ou}, w_c$ and bias terms $b_{Fo}, b_{In}, b_{Ou}, b_c$ are all part of initialization. The algorithm determines the forget gate Fo_t , input gate In_t , and candidate cell state \bar{c}_t for each time step (t). The sigmoid (σ) and hyperbolic tangent \tanh functions are used in these calculations. Based on the input and forget gates, the cell state is updated. After determining the output gate Ou_t , the hidden state's current value h_t is calculated by multiplying the product of the output gate and the hyperbolic tangent of the updated cell state. The current hidden state h_t and cell state c_t are produced as the result of repeating this process at each time step. EngageLearnPro's LSTM algorithm allows it to dynamically modify its learning approach, capturing and remembering crucial data for efficient English learning.

3.1.2. ACO based Decision Making. ACO is a useful algorithmic tool to improve learning efficiency in the context of EngageLearnPro. The application of ACO, which is used to optimize decision-making processes

within the platform, is inspired by the foraging behavior of ants. The ACO in EngageLearnPro assists in identifying the most efficient learning strategies, much like ants leave pheromone trails to communicate and direct others toward the best paths. It functions by modeling the cooperative investigation of multiple learning trajectories, where each trajectory stands for a possible learning decision. Based on user interactions, the algorithm assesses the effectiveness of various paths and gradually improves them to accommodate unique learning styles. EngageLearnPro's ability to integrate ACO allows it to dynamically modify its approach, guaranteeing that the learning environment is efficient, adaptable, and customized to users' changing needs and preferences. The ACO algorithm is adapted from the study [19].

Algorithm 13 ACO based Decision Making

```

1:  $global\_best \leftarrow$  Build initial solution
2: Calculate pheromone trails limits:  $\tau_{min}$  and  $\tau_{max}$ 
3: Set pheromone trails values to  $\tau_{max}$ 
4:  $source\_solution \leftarrow global\_best$ 
5: for  $i \leftarrow 1$  to  $\#iterations$  do
6: for  $j \leftarrow 0$  to  $\#ants - 1$  do
7:  $route_{ant(j)}[0] \leftarrow u\{0, n - 1\}$  // Select first node randomly
8:  $min\_new\_edges \leftarrow calc\_num\_new\_edges()$ 
9:  $new\_edges \leftarrow 0$ 
10:  $K \leftarrow 1$ 
11: while  $K < n$  do
12:  $u \leftarrow route_{ant(j)}[K - 1]$ 
13:  $v \leftarrow select\_next\_node\ u \leftarrow route_{ant(j)}$ 
14:  $route_{ant(j)}[K] \leftarrow v$ 
15:  $K \leftarrow K + 1$ 
16: if  $(u, v) \notin source\_solution$  then
17:  $new\_edges \leftarrow new\_edges + 1$ 
18: Add  $v$  to  $LS\_checklist$ 
19: if  $new\_edges \geq min\_new\_edges$  then
▷ Complete  $route_{ant(j)}$  following source solution
20:  $u \leftarrow succ(source\_solution, v)$  // ...forward
21: while  $u \notin route_{ant(j)}$  do
22:  $route_{ant(j)}[K] \leftarrow u$ 
23:  $u \leftarrow succ(source\_solution, u)$ 
24:  $K \leftarrow K + 1$ 
25:  $u \leftarrow pred(source\_solution, u)$  // ...or backward
26: while  $u \notin route_{ant(j)}$  do
27:  $route_{ant(j)}[k] \leftarrow u$ 
28:  $u \leftarrow pred(source\_solution, u)$ 
29:  $k \leftarrow k + 1$ 
30:  $local\_search(route_{ant(j)}, LS\_checklist)$ 
31:  $iter\_best \leftarrow select\ shortest\ route_{ant(0)} \dots, route_{ant(\#ants-1)}$ 
32: if  $global\_best = \emptyset$  or  $iter\_best$  is shorter than  $global\_best$  then
33:  $global\_best \leftarrow iter\_best$ 
34: Update pheromone trails limits  $\tau_{min}$  and  $\tau_{max}$  using  $global\_best$ 
35: Evaporate pheromone according to  $\rho$  parameter
36:  $source\_solution \leftarrow Choose\ between\ global\_best\ and\ iter\_best$ 
37: Deposit pheromon

```

With its foundation in ACO, the EngageLearnPro decision-making algorithm prioritizes improving the learning paths available on the platform. The first step in the process is to initialize a global best solution, which serves as the foundation for the lessons that follow. Pheromone trails are created with predetermined boundaries that are initially set at maximum values, reflecting the allure of different learning paths. Counting the number of new edges in the learning route, individual ants representing different learning paths navigate

randomly chosen nodes over the course of iterations. The algorithm improves the flexibility and variety of learning paths by updating the route by taking into account nodes that are not present in the source solution. Interestingly, the algorithm uses local search to improve the efficiency of learning routes. This leads to a global update where the global best solution is adjusted based on its emptiness or the superiority of the iteration's best route. The iteration's best route, representing the shortest path among the ant population, is identified. Then, based on this global best solution, pheromone trail limits are updated, impacting the appeal of particular routes. Pheromone deposition and evaporation are steps that follow. Pheromone deposition is influenced by the decision to select between the global best and the iteration's best solution, which shapes the learning paths. The goal of this iterative and adaptive process is to continuously improve the learning paths in EngageLearnPro by dynamically adapting to changing user preferences and needs. In the end, the algorithm aims to maximize the educational process by creating more efficient and interesting avenues for language learning.

Providing in-app surveys, direct feedback forms, social media interactions, and email communications. This variety ensures a broad spectrum of insights, from usability issues to suggestions for new features. Collected feedback is categorized into various segments such as app performance, user interface (UI) design, learning content quality, and algorithmic suggestions. Advanced data analysis techniques, including natural language processing (NLP), are employed to sift through the feedback, identifying common themes, user needs, and potential areas for enhancement.

4. Tests and Validation.

4.1. Simulation Plan. This section examines the effectiveness of EngageLearnPro using a simulation of English classroom instruction; the dataset's original source is taken from the study [22].

This simulation plan's goal is to thoroughly assess how the EngageLearnPro Mobile App affects students' learning outcomes and levels of interest in the context of teaching English in a classroom. There are two classes of 61 business English students participating in the experimental design. While the Control Class uses conventional teaching techniques, the Experimental Class makes use of the EngageLearnPro mobile app. In order to determine the starting proficiency level of students in both classes, pre-experiment tests are administered, and information on smartphone ownership and mobile network status is gathered. During the implementation stage, instructors in the Experimental Class use the EngageLearnPro Mobile App in their English lessons, while the Control Class follows traditional teaching strategies that place limitations on the use of mobile devices. The main goals of observation techniques are to evaluate student participation, classroom performance, and EngageLearnPro's real-time interactivity. Learning outcomes are measured using post-experiment tests administered at the end of a semester. Improvement is assessed by comparing the test results with pre-experiment data. Evaluation metrics encompass comparing test results from before and after the experiment, gauging student involvement and interaction, and analyzing the general dynamics and efficacy of the classroom. Improved learning outcomes in the Experimental Class, higher interest and participation attributable to EngageLearnPro's interactive features, and observational data offering insights into the effect on real-time interactivity are among the anticipated results.

4.2. Evaluation Criteria.

4.2.1. Comparison before testing. Figure 4.1 presented performance metrics for two different classes: the Experimental class and the Control class that cover a range of performance categories. These categories include the number of students who scored below 60, the number of students who scored between 60 and 80, the number of students who scored between 80 and 100, and the average score attained. Each metric provides information on the performance of the classes as well as the distribution of proficiency, engagement levels, and overall academic achievement of the students. This extensive dataset makes it possible to assess the educational dynamics of the classes in a more nuanced way, offering insights into things like academic proficiency, student motivation, and the distribution of scores across various performance thresholds. The scores are collected before testing the proposed app of EngageLearnPro.

4.3. Post Testing Results.

4.3.1. Experimental class test results using EngageLearnPro App. The EngageLearnPro app's performance for the Experimental class are shown in Figure 4.2 across various performance categories. Of the

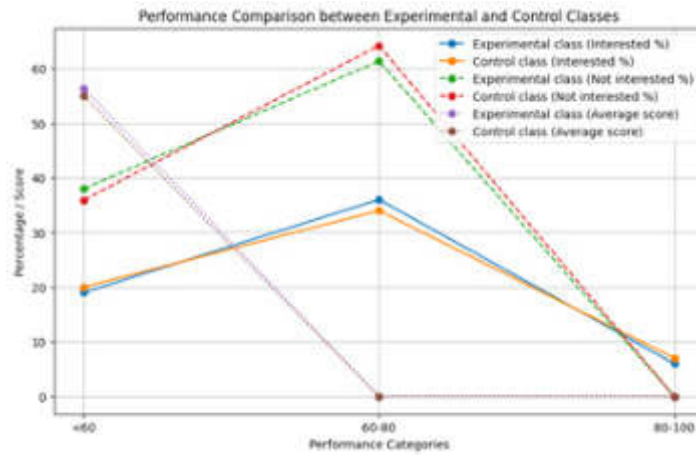


Fig. 4.1: Comparison results before testing

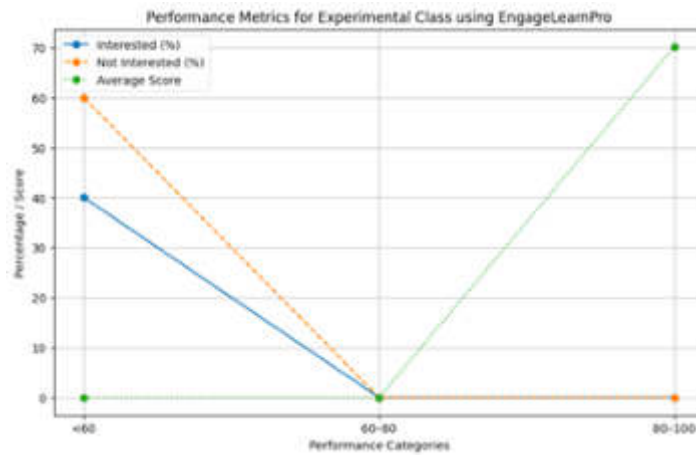


Fig. 4.2: Experimental using EngageLearnPro App

students in the Experimental class, 40% show an interest in learning, according to the 'Interested Percentage' line; the remaining 60% do not show any interest. The 'Average Score' line indicates a high average score of 70.2%, indicating a positive correlation between academic performance and the EngageLearnPro app. This pattern shows that students are distributed evenly across proficiency levels, demonstrating the effectiveness of the app in capturing users' attention and creating a positive learning environment.

4.3.2. Control Class test results (Without EngageLearnPro App). The results of the Control Class show that it can be difficult to engage a sizable portion of students when employing traditional teaching methods, with 33.50% of students expressing interest and 65.50% not interested was shown in Figure 4.3. The marginally lower percentage of interested students than in the Experimental Class may point to the inadequacies in traditional teaching strategies for maintaining high interest. The Experimental Class's average score of 55.8 indicates that the Control Class attains similar academic results. The aforementioned data underscores the constraints of conventional pedagogical approaches in sustaining widespread student interest and involvement. The outcomes of the Control Class highlight the potential of the EngageLearnPro app, which can address issues with traditional teaching methods while producing similar academic results thanks to its interactive

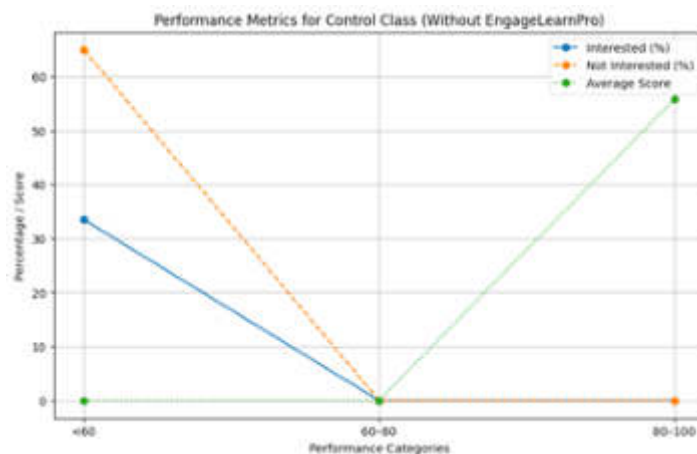


Fig. 4.3: Control Class results without using EngageLearnPro App

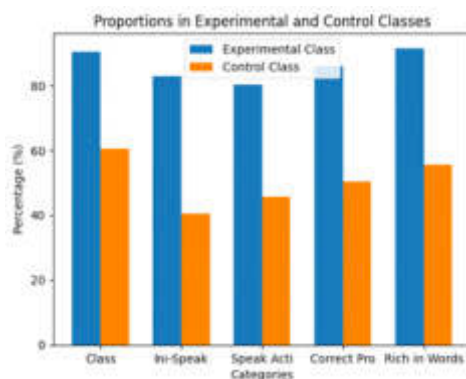


Fig. 4.4: Overall Performance indications based on Experimental and Control Class of EngageLearnPro

and adaptive features. This demonstrates how effective EngageLearnPro is at fostering a more dynamic and captivating learning environment in the context of teaching English to the general public.

4.3.3. Performance Analysis. The effectiveness of the EngageLearnPro app is demonstrated by the Figure 4.4, which compares and contrasts important language learning categories between the Experimental and Control classes. The Experimental class shows a significantly higher percentage (82.74%) in the "Initiative to Speak" category than the Control class (40.34%), suggesting that the app encourages students to be proactive in their verbal communication. In similar fashion, the Experimental class does better than the Control class in areas like "Rich in Words" (91.46% vs. 55.47%), "Speak Actively" (80.25% vs. 45.77%), and "Correct Pronunciation" (85.89% vs. 50.22%). These differences highlight how the app promotes active engagement, improves pronunciation, and enhances vocabulary acquisition. The increasing percentages in these language proficiency categories show that EngageLearnPro is more effective than traditional teaching methods at creating a dynamic and engaging learning environment. The app's beneficial effects on various facets of language learning in the Experimental class are clearly depicted in Figure 4.4, which supports the idea that it can improve language learning outcomes in general.

5. Conclusion. In conclusion, the research on the effectiveness of the EngageLearnPro app for language learning provides strong proof of its beneficial effects on student proficiency and engagement when compared

to conventional teaching techniques. The app's success in fostering a lively and engaging learning environment was demonstrated by the Experimental class's consistent displays of increased interest, active participation, and improved language skills. To guarantee generalizability, a larger dataset and a variety of learner groups are nevertheless required. Furthermore, the study concentrated on immediate results; a longitudinal approach would reveal information about the app's long-term effects. Subsequent investigations may examine the incorporation of increasingly sophisticated technologies, evaluate the app's flexibility in various educational settings, and examine particular aspects that enhance its effectiveness. Notwithstanding these drawbacks, the study's encouraging results imply that EngageLearnPro has potential as a cutting-edge tool for improving language learning encounters, opening the door for more developments in technology-driven education.

REFERENCES

- [1] M. AMERI, *The use of mobile apps in learning english language*, Budapest International Research and Critics in Linguistics and Education (BirLE) Journal, 3 (2020), pp. 1363–1370.
- [2] P. ASTHANA AND B. HAZELA, *Applications of machine learning in improving learning environment*, Multimedia big data computing for IoT applications: concepts, paradigms and solutions, (2020), pp. 417–433.
- [3] Y. BAO, Y. ZHU, AND F. QIAN, *A deep reinforcement learning approach to improve the learning performance in process control*, Industrial & Engineering Chemistry Research, 60 (2021), pp. 5504–5515.
- [4] L. A. CHANNA, *English in pakistani public education: Past, present, and future*, Language Problems and Language Planning, 41 (2017), pp. 1–25.
- [5] X. CHEN, *Evaluating language-learning mobile apps for second-language learners*, Journal of Educational Technology Development and Exchange (JETDE), 9 (2016), p. 3.
- [6] C. DIAZ LARENAS, P. ALARCON HERNANDEZ, M. ORTIZ NAVARRETE, ET AL., *A case study on efl teachers beliefs about the teaching and learning of english in public education*, (2015).
- [7] X. FAN, K. LIU, X. WANG, AND J. YU, *Exploring mobile apps in english learning*, Journal of Education, Humanities and Social Sciences, 8 (2023), pp. 2367–2374.
- [8] P. GOU, *Teaching english using mobile applications to improve academic performance and language proficiency of college students*, Education and Information Technologies, (2023), pp. 1–15.
- [9] Y. HAO, K. S. LEE, S.-T. CHEN, AND S. C. SIM, *An evaluative study of a mobile application for middle school students struggling with english vocabulary learning*, Computers in Human Behavior, 95 (2019), pp. 208–216.
- [10] X. HU, *Micro course model in college english reform based on ant colony algorithm*, in International Conference on Frontier Computing, Springer, 2022, pp. 1193–1199.
- [11] K. ISHAQ, F. ROSDI, N. A. M. ZIN, AND A. ABID, *Usability and design issues of mobile assisted language learning application*, International Journal of Advanced Computer Science and Applications, 11 (2020).
- [12] T. LI, Q. XIE, AND H. ZHANG, *Design of college scheduling algorithm based on improved genetic ant colony hybrid optimization*, Security and Communication Networks, 2022 (2022).
- [13] B. LU AND Y. HE, *Influence of teaching and course evaluation of performing arts students based on improved ant colony algorithm and data fusion*, Security and Communication Networks, 2022 (2022).
- [14] T. POORNAPPRIYA AND R. GOPINATH, *Application of machine learning techniques for improving learning disabilities*, Int. J. Electr. Eng. Technol.(IJEET), 11 (2020), pp. 392–402.
- [15] M. PUSHPA, *Aco in e-learning: Towards an adaptive learning path*, International Journal on Computer Science and Engineering, 4 (2012), p. 458.
- [16] P. SAYER, *more & earlier: Neoliberalism and primary english education in mexican public schools*, L2 Journal, 7 (2015).
- [17] Q. SHAN, *Intelligent learning algorithm for english flipped classroom based on recurrent neural network*, Wireless Communications and Mobile Computing, 2021 (2021), pp. 1–8.
- [18] L. SHI, M. DUJIANG, AND P. GAO, *A high performance computing technology powered multimedia fusion model in university english translation*, PeerJ Computer Science, 9 (2023), p. e1608.
- [19] R. SKINDEROWICZ, *Improving ant colony optimization efficiency for solving large tsp instances*, Applied Soft Computing, 120 (2022), p. 108653.
- [20] C. SRIHARSHA, S. RITHWIK, K. P. PRAHLAD, AND L. S. NAIR, *Intelligent learning assistant using bert and lstm*, in 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), IEEE, 2021, pp. 1–6.
- [21] L. SUN, *College english teaching evaluation with neural network*, Mathematical Problems in Engineering, 2022 (2022).
- [22] H. TU, *Application of mobile app in english teaching in an intelligent environment*, Mobile Information Systems, 2021 (2021), pp. 1–9.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Jan 6, 2024

Accepted: Feb 9, 2024



RESEARCH ON THE APPLICATION OF MOOCS BASED ON REINFORCEMENT LEARNING IN COLLEGE ENGLISH TEACHING

YU GU*

Abstract. In the field of teaching college English, this study explores the integration of reinforcement learning concepts with Massive Open Online Courses (MOOCs). "LearnFlex," the suggested framework, is intended to support an environment that is dynamic and flexible for learning. By offering thorough English language courses and utilizing reinforcement learning techniques, LearnFlex leverages the inherent benefits of MOOCs to customize and enhance the learning process for every student. This study's main goal is to assess how well LearnFlex works in the context of teaching college English to improve student performance, engagement, and general satisfaction. Through the integration of educational technology, machine learning, and pedagogical methodologies, LearnFlex aims to offer significant insights that support the ongoing development of efficient and customized online learning. The study contributes to the larger objective of improving teaching strategies by utilizing cutting-edge technologies to build a learning ecosystem that is more adaptable and focused on the needs of students. This study aims to provide insights for future improvements in online education, specifically in the area of language training, by conducting a thorough examination of LearnFlex's effects.

Key words: Reinforcements learning, MOOCs, college English teaching, student engagement, personalized online education

1. Introduction. Massive Open Online Courses (MOOCs) have become a powerful and revolutionary force in the modern educational scene. MOOCs are a paradigm change in education, providing a wide range of worldwide audiences with flexible and easily available learning options [2, 15]. The fundamental quality of MOOCs is their capacity to democratize education by bridging the gap between socioeconomic and geographic constraints that frequently inhibit traditional learning paradigms [15]. MOOCs facilitate self-paced and self-directed learning experiences by giving learners unparalleled access to a plethora of educational content through the use of digital platforms. This method has shown to be very helpful in meeting the different requirements and interests of students, encouraging lifelong learning across a range of subjects. MOOCs have become a powerful and revolutionary force in the modern educational scene [13]. These are a paradigm change in education, providing a wide range of worldwide audiences with flexible and easily available learning options. The fundamental quality of MOOCs is their capacity to democratize education by bridging the gap between socioeconomic and geographic constraints that frequently inhibit traditional learning paradigms. MOOCs facilitate self-paced and self-directed learning experiences by giving learners unparalleled access to a plethora of educational content through the use of digital platforms [6, 16]. This method has shown to be very helpful in meeting the different requirements and interests of students, encouraging lifelong learning across a range of subjects.

The use of MOOCs in college English instruction offers a flexible and dynamic method of teaching the language [19]. MOOCs provide a wide range of benefits for teaching college English that support the advancement of conventional pedagogical approaches [10, 3]. The availability of top-notch English language instruction to a multicultural and international student body is one significant benefit. Students can interact with rich, standardized content regardless of where they are in the world, guaranteeing a thorough and consistent educational experience. Due to MOOCs' inherent flexibility, students can advance at their own speed and according to their own preferences and learning styles [18]. Furthermore, MOOCs' interactive format encourages active engagement and participation through discussion boards, multimedia components, and group projects. These elements not only improve how students engage with the curriculum, but they also help students feel more like a community. MOOCs' flexibility is essential for meeting students' diverse skill levels because they offer

*Zhengzhou University of Industrial Technology, Zhengzhou, 450000, China (yuguindustriail12@outlook.com)

customized assessments and content to meet each student's needs. Because of their scalability and affordability, MOOCs significantly improve access to high-quality English language instruction while providing a significant response to the drawbacks of conventional teaching techniques [15, 7]. This study investigates how these benefits can be increased by incorporating MOOCs that are enhanced by the principles of reinforcement learning, opening the door to a novel and successful method of teaching college English.

MOOCs have many benefits, but they are not without drawbacks. A noteworthy obstacle pertains to learner engagement and completion rates [23]. MOOCs frequently encounter elevated dropout rates, which can be attributed to various factors, including inadequate personalization, restricted interactivity, and inadequate flexibility to meet individual learning requirements [9, 1]. The integration of reinforcement learning with the machine learning-based C4.5 algorithm is suggested as a solution to these drawbacks. This combined strategy, called "LearnFlex," aims to add personalization and adaptability to MOOCs to increase their efficacy in teaching college English. LearnFlex uses reinforcement learning to dynamically modify learning paths and content according to each student's progress, encouraging long-term engagement [4, 11]. The C4.5 algorithm's integration makes a further contribution by examining learner data to find patterns that allow the system to offer customized interventions and recommendations [5]. Thus, LearnFlex is a creative step toward reducing the drawbacks of traditional MOOCs in the context of teaching college English and fostering a more responsive, flexible, and student-centered MOOC environment.

The burgeoning intersection of educational technology and language training presents a unique opportunity to revolutionize the way college English is taught and learned. This research is motivated by the potential to harness Massive Open Online Courses (MOOCs), enriched with reinforcement learning techniques, to create a more dynamic, personalized, and effective learning environment. "LearnFlex," our proposed framework, stands at the forefront of this educational innovation, aiming to redefine college English teaching through the strategic integration of MOOCs and machine learning principles. The core objective of this study is to meticulously evaluate the efficacy of LearnFlex in enhancing college English education by focusing on three primary outcomes: student performance, engagement, and overall satisfaction. By leveraging the scalability and accessibility of MOOCs, coupled with the adaptive capabilities of reinforcement learning, LearnFlex is designed to offer a tailored educational experience that meets the diverse needs of learners, thereby overcoming the limitations of traditional one-size-fits-all approaches.

Central to our motivation is the belief that every student's learning journey is unique. Traditional educational models often fail to accommodate individual learning styles, pacing, and preferences, leading to suboptimal outcomes. LearnFlex seeks to address these challenges by employing reinforcement learning algorithms that adapt the learning content and pathways based on real-time feedback from student interactions. This approach ensures that the learning process is continuously optimized for each student, fostering a deeper understanding and mastery of the English language.

The contribution of the paper as follows:

1. The goal of the study is to improve the flexibility and customization of online learning by introducing a novel approach called "LearnFlex".
2. The suggested LearnFlex combines the reinforcement-based DQN technique with the MOOCs-based C4.5 algorithm.
3. Proposed LeanFlex uses reinforcement learning to modify learning paths and content according to each student's progress, introducing dynamic adaptability.
4. The C4.5 algorithm's integration yields analytical insights through learner data analysis, pattern recognition, and customized recommendation generation.
5. The efficacy of the proposed LearnFlex is proved with valid experiments.

1.1. Related work. This study [12] highlights problems with low intelligence and ineffective teaching effects in online college English cross-cultural instruction using MOOCs. Through the use of artificial intelligence and cloud computing, the research presents an enhanced MOOC model and algorithm designed to satisfy the needs of online education. Requirements analysis is used to build functional modules, and control experiments verify the model's functionality and show how effective it is at improving the effectiveness of English language instruction across cultural boundaries in virtual settings. This study [14] investigates how to improve English instruction in China by integrating a Constructive English MOOC system based on the RBF algorithm.

The system runs smoothly and efficiently, completing tasks in 3–7 seconds with an astounding 98% efficiency. The platform encourages students to participate actively in their education by developing their capacity for independent research and piquing their interest in English studies by replacing traditional teaching methods with a technology-enhanced approach. The results of the study demonstrate the benefits of using the RBF algorithm and effective teaching techniques, indicating important new directions in the field and the possibility of revolutionizing English instruction. This study [22] uses artificial intelligence (AI) emotion recognition and neural network algorithms to overcome the shortcomings of conventional cross-cultural English teaching models. The created cross-cultural O2O English teaching system uses background models to track and identify students' emotions along with intelligent recognition and management. Robust performance and efficient online teaching control are demonstrated by the comprehensive O2O teaching model that integrates both online and offline components. The model has been successful in raising student emotional engagement and teaching effectiveness in cross-cultural English instruction, according to the study's statistical tests.

This study [21] improves on traditional teaching quality evaluation by introducing a comprehensive evaluation model for MOOC teaching in accounting using the Rete algorithm. In order to provide a thorough quality assessment, the model evaluates the MOOC teaching mode, compares teaching data with the Rete algorithm, and establishes evaluation standards and weight calculations. The model's usefulness in controlling MOOC accounting teaching quality is demonstrated by its practical application, which also indirectly improves teaching outcomes.

This study [17] focuses on learning objectives and Bloom's taxonomy to address the difficulty of automating the pedagogical classification of MOOCs. The research uses transfer learning through BERT to achieve large-scale and automatic annotation even with a small annotated dataset. The results of the experiments show that the classifier's complexity has little effect on performance; the best results are obtained when dense layers are added to BERT, dropout is included, and ReLU activation functions are used. In the context of MOOCs, the study demonstrates the value of transfer learning for pedagogical annotation, opening the door to better quality control and comprehension of their pedagogical models. The underuse [8] of a wealth of university data is the subject of this study, which focuses on forecasting the chance of withdrawal for incoming students. The application uses the C4.5 algorithm to convert large amounts of data into a decision tree, which allows for the rule-based classification of new students. The tree structure of the system makes it easier to identify possible student withdrawals early on, which helps management make decisions more quickly. The application, which was created using the waterfall model and PHP and MySQL, attempts to improve strategic planning and lower the chance of student attrition by using insights from data.

2. Methodology.

2.1. LearnFlex Overview. In order to create an intelligent and adaptable learning environment for college English teaching, LearnFlex's system architecture integrates DQN and MOOC-based C4.5 techniques was illustrated in Figure ???. To guarantee a smooth integration of these technologies, this entails identifying system components, data flows, and interactions. Data privacy and ethical considerations are given top priority while a plan is developed to collect learner information, behavioral data, and pertinent contextual information. The next step of the preparation process involves creating a training dataset D that is thorough and includes learner attributes, past interactions, and pertinent features. After that, the MOOC-based C4.5 algorithm is put into practice, beginning with the training dataset's initialization and the feature-representing attributes' definition. To improve the classification process, the C4.5 algorithm is used, which includes creating decision trees, using TG-C4.5 to select the optimal attribute features, parallel processing with Hadoop for scalability, and using MapReduce to calculate the information gain ratio. After that, LearnFlex incorporates the DQN reinforcement learning techniques. This entails putting the DQN training procedure into practice as well as initializing the Q-value function and replay memory rm . C4.5 and DQN work together to enable adaptive decision-making, which improves the system's capacity to adapt to shifting learning dynamics. Various user roles, including instructors, administrators, and students, are taken into account in user-centric design considerations. The design of role-based systems adapted to particular business requirements is informed by an analysis of the behavioral traits and basic information about the user.

Next, the system is deployed in the College English Teaching environment after being initialized and incorporating both DQN and C4.5 techniques based on MOOCs. A loop for continuous improvement is created to

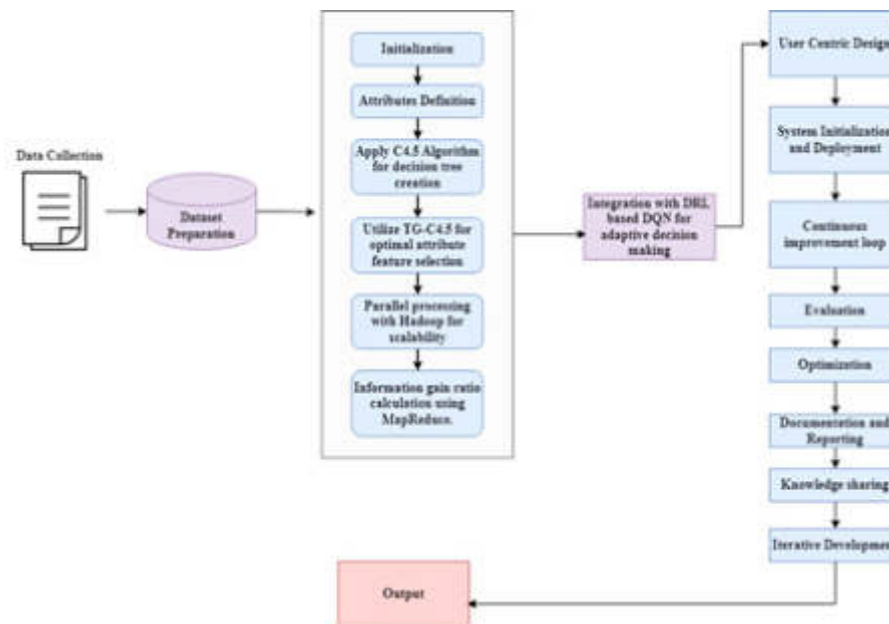


Fig. 2.1: Proposed LearnFlex Design

track system performance, collecting user input by watching and iteratively improving the LearnFlex algorithm in response to input. During the evaluation phase, the LearnFlex system's efficacy is evaluated with respect to user satisfaction, learning outcomes, and adaptability to changing educational needs. After that, optimization enables parameter and algorithm improvements based on the evaluation's findings. The process relies heavily on reporting and documentation, with the methodology, algorithms, and system components all having extensive documentation. The LearnFlex system generates thorough reports that include recommendations and results. In order to spread discoveries and insights, academic publications, conferences, and knowledge-sharing platforms place a strong emphasis on knowledge sharing. The LearnFlex system is continuously evolving and being improved with the iterative development approach. Finally this allows the provision of a customized and successful learning environment.

2.2. Proposed LearnFlex Approach.

2.2.1. MOOCs based C4.5 algorithm. This section discusses the C4.5 concept, which is based on MOOCs and was taken from the study [4]. The MOOCs-based C4.5 diagrammatic flow is depicted in Figures 2.1 & 3.1 and 3.2 of the study. We now carry out the simple algorithmic steps as follows.

The MOOCs-based C4.5 algorithm takes into account various learning styles, behaviors, and course content in order to manage the complex data from online courses. This tool performs a few crucial tasks. In order to help the system determine what might be most effective for each individual, it first arranges and makes sense of the data about each student. It's similar to customizing the educational process to meet each student's needs. Based on the C4.5 algorithm, this tool is also highly effective in generating decision trees, which aid in the system's intelligent decision-making regarding how to teach, what content to deliver, and even suggesting customized learning plans. Furthermore, it excels at managing multiple types of data simultaneously, which helps it overcome the difficulties of online learning. It increases the precision of identifying students' areas of strength and weakness, which is critical in the teaching of English. It also becomes even more potent when combined with Hadoop, a big data platform, allowing it to handle massive volumes of data effectively. All things considered, LearnFlex is made smarter by this tool the MOOCs-based C4.5 algorithm which aids in the efficient use of data to produce a customized and successful learning environment for college students studying English.

The system employs advanced load balancing techniques to distribute traffic evenly across servers, preventing any single server from becoming a bottleneck. This not only enhances performance but also ensures a smooth and responsive experience for all users. LearnFlex leverages CDNs to cache and deliver content from servers closest to the user’s location. This significantly reduces bandwidth usage and lowers the server load, enabling faster content delivery even in high-demand scenarios.

The MOOCs-based C4.5 algorithm in LearnFlex starts with initializing the training dataset D and defining attributes A . Attribute selection is performed using the TG-C4.5 algorithm, denoted as $TG - C4.5 = \text{taylorseries}(C4.5, GINIIndex)$ here $GINIIndex$ is introduced to enhance classification performance. The $GINI$ index (Gini) is defined as

$$Gini = 1 - \sum_{i=1}^n \left(\frac{|DC_I|}{|D|} \right)^2$$

capturing the impurity of a dataset. For a specific attribute A , $GiniSplitA(D)$ is computed as $GiniSplitA(D) = \sum_{j=1}^m \left(\frac{|D_j|}{D} Gini(D_j) \right)$. The mean sum of GINI indices ($Sum - GiniAF(D)$) is then calculated as

$$\frac{1}{s} \sum_{i=1}^s \sum_{j=1}^x \left(\frac{|D_{ij}|}{|D|} Gini(D_{ij}) \right)$$

To improve $GainRatio(A)$ calculation, the algorithm considers this mean sum:

$$GainRatio(A) = \frac{Gain(A)}{SplitInfoA(T) - \alpha \cdot Sum - Gain.SplitAF(D)}$$

Here, α is a tuning parameter. Decision trees are designed in the root node using C4.5 (D, A), and TG - C4.5 is applied for optimal attribute feature selection. The algorithm integrates parallel processing with Hadoop, employing $HD - TG - C4.5$ for large-scale data processing. The Gain Ratio Calculation with MapReduce in the proposed MOOCs-based C4.5 algorithm involves a series of steps aimed at optimizing the decision tree structure and enhancing the algorithm’s effectiveness. In next step parallel statistics are employed to calculate the information gain ratio for a specific attribute feature A using the MapReduce paradigm. The $GainRatio(GainRatio(A))$ is determined as the ratio of the Gain for attribute A to the Split Information for A subtracted by a weighted sum. Moving to Step 16, the algorithm adjusts or replaces misclassification results in the training set with probability errors, refining the classification accuracy by considering the likelihood of errors. Next, it introduces the calculation of the probability $P_{wj} \in \left[\frac{nwj}{V+s}, \frac{nwj+s}{V+s} \right]$ for each node using the Integrated Error Probability (IEP) model, contributing to a probabilistic approach to pruning. It focuses on estimating values for the number of nodes and subsets $n(t)$ a crucial step in determining the structure of the decision tree. Pruning decisions are made in the next step based on a condition related to the number of nodes $n(t) \leq n(T_i) + SE[n(T_i)]$ leading to the removal of specific branches and optimizing the decision tree structure. The algorithm analyzes user basic information and behavioural characteristics, highlighting the importance of understanding individual learner attributes for personalized learning experiences. Next, it integrates the algorithm into the design of an education system for predicting performance, emphasizing practical applications in the educational environment. Consideration of different user roles and their corresponding business requirements is addressed, ensuring a tailored approach to diverse stakeholders. The final step, involves continuous monitoring of system performance and gathering feedback, facilitating an iterative process for enhancements and refinements based on real-world outcomes and user experiences.

Algorithm 1 focuses on estimating values for the number of nodes and subsets $n(t)$ a crucial step in determining the structure of the decision tree. Pruning decisions are made in the next step based on a condition related to the number of nodes n leading to the removal of specific branches and optimizing the decision tree structure. The algorithm analyzes user basic information and behavioural characteristics, highlighting the importance of understanding individual learner attributes for personalized learning experiences. Next, it integrates the algorithm into the design of an education system for predicting performance, emphasizing practical applications

Algorithm 14 MOOCs based C4.5 algorithm

- 1: Initialize the training dataset D with learner information and behavioral characteristics.
- 2: Define attributes A representing features in the dataset.

Attribute selection using TG-C4.5 algorithm

- 3: Apply Taylor series based C4.5 Algorithm to calculate information gain rate.
- 4: Utilize TG-C4.5 algorithm for optimal attribute feature selection.

$$TG - C4.5 = \text{taylorseries}(C4.5, \text{GINIIndex})$$

GINI Index and splitting information

- 5: Introduce GINI index to improve classification performance
- 6: Define the GINI index as $Gini = 1 - \sum_{i=1}^n \left(\frac{|DC_I|}{|D|} \right)^2$
- 7: For attribute A , compute $GiniSplitA(D) = \sum_{j=1}^m \left(\frac{|D_j|}{|D|} Gini(D_j) \right)$
- 8: Calculate the mean sum of GINI indices using $Sum - GiniAF(D) = \frac{1}{s} \sum_{i=1}^s \sum_{j=1}^x \left(\frac{|D_{ij}|}{|D|} Gini(D_{ij}) \right)$
- 9: Improve the GainRatio calculation by considering the mean sum of GINI indices
 Compute $GainRatio(A) = \frac{Gain(A)}{SplitInfoA(T) - \alpha \cdot Sum - GainSplitAF(D)}$
- 10: Design decision trees in the root node to select and train optimal attribute features by utilizing $DecisionTree = C4.5(D, A)$
- 11: Apply the TG-C4.5 algorithm for optimal attribute feature selection.

Parallel Processing with Hadoop

- 12: Implement parallel decision algorithms using the Hadoop platform framework.
- 13: Design HD-TG-C4.5 for processing large scale data

$$HD - TG - C4.5 = \text{ParallelProcessing}(D, \text{Hadoop})$$

- 14: Utilize Hadoop Distributed File System (HDFS) and Hadoop MapReduce for efficient data processing.
 Gain Ratio Calculation with MapReduce
- 15: Perform Parallel statistics on the information gain ratio of attribute feature

$$GainRatio(A) = \text{MapReduce}(\text{InformationGainRatio})$$

$$GainRatio(A) = \frac{Gain(A)}{SplitInfoA(T) - \alpha \cdot Sum - GainSplitAF(D)}$$

$$Sum - GiniAF(D) = \frac{1}{s} \sum_{i=1}^s \sum_{j=1}^x \left(\frac{|D_{ij}|}{|D|} Gini(D_{ij}) \right)$$

- 16: Replace the misclassification result of the training set with probability error.
- 17: Calculate the probability of each node using the IEP model

$$P_{wj} \in \left[\frac{nwj}{V+s}, \frac{nwj+s}{V+s} \right]$$

Probability based Pruning

- 18: Estimate the values for the number of nodes and subsets

$$n(t) = \max\{e(pt) + 0.5 | pt \in kt(w)\}$$

MOOCs based algorithm output

- 19: Prune when the condition is satisfied

$$n(t) \leq n(T_i) + SE[n(T_i)]$$

- 20: Analyse the user basic information and behavioural characteristics.
 - 21: Design as education system for performance prediction.
 - 22: Consider different user roles and their corresponding business requirements.
 - 23: Monitor system performance and gather feedback.
-

in the educational environment. Consideration of different user roles and their corresponding business requirements is addressed, ensuring a tailored approach to diverse stakeholders. The final step, involves continuous monitoring of system performance and gathering feedback, facilitating an iterative process for enhancements and refinements based on real-world outcomes and user experiences.

2.2.2. Integrating DQN for adaptive decision making. One of the main functions of the Deep Q-Network (DQN) integration in the LearnFlex system is to improve adaptive decision-making. Using techniques from reinforcement learning, DQN is used to optimize workload scheduling decisions. The main goal is to develop a dynamic and intelligent system that can adjust on its own to changing circumstances in the context of teaching college English. With the help of DQN, LearnFlex is able to maximize expected rewards when making decisions by drawing on past experiences that are stored in a replay memory. LearnFlex can now customize its recommendations and responses thanks to this integration, giving students a more successful and individualized learning experience. LearnFlex aspires to provide an adaptive, intelligent, and data-driven educational platform that meets the specific needs of each individual learner by fusing MOOC-based C4.5 algorithms with DQN.

```

1: Input:  $n_{min}, n_{max}, \Delta, \alpha, \gamma, \delta$ 
2: Output: Workload Scheduling Decision
3: Initialize replay memory  $rm$  to capacity  $n$ 
4: Initialize action value function  $Q$  with random weights  $\theta$ 
5: Initialize target action value function  $\hat{Q}$  with weights  $\theta = \theta^-$ 
6: for  $e$  dopisode = 1,  $m$  do
7:   Initialize sequence  $s_1 = \{x_1\}$  and preprocessed sequence  $\sigma_1 = \sigma_1(s_1)$ 
8:   for  $dot$  = 1,  $T$  do
9:     with probability  $\delta$  select a random action  $a_t$ 
10:    otherwise select  $a_t = \max_a Q(s_t, a, \theta)$ 
11:    Execute action  $s_t$  and observe reward  $r_t$  and next state  $x_{t+1}$ 
12:    Set  $s_{t+1} = s_t, a_t, x_{t+1}$  and preprocess  $\sigma_{t+1} = \sigma(s_{t+1})$ 
13:    Store transition  $(\sigma_t, a_t, r_t, \sigma_{t+1})$  in replay memory  $rm$ 
14:    Sample random minibatch of transitions  $(\sigma_J, a_J, r_J, \sigma_{J+1})$  from replay memory  $rm$ 
15:    Set  $y_j = \left\{ r_{J+\gamma} \max_a \hat{Q}_{(s_{J+1}, a', \theta^-)}^{r_J} \right\}$ 
16:    Perform gradient descent step on  $(Y_J - Q(s_J, a_J, \theta))^2$  with respect to the network
    Parameters  $\vartheta$ 
17:    Every  $C$  steps reset  $\hat{Q} = Q$ 
18:   end for
19: end for
Online making workload scheduling decision
20: Load the parameters  $\vartheta$ ;
21: Calculate action-value  $Q(s_t, a; \theta)$ 
22: Output  $a_t = \operatorname{argmax} Q(s_t, a; \theta)$ 

```

In the LearnFlex system, the algorithm entails training a DQN for adaptive workload scheduling decisions. First, two action-value functions, Q and \hat{Q} , are initialized with random weights in a replay memory rm . Episodes are used in the training process to select states and actions at random or in an attempt to maximize the Q-value. Transitions are recorded in the replay memory, and rewards are tracked. The Q-network parameters are updated using a gradient descent step to minimize the temporal difference between the target and predicted Q-values after a random minibatch of transitions is sampled on a regular basis. Periodically, the target network is reset. The algorithm moves into the online decision-making stage after training. The trained parameters are loaded, and action values for the possible actions and the current state are computed. The action with the highest computed Q-value is chosen to make the decision. During online execution, this procedure is repeated, giving the LearnFlex system the ability to schedule workload adaptively using the learned Q-values.

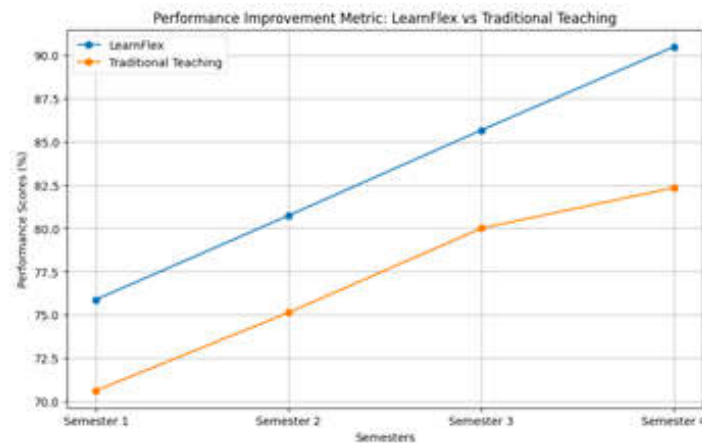


Fig. 3.1: Performance Comparison

3. Results and Experiments.

3.1. Simulation Setup. We move forward with the evaluation of the suggested LearnFlex based on the study [20]. Based on the dataset that comprises universities, colleges, and schools. Here, we evaluate the proposed LearnFlex using data from colleges.

3.2. Evaluation Criteria. The Figure 3.1 shows that the LearnFlex framework is significantly more effective than traditional teaching methods in improving student performance. LearnFlex continuously outperformed conventional methods in terms of academic achievement during the observed periods. The LearnFlex scores show a consistent upward trend, ranging from 75.89 to 90.47. The scores for the traditional teaching approach, on the other hand, range from 70.64 to 82.34, indicating a somewhat slower and less noticeable improvement trajectory. The significant improvement in performance highlights LearnFlex's beneficial effects on college English learning outcomes for students. The upward trend in LearnFlex scores indicates the platform's efficacy and adaptability in meeting the varied needs of students, which in turn creates an environment that promotes better learning outcomes and more fulfilling educational experiences in general.

The efficacy of LearnFlex in improving student engagement is demonstrated by multiple metrics, indicating noteworthy improvements over conventional teaching approaches were present in Figure 3.2. First off, LearnFlex outperforms the conventional approach with an astounding participation rate of 85.77% compared to 70.27% for the former. This notable rise suggests that LearnFlex successfully promotes a greater degree of student participation in learning activities. When it comes to the interaction of materials, LearnFlex performs exceptionally well, scoring 90.22%, while traditional approaches fall short at 75.89%. This disparity indicates a more in-depth examination of learning resources as LearnFlex users interact with course materials in a more comprehensive manner. When it comes to learning activities, LearnFlex is still ahead of the competition, with an 80.34% score as opposed to 79.42% for the traditional method. While both strategies demonstrate respectable levels of student involvement, LearnFlex guarantees that students stay actively engaged throughout a variety of educational tasks, making for a more dynamic learning environment. Another area where LearnFlex excels is communication, where it achieves an amazing 95.06%, compared to 80.44% for traditional methods. LearnFlex's capacity to establish an engaging and cooperative learning environment is demonstrated by its exceptional communication-fostering capabilities. Traditional approaches, on the other hand, show a clear weakness in this area. With an overall engagement metric of 80.12% as opposed to the traditional method's 60.75%, LearnFlex is clearly a comprehensive solution. This notable difference suggests that LearnFlex's dynamic and adaptable features greatly enhance students' overall engagement in a comprehensive way. Taken as a whole, the metrics show that LearnFlex improves engagement in a number of ways, which makes it a more impactful and promising approach than more conventional teaching techniques.

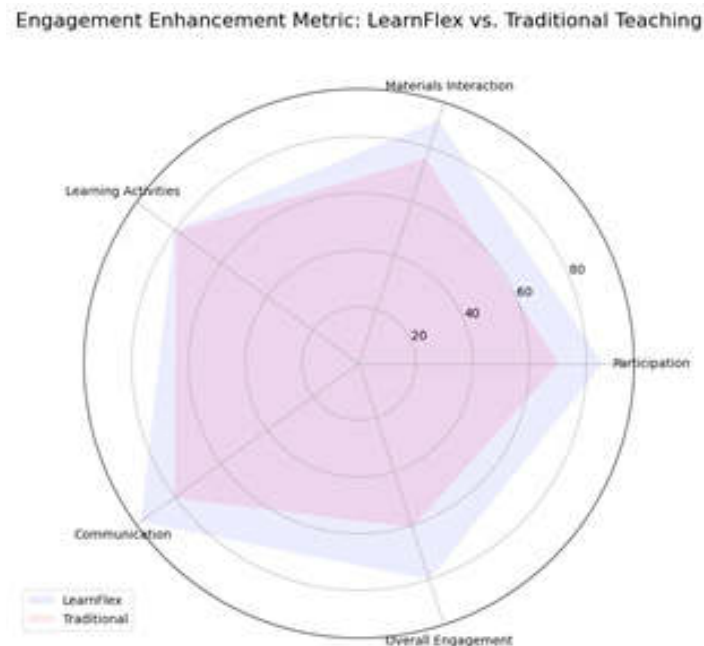


Fig. 3.2: Student Engagement Comparison

An extensive examination of user satisfaction ratings shows that the suggested LearnFlex is demonstrably effective when compared to traditional teaching methods, illustrated Figure 3.3. These scores, which range from 1 to 5, represent how satisfied students are overall with different aspects of their educational experience. LearnFlex has a usability score of 4.2, which is higher than traditional methods' score of 3.8. This discrepancy highlights the LearnFlex platform's greater usability by showing that students believe it to be more approachable and user-friendly. In terms of content relevance, LearnFlex performs better than traditional methods, scoring 4.5 out of 4.0. This discrepancy suggests that learners view the information in LearnFlex as more individualized and relevant to their particular learning requirements, which raises the overall significance of the course content. LearnFlex's learning process receives an astounding 4.8, vastly surpassing the 4.1 score assigned to conventional methods. This significant difference emphasizes that LearnFlex environments provide students with a more engaging and enriching educational experience than traditional classroom settings. Regarding assistance, LearnFlex receives a satisfaction rating of 4.3, which is higher than the 3.9 that traditional approaches receive. This indicates that LearnFlex does a great job of offering the required frameworks for support, creating an atmosphere in which students feel sufficiently assisted throughout their educational journey. LearnFlex is clearly superior, as evidenced by the overall satisfaction metric 4.6, which averages scores from a variety of categories. Traditional methods score 4.2. This conclusion demonstrates that LearnFlex provides students with a thorough and satisfying learning experience, not only meeting but exceeding their expectations. In essence, LearnFlex is an excellent pedagogical framework that is responsive to the changing needs of students and goes above and beyond conventional approaches to create a positive and productive learning environment.

4. Conclusion. In conclusion, the LearnFlex framework that has been suggested, which combines DRL-based DQN methods with MOOCs-based C4.5 algorithms, is a revolutionary development in the field of teaching college English. The well-thought-out system architecture uses C4.5 to make efficient decisions based on learner characteristics and integrates MOOCs to improve adaptability. This procedure is further improved by the TG-C4.5 algorithm, which takes into account the peculiarities of online learning environments. Simultaneously, the incorporation of DQN methodologies incorporates reinforcement learning, guaranteeing flexible decision-making and customized experiences. The user-centric design, which accommodates the various roles of teachers, admin-

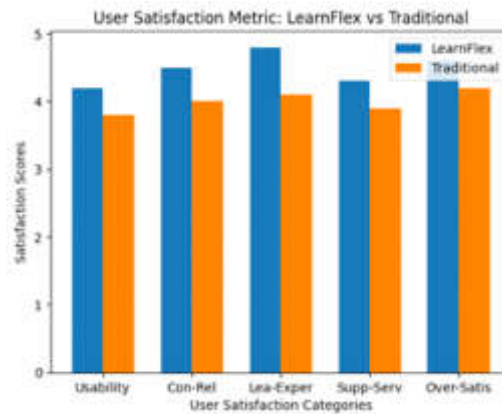


Fig. 3.3: User Satisfaction

istrators, and students, complements the algorithmic prowess. LearnFlex shows efficacy in enhancing learning outcomes, user satisfaction, and adaptability to changing educational needs through methodical initialization, continuous improvement loops, and a strong evaluation framework. The C4.5 algorithm, which is based on MOOCs, greatly enhances personalized learning by being highly scalable and adept at creating decision trees. Massive datasets in MOOCs present challenges for scalability, but Hadoop's parallel processing capabilities further improve it. All things considered, LearnFlex shows itself to be an intelligent, scalable, adaptive system that is ready to transform the way college English is taught by fusing state-of-the-art algorithms and technologies into a framework that is performance-driven and centered on the needs of the user.

5. Limitations and Discussions. LearnFlex is a promising framework for bettering college English instruction, but it has a number of issues that need to be acknowledged and discussed carefully in order to be improved. The significant reliance on learner data raises privacy concerns, necessitating a careful balancing act between personalization and privacy. Algorithmic complexity is introduced by the combination of DRL-based DQN and MOOCs-based C4.5, which presents difficulties for system upkeep and user-friendliness. Further research is necessary to generalize LearnFlex's success outside of the English classroom, taking into account the various ways that students learn different subjects. In areas with inadequate infrastructure, accessibility issues, such as technology requirements, may limit its effectiveness. Further investments in professional development are warranted because teacher training is essential for effective deployment. Important topics for discussion include striking a balance between standardized quality and personalization, creating feedback loops, addressing cost-effectiveness and scalability, and taking social and cultural nuances into account. Participating in these conversations will help to improve LearnFlex and create a dynamic, flexible learning environment for college English instructors. Future online learning platforms will likely incorporate more sophisticated tools for collaboration and community building, mimicking the social learning environments of traditional classrooms. These tools will support real-time collaboration on projects, peer-to-peer learning, and mentorship opportunities, breaking down the barriers of isolation often associated with online education.

REFERENCES

- [1] S. S. AHMED, E. KHAN, M. FAISAL, AND S. KHAN, *The potential and challenges of moocs in pakistan: a perspective of students and faculty*, Asian Association of Open Universities Journal, 12 (2017), pp. 94–105.
- [2] M. H. BATURAY, *An overview of the world of moocs*, Procedia-Social and Behavioral Sciences, 174 (2015), pp. 427–433.
- [3] J. CHEN, *An e-portfolio-based model for the application and sharing of college english esp moocs.*, Higher Education Studies, 7 (2017), pp. 35–42.
- [4] X. CHEN, *Design and research of mooc teaching system based on tg-c4. 5 algorithm*, Systems and Soft Computing, 5 (2023), p. 200064.

- [5] J. DAI, *Selective forwarding unit placement in edge computing based on dqn for real-time communications*, in GLOBECOM 2022-2022 IEEE Global Communications Conference, IEEE, 2022, pp. 4510–4516.
- [6] R. DENG, P. BENCKENDORFF, AND D. GANNAWAY, *Progress and new directions for teaching and learning in moocs*, Computers & Education, 129 (2019), pp. 48–60.
- [7] Y. DING AND H.-Z. SHEN, *English language moocs in china: Learners' perspective.*, The EuroCALL Review, 28 (2020), pp. 13–22.
- [8] D. ERLAN, *C4. 5 algorithm application for prediction of self candidate new students in higher education*, JOIN (Jurnal Online Informatika), 3 (2018), pp. 22–28.
- [9] H. FOURNIER, R. KOP, AND G. DURAND, *Challenges to research in moocs*, MERLOT Journal of Online Learning and Teaching, 10 (2014).
- [10] B. GAO, *Highly efficient english mooc teaching model based on frontline education analysis*, International Journal of Emerging Technologies in Learning (Online), 14 (2019), p. 138.
- [11] Z. HUANG, Q. LIU, C. ZHAI, Y. YIN, E. CHEN, W. GAO, AND G. HU, *Exploring multi-objective exercise recommendations in online education systems*, in Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 1261–1270.
- [12] X. HUIYING AND M. QIANG, *College english cross-cultural teaching based on cloud computing mooc platform and artificial intelligence*, Journal of Intelligent & Fuzzy Systems, 40 (2021), pp. 7335–7345.
- [13] S.-W. KIM, *Moocs in higher education*, Virtual learning, 1 (2016), pp. 1–17.
- [14] D. LIU, *Development and application of constructive english mooc system based on rbf algorithm*, in 2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), IEEE, 2023, pp. 1–6.
- [15] M. MELLATI AND M. KHADEMI, *Mooc-based educational program and interaction in distance education: Long life mode of teaching*, Interactive Learning Environments, 28 (2020), pp. 1022–1035.
- [16] M. PÉREZ-SANAGUSTÍN, I. HILLIGER, C. ALARIO-HOYOS, C. D. KLOOS, AND S. RAYAN, *H-mooc framework: reusing moocs for hybrid education*, Journal of Computing in Higher Education, 29 (2017), pp. 47–64.
- [17] H. SEBBAQ AND N.-E. EL FADDOULI, *Fine-tuned bert model for large scale and cognitive classification of moocs*, International Review of Research in Open and Distributed Learning, 23 (2022), pp. 170–190.
- [18] H. M. SHALATSKA, *The efficiency of moocs implementation in teaching english for professional purposes*, Information technologies and teaching aids, 66 (2018), pp. 186–196.
- [19] K. WANG AND C. ZHU, *Mooc-based flipped learning in higher education: students participation, experience and learning performance*, International Journal of Educational Technology in Higher Education, 16 (2019), pp. 1–18.
- [20] R. ZANG AND L. WANG, *Personalized teaching model of college english based on big data*, in Journal of Physics: Conference Series, vol. 1852, IOP Publishing, 2021, p. 022013.
- [21] L.-M. ZHANG, *Comprehensive evaluation model of mooc teaching quality of accounting major based on rete algorithm*, in e-Learning, e-Education, and Online Training: 7th EAI International Conference, eLEOT 2021, Xixiang, China, June 20–21, 2021, Proceedings Part II 7, Springer, 2021, pp. 379–391.
- [22] M. ZHANG AND L. ZHANG, *Cross-cultural o2o english teaching based on ai emotion recognition and neural network algorithm*, Journal of Intelligent & Fuzzy Systems, 40 (2021), pp. 7183–7194.
- [23] S.-H. ZHONG, Q.-B. ZHANG, Z.-P. LI, AND Y. LIU, *Motivations and challenges in moocs with eastern insights*, International Journal of Information and Education Technology, 6 (2016), p. 954.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Jan 6, 2024

Accepted: Feb 9, 2024



SMART FISH PASSAGE DESIGN AND APPLICATION OF HYDROACOUSTIC COMMUNICATION TECHNOLOGY IN AQUATIC ECOSYSTEM RESTORATION

CHAO YUE^{*}, MENGEN ZHU[†], LEI YANG[‡] AND LEI LI[§]

Abstract. The demand for creative solutions to help migrate fish and conservation efforts grows as aquatic environments experience more and more pressure from humans and fragmentation of habitat. The utilization of hydroacoustic technology for communication in conjunction with smart fish pathway architecture is the main emphasis of this research to improve rehabilitation efforts for aquatic environments. Using sophisticated systems for tracking and regulation to improve migratory pathways, the study investigates cutting-edge solutions in engineering for fish passage. Real-time data capture and transmission are made possible by the application of hydroacoustic technology for communication, which means that fish populations and monitoring systems can effectively communicate. The creation of intelligent fish passage structures with actuators, sensors, and communications components is a major focus of the research. The best passage efficiency of these structures is ensured by their dynamic adaptation to fish behavior and variables in the environment. An essential interface for gathering information on behavior, evaluating migratory trends, and putting adaptive management plans into practice is hydroacoustic communications technology. In order to assess the efficacy of the hydroacoustic communication technology and smart fish passage design in a variety of aquatic habitats, a thorough field investigation is part of the suggested methodology. To evaluate the effect on migration of fish rates of achievement, species diversity, and general well-being of the ecosystem, field data will be studied.

Key words: Marine Internet of Things; Internet of Underwater Things; protocols; smart fish, passage, hydroacoustic communications, aquatic ecosystem restoration

1. Introduction. In recent times, Earth observation, alterations in marine ecosystems, and changing climates have garnered human interest and have had a major effect on human productivity [8]. Underwater recognition of targets is currently a growing field of study due to the increasing demand for underwater identification. It has applications in the areas of ship noise categorization [7], underwater target localization and recognition [15], and aquatic environment surveying and demonstrating [14]. Underwater target recognition has new opportunities due to the rapid growth of artificial intelligence methods like machine learning and deep learning, the development of supercomputing, the substantial rise in math authority, and the rapid expansion of big-data-processing computation.

Researchers in this field are swiftly implementing research findings to improve technological iterations. Nevertheless, challenges remain in the application of deep learning for underwater target recognition, including small data amounts, limited flexibility of traditional visual system computations, complex pre-processing procedures, and deep learning patterns that are still too intricate to offer excellent generalizability. It is worth noting that acoustic and various signal-filtering techniques are useful for detecting pipeline leaks, and that deep learning algorithms are also widely employed, suggesting that models based on deep learning have excellent generality and swear in underwater acoustics [11].

To deploy IoUT, follow these three steps [24]. Developing a dynamic, ongoing, all-encompassing, and intelligent real-time view of the underwater world is the first stage. Large-scale, long-term, continuous oceanographic data collection has been made possible in recent decades by underwater sensor networks made up of a range of equipment, including conductivity, and heat and depth detectors, microbial sensors, and current meters [6]. Innovative uses that utilize human–robot interactions, such as undersea pipeline assessment, undersea volcanic

^{*}Hydropower and Water Conservancy Engineering Institute, POWERCHINA HUADONG Engineering Corporation Limited, Hangzhou, Zhejiang, 311122, China, email: chaoyueresad@outlook.com

[†]Dagu Hydropower Branch of Huadian Xizang Energy Co., Ltd., Shannan, Xizang, 856000, China

[‡]Hydropower and Water Conservancy Engineering Institute, POWERCHINA HUADONG Engineering Corporation Limited, Hangzhou, Zhejiang, 311122, China

[§]Dagu Hydropower Branch of Huadian Xizang Energy Co., Ltd., Shannan, Xizang, 856000, China

activity and hydrothermal source research, seabed visualisation, strategic surveillance, and underwater rescue, are driving up demand for real-time multimedia data [22].

Large-scale real-time underwater data transmission is the second phase of the IoUT deployment. Undersea gliders, remotely controlled unmanned underwater vehicles (ROVs), and automated unmanned underwater vehicles (UUVs) are examples of movable systems that have made it possible to build mobile underwater networks and are essential for high-quality surveillance footage [26, 27, 17, 19]. The smart analysis of large amounts of underwater data is the third step in the deployment of IoUT. The amount of maritime data that was acquired in the past was limited because of a lack of equipment and minimal investment, which caused the process to take years or months.

The motivation behind this research stems from the ever-increasing need for innovative solutions to address the challenges faced by aquatic environments due to human activities and habitat fragmentation. As human impact on aquatic ecosystems intensifies, it becomes imperative to find creative and effective ways to support fish migration and conservation efforts. This research is driven by the pressing demand to enhance the rehabilitation of aquatic environments, offering a lifeline to struggling fish populations.

This study focuses on a novel approach that combines hydroacoustic technology with smart fish pathway architecture to revolutionize the way we facilitate fish migration and conservation. By harnessing cutting-edge engineering solutions, including real-time data capture and transmission through hydroacoustic technology, we enable seamless communication between fish populations and monitoring systems. The novelty lies in the creation of intelligent fish passage structures equipped with advanced actuators, sensors, and communication components, allowing them to adapt to fish behavior and changing environmental conditions dynamically.

The main contribution of the proposed method is given below:

1. Using hydroacoustic data, a DNN-LSTM model is trained to identify complex temporal correlations in fish movements.
2. The trained model is then included into the smart fish passageways control system, enabling ongoing adaptation in response to change fish behavior and environmental variables.
3. The goal of the study is to show how DNN-LSTM can be used to provide fish passage systems with a smart, self-learning framework that minimizes ecological disturbance while maintaining efficient passage.

Remaining sections of this paper are structured as follows: Section 2 discusses about the related research works, Section 3 describes the Smart fish passage design, application of hydroacoustic communication technology and Deep Neural Networks, Section 4 discusses about the experimented results and comparison and Section 6 concludes the proposed optimization method with future work.

2. Related Works. The future's top innovations for implementing smart data processing on massive scales will be big data, cloud computing, artificial intelligence, and virtual reality [18]. Therefore, one of the key areas of research for upcoming applications of human-robot interaction is the real-time changing and visual tracking of underwater landscapes. Most commercial marine IoT applications concentrate on both surface and subsea IoT technology to measure and monitor business activity. It is possible to identify and isolate the IoUT market sector for usage in aquaculture and fishing, that works with fish breeding in small spaces [5].

In response, the central server analyzes the information and creates management and choice-making strategies. Large-scale fish and marine creature searches and analyses in open waterbodies comprise another set of tasks [28, 23, 12, 2]. It should be noted that the notion of Internet of Things Ocean (IoTO) [10] or Ocean of Things is mentioned in the literature in addition to Internet of Things (IoUT). One way to conceptualize the Internet of Things (IoT) technology is as an intelligent underwater item network. One potential technique for the organized administration of various marine data types is IoTO.

The International Maritime Organization (IMO) first established the concept of "e-Navigation" [13, 21] to enable different kinds of navigation services, which is akin to Maritime IoT and was intended to improve the shipping sector. Other services were subsequently added; these are mentioned above. Thus, maritime IoT refers to a dispersed hardware–software complex that allows for transmitted from different above-water ocean engineering structures and items gadgets, used via a unified machine-type communication via a data system (typically the Internet, or a marine information network, or a network of underwater services).

The wavelet transform approach is not without flaws, though. The shortcomings of various discrete wavelet

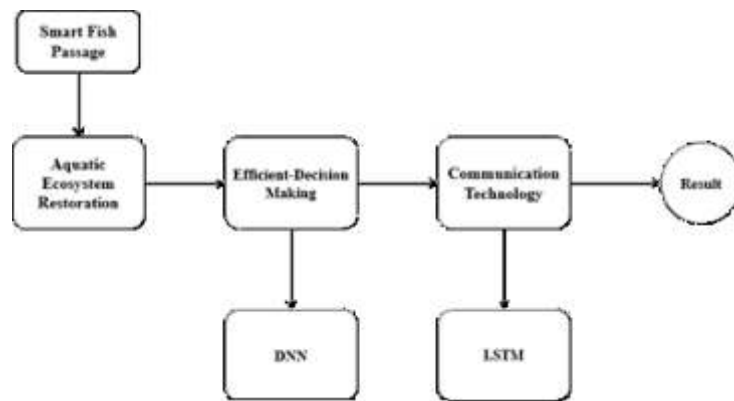


Fig. 3.1: Architecture of proposed method

transform (DWT) system designs were noted by author in [3, 9]. The outstanding characteristics of the wavelet transform (WT) in one dimension are not transferable to two dimensions or higher, and it shows a lack of adaptivity to additional modality disintegration techniques, such as EMD, LMD, VMD, SGMD, etc. The one-dimensional characteristic vector obtained by the wavelet transform is frequently not enough to provide optimal features because of the intricate nature of the hydroacoustic surroundings [1]; therefore, finding a way to enhance the selection of multidimensional or optimal features has emerged as a potential area of research.

To reclaim the breakpoints in the LOFAR spectrum and achieve an exceptional identification rate in the CNN network, the author [25] developed a multi-step decision-algorithm-based improvement technique based on LOFAR spectrum improvement for underwater detection of targets. After implementing these spectrograms into the AlexNet network, the researchers [20, 16, 4] analyzed typical spectrum maps, such as LOFAR, Audio, The demon, a histogram etc. and discovered that the LOFAR spectra had the best identification rate.

The wavelet transform approach, particularly the discrete wavelet transform (DWT) system designs, has been found to have shortcomings in handling multidimensional data. While it may perform well in one dimension, it lacks adaptivity when applied to two or more dimensions. This limitation hinders its effectiveness in dealing with the complex nature of hydroacoustic environments, where multidimensional data is often encountered. The one-dimensional characteristic vector obtained from the wavelet transform may not provide optimal features for underwater target detection due to the intricate nature of hydroacoustic surroundings. This inadequacy highlights the need for improved techniques to select multidimensional or optimal features that can better capture the nuances of the underwater environment.

3. Proposed Methodology. The proposed methodology uses deep learning method for Smart fish passage design and application of hydroacoustic communication technology in aquatic ecosystem restoration. It uses Deep Neural Networks based Long-Short Term Memory (LSTM) for smart fish passages and intelligent decision-making. Field tests in various aquatic habitats will be carried out to evaluate the effectiveness of the smart fish tunnels augmented by DNN-LSTM. The general effect on the recovery of aquatic ecosystems, species-specific adaptation, and migration success rates are examples of key performance indicators. The goal of the study is to show how DNN-LSTM can be used to provide fish passage systems with an intelligent, self-learning architecture that minimizes ecological disturbance while maintaining efficient passage. In figure 3.1 shows the architecture of proposed methodology.

3.1. Smart Fish Passage and hydroacoustic communication for Aquatic Restoration. With the overall objective of fostering successful aquatic ecosystem restoration, this research focuses on the integration of Deep Neural Networks (DNN) and Long Short-Term Memory (LSTM) networks to create an intelligent system for smart fish passage design and hydroacoustic communication. Through real-time, data-driven communication between fish populations and fish passage structures, the study hopes to improve the responsiveness and flexibility of fish passage structures by utilizing the temporal learning capabilities of DNN-LSTM architectures.

The suggested approach uses hydroacoustic data to train a DNN-LSTM model, which then uses the data to capture and evaluate the complex temporal dynamics of fish movement patterns. The trained model is the central intelligence element of smart fish passageways, allowing for dynamic modifications based on the hydroacoustic monitoring system's real-time data. At the same time, a bidirectional communication channel is established between the fish community and the smart channels using hydroacoustic communication technology.

A variety of aquatic conditions will be used for field testing to assess how well the DNN-LSTM-based smart fish passage system works. To verify the efficacy of the suggested strategy, important indicators like ecological impact, species-specific adaptation, and migration success rates will be evaluated. The goal of the study is to show how DNN-LSTM can be an effective tool for developing smart, self-learning fish passage solutions that maximize fish migration and ecosystem restoration initiatives.

The results obtained from this study have wider ramifications for the development of technology-based conservation tactics. The suggested framework demonstrates a comprehensive strategy for improving aquatic ecosystems by fusing deep learning methods with hydroacoustic communication, tackling the problems brought on by habitat destruction and dispersion. The findings highlight the possibility for innovative technologies to play a critical role in the ecological restoration of aquatic ecosystems and add to the expanding field of intelligent ecological management and monitoring.

3.2. Deep Neural Networks (DNN). Artificial neural networks that analyze data using numerous layers—hence the name "deep"—are known as deep neural networks (DNNs). Nodes, sometimes referred to as neurons or units, are present in every layer of the network and are connected by weighted connections. Deep neural networks (DNNs) are a subclass of machine learning models that fall within the deep learning category.

To enable them to learn hierarchical data representations, DNNs usually consist of numerous hidden layers positioned among the input and output layers. The network can catch intricate patterns and characteristics because of its depth. Backpropagation is an optimization algorithm used in DNN training. Through training, the network attempts to reduce the discrepancy between target values and projected outputs by adjusting the weights of its connections.

Activation functions are applied by nodes in each layer to the weighted total of their inputs. Sigmoid, hyperbolic tangent (tanh), and rectified linear unit (ReLU) are examples of common activation functions. By adding non-linearity to the system, these functions enable it to learn intricate mappings. Several deep learning structures, including TensorFlow, PyTorch, Keras, and others, are used to implement DNNs. Deep learning model construction, training, and deployment tools are offered by these platforms.

Natural language processing, recommendation systems, picture and speech recognition, and other industries have shown impressive performance using DNNs. When given tasks that require a lot of data to be trained, they perform exceptionally well. Large volumes of labeled data are needed for the computationally demanding process of training deep neural networks. Another frequent issue which must be handled is overfitting, which occurs when a model learns noise in the data used for training rather than the real patterns.

3.3. Long-Short Term Memory (LSTM). One kind of recurrent neural network (RNN) architecture called Long Short-Term Memory (LSTM) was created to solve the gradient that diminishes issue, which is a prevalent problem with conventional RNNs. LSTMs are especially useful for problems requiring time-series data, processing natural languages, and sequential pattern identification since they were developed to recognize dependencies that persist in sequential data.

The capacity of LSTMs to preserve cells with memories which can store and retrieving information across extended sequences is essential for avoiding the loss of pertinent data during training. This is made possible by a group of gates that control information flow into, out of, and inside the memory cell. These gates include an input gate, an output gate, and a forget gate.

Cell State (Ct): The long-term information storage cell in the memory.

The output generated by the LSTM at a specific time step is known as the Hidden State (Ht).

How much of the new data should be saved in the memory cell is decided by the input gate (i).

The Forget Gate (f) determines the amount of data that should be removed from the memory cell. The output gate (o) controls the amount of memory cell content that is utilized to produce the output.

Long-term dependency capture is a key component of LSTMs' performance in a variety of uses, such as time-series prediction, translation by machine, and speech recognition. An LSTM could be used to simulate

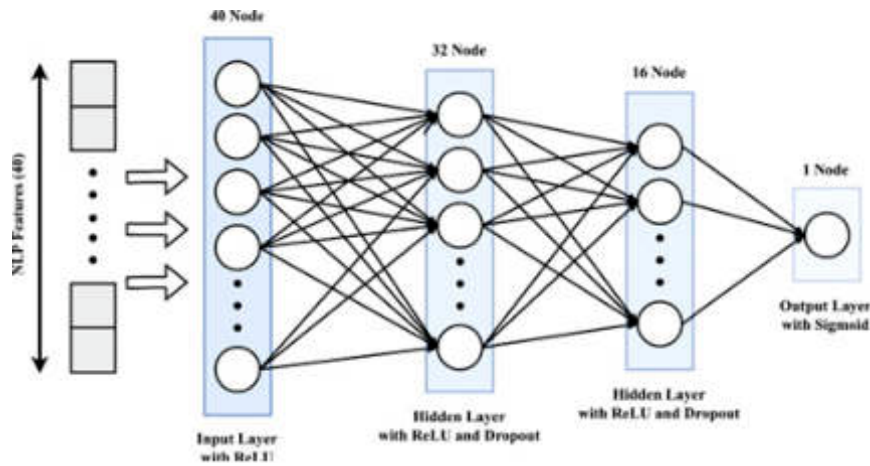


Fig. 3.2: Structure of DNN-LSTM

and foresee the actions of fish over time in the context of your mentioned study on hydroacoustic communication technology and smart fish passage design. This would allow the smart travel structures to change their configuration based on the evolving patterns seen in the hydroacoustic data. In figure 3.2 shows the structure of DNN-LSTM.

Hydroacoustic technology allows for non-intrusive monitoring of fish populations without physically disturbing their habitat. This is in contrast to some traditional methods like electrofishing, which can be invasive. It can cover large areas, making them suitable for monitoring fish migration in expansive aquatic environments such as rivers, lakes, and oceans. This wide coverage is often difficult to achieve with manual methods. They provide real-time data, allowing researchers to track fish movements and behaviors as they occur. This immediacy can be crucial for making timely management decisions.

Hydroacoustic systems can distinguish between different fish species based on their acoustic signatures. This capability is valuable for studying specific species' migration patterns and behaviours. Compared to physical interventions like fish ladders or traps, hydroacoustic technology typically has a lower ecological impact since it does not require altering the natural flow of water or physical structures.

4. Result Analysis. Although short-time Fourier, Meier, Hilbert-Yellow, and additional methods of processing have been put forth to address some aspects of indicate extraction of features, single signal processing for feature extraction is no longer able to increase the classifier's effectiveness due to the flaws in the various algorithms. Therefore, one path for the advancement of hydroacoustic signal detection will be multi-spectrum feature fusion. The dataset is taken from Kaagle for evaluation.

The evaluation parameters such as accuracy, precision, recall and f1-score is measured. The proposed method achieves better result in all parameter metrics.

The simulation's accuracy, which is expressed as follows in Equation (4.1), indicates how effectively the model works across classes.

$$Accuracy = \frac{Total\ number\ of\ truly\ classified\ samples}{Total\ Samples} \tag{4.1}$$

The precision of the simulations is an assessment of their capacity to detect true positives, and it is computed using Equation (4.2).

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

The proportion of projected true positive and false negative values to true positive prediction values is

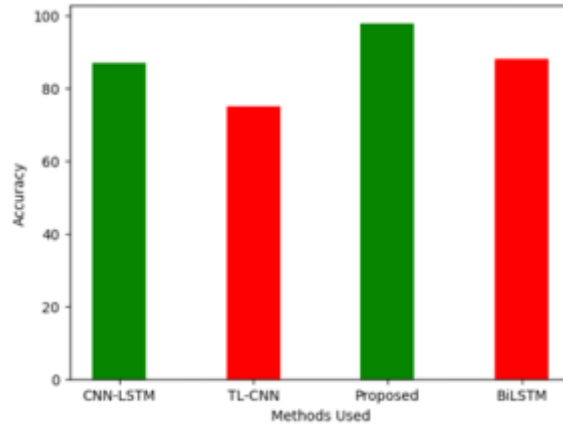


Fig. 4.1: Accuracy

known as the recall. Equation (4.3) represents the calculation.

$$Recall = \frac{TP}{TP + FP} \quad (4.3)$$

The model's total accuracy, or F1 score, strikes a positive class balance between recall and precision. Equation (4.4), which represents the calculation, is used.

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4.4)$$

This has to do with how well models (like DNN-LSTM) forecast fish behavior from hydroacoustic data. The capacity of the model to precisely predict fish behaviors and movement patterns is crucial for creating intelligent fish passage systems that work well.

Analyse the degree to which the intelligent fish passage systems can adjust in real time to changes in fish behavior and surrounding circumstances. This entails evaluating how accurately the passages alter their configurations in response to the integrated models' predictions.

Analyse how accurate the overall influence on the restoration of aquatic ecosystems is. Assessing alterations in diversity of species, health of ecosystems, and other pertinent indicators of ecology is part of this. In this case, accuracy refers to how successfully the technologies being used support the recovery of the ecosystem.

Evaluate the precision of data transfer between the smart passageways and the hydroacoustic communication technology. This entails assessing the accuracy and dependability of the communication links used to send information on fish behavior and system modifications. In figure 4.1 shows the evaluation of accuracy.

In the setting of Fish Passage Design and Hydroacoustic Communication, as well as recall relates to the system's capacity to accurately detect and meet the needs of migrating fish. It concerns the percentage of real positive cases (fish passes that are successful or pertinent hydroacoustic signals) that the equipment accurately detects. Regarding Fish Passage Designs, a high recall rate suggests that a considerable proportion of the fish population can move via the structures without any hindrance. Ensuring that the planned target species are accommodated, and their migration is facilitated by the constructed routes is crucial for the successful restoration of aquatic ecosystems.

To summarize, a high recall in hydroacoustic communication means that the system correctly recognizes and interprets relevant underwater signals, which contributes to a greater awareness of fish behavior to enhance passage design and overall aquatic restoration efforts. Conversely, a high recall in fish passage design means that the structures in question are effectively facilitating fish migration. In figure 4.3 shows the evaluation of recall.

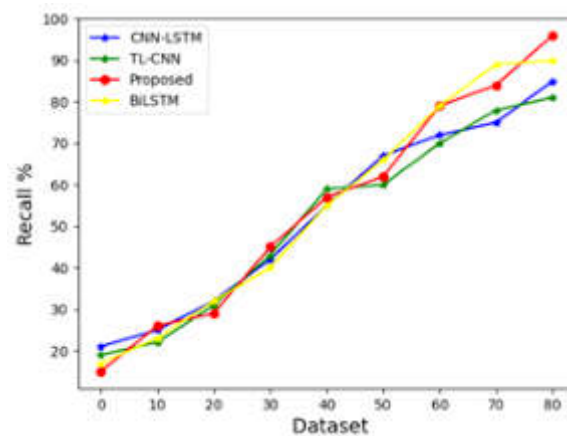


Fig. 4.2: Recall

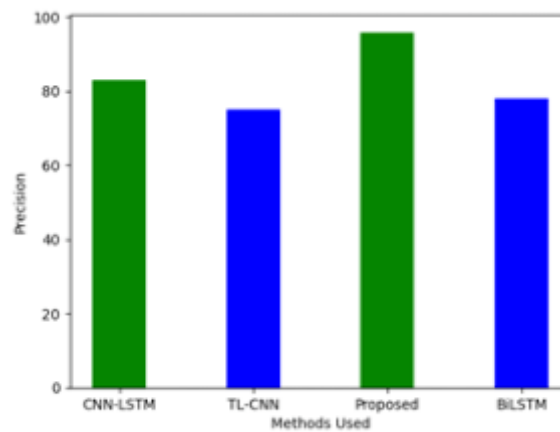


Fig. 4.3: Precision

Recall in Hydroacoustic Communication is the system's capacity to identify and decipher pertinent signals in the submerged acoustic environment. This includes recognizing hydroacoustic cues, such as movement patterns or communication signals, accurately in fish. To collect thorough and precise data on fish behavior—data that can later be utilized to influence the adaptive characteristics of smart fish passages—high recall in hydroacoustic communication is necessary.

When discussing Fish Passage Design and Hydroacoustic Communication, as well as precision pertains to the precision and dependability of the technologies and systems that support fish migration and interaction in aquatic settings. It is imperative to guarantee that the solutions put into practice accurately and successfully tackle the problems related to fish movement and hydroacoustic communications. Fish tunnels must be precisely designed so that the structures can adjust to changing environmental conditions and the unique behaviors of various fish species.

The precise design of the passage structures guarantees the smooth and effective passage of fish, reducing the amount of stress and energy that the aquatic species must expend. Elements that preferentially aid the movement of fish species while discouraging non-target species can be included with precision in design to improve the balance of nature. In figure 4.3 shows the evaluation of precision.

In binary classification problems, the F1-score—also referred to as the F1 measure or F1-value—is a statistic

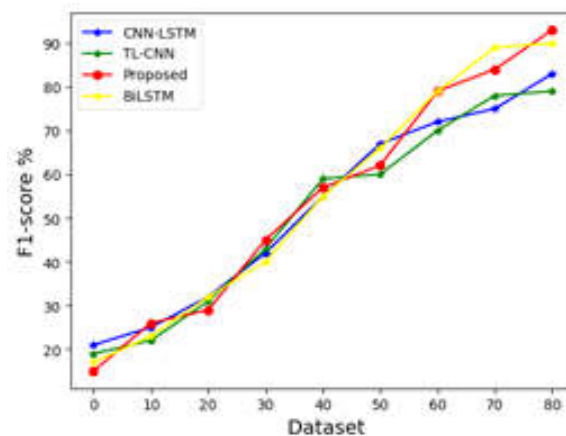


Fig. 4.4: F1-score

that is frequently employed. Recall and precision are used to give a fair assessment of a model's effectiveness. The F1-score could be used to assess how well the system recognizes successful fish passage events in the context of fish route design and hydroacoustic communications. A high F1-score suggests that the algorithm is successfully recognizing effective fish passages while avoiding false positives and false negatives in the setting of fish passage design and hydroacoustic communications. This statistic offers a thorough evaluation of the model's capacity to recognize and react to fish movements inside the intended passageways, accounting for the intricacies of aquatic communications and surroundings. In figure 4.4 shows the evaluation of F1-score.

5. Conclusion. As human pressure on aquatic ecosystems increases and habitat becomes more fragmented, there is an increasing need for innovative solutions to support fish migration and conservation efforts. The focus of this project is to improve rehabilitation efforts for aquatic ecosystems by using smart fish pathway architecture in conjunction with hydroacoustic technologies for communication. The project explores state-of-the-art engineering solutions for fish passage by utilizing complex tracking and regulating systems to enhance migratory paths. Fish populations and monitoring systems can efficiently communicate thanks to the adoption of hydroacoustic technology for communication, which enables real-time data capture and transmission. One main goal of the project is to create intelligent fish passage structures that include actuators, sensors, and communications components. These structures' dynamic response to fish behavior and environmental factors ensures optimal passage efficiency. Hydroacoustic communications technology is a vital interface for behavior data collection, migration trend assessment, and the implementation of adaptive management strategies. A comprehensive field study is part of the recommended methodology to evaluate the effectiveness of the hydroacoustic communication technology and smart fish passage design in a range of aquatic settings. Field data will be examined to assess the impact on fish migration rates of accomplishment, species variety, and overall ecosystem health. The future research directions can contribute to the ongoing development and refinement of hydroacoustic technology for fish migration and monitoring, ultimately advancing the field of aquatic conservation and sustainable management.

Acknowledgement. This work was sponsored in part by National Natural Science Foundation of China (2345678)

REFERENCES

- [1] M. F. ALI, D. N. K. JAYAKODY, Y. A. CHURSIN, S. AFFES, AND S. DMITRY, *Recent advances and future directions on underwater wireless communications*, Archives of Computational Methods in Engineering, 27 (2020), pp. 1379–1412.
- [2] M. F. ALI, D. N. K. JAYAKODY, AND Y. LI, *Recent trends in underwater visible light communication (uvlc) systems*, IEEE Access, 10 (2022), pp. 22169–22225.

- [3] S. ANEES, S. R. BARUAH, AND P. SARMA, *Hybrid rf-fso system cascaded with uwoc link*, International Journal of Innovative Technology and Exploring Engineering, 8 (2019), pp. 2278–3075.
- [4] H. S. DOL, P. CASARI, T. VAN DER ZWAN, AND R. OTNES, *Software-defined underwater acoustic modems: Historical review and the nilus approach*, IEEE Journal of Oceanic Engineering, 42 (2016), pp. 722–737.
- [5] A. A. A. EL-BANNA AND K. WU, *Machine learning modeling for IoUT networks: Internet of underwater things*, Springer Nature, 2021.
- [6] I. GOODFELLOW, J. POUGET-ABADIE, M. MIRZA, B. XU, D. WARDE-FARLEY, S. OZAIR, A. COURVILLE, AND Y. BENGIO, *Generative adversarial networks*, Communications of the ACM, 63 (2020), pp. 139–144.
- [7] T. GUO, Y. SONG, Z. KONG, E. LIM, M. LÓPEZ-BENÍTEZ, F. MA, AND L. YU, *Underwater target detection and localization with feature map and cnn-based classification*, in 2022 4th International Conference on Advances in Computer Technology, Information Science and Communications (CTISC), IEEE, 2022, pp. 1–8.
- [8] A. KABANOV AND V. KRAMAR, *Marine internet of things platforms for interoperability of marine robotic agents: An overview of concepts and architectures*, Journal of Marine Science and Engineering, 10 (2022), p. 1279.
- [9] A. KATARIA, S. GHOSH, V. KARAR, T. GUPTA, K. SRINIVASAN, AND Y.-C. HU, *Improved diver communication system by combining optical and electromagnetic trackers*, Sensors, 20 (2020), p. 5084.
- [10] K. KEBKAL, A. KEBKAL, E. GLUSHKO, V. KEBKAL, L. SEBASTIÃO, A. PASCOAL, J. RIBEIRO, H. SILVA, M. RIBEIRO, AND G. INDIVERI, *Underwater acoustic modems with synchronous chip-scale atomic clocks for scalable tasks of auw underwater positioning*, Gyroscopy and Navigation, 10 (2019), pp. 313–321.
- [11] M. KHISHE AND M. R. MOSAVI, *Chimp optimization algorithm*, Expert systems with applications, 149 (2020), p. 113338.
- [12] S. KUMARA AND C. VATSB, *Underwater communication: A detailed review*, in CEUR Workshop Proceedings, 2021.
- [13] I. LEBLOND, S. TAUVRY, AND M. PINTO, *Sonar image registration for swarm auws navigation: Results from swarms project*, Journal of Computational Science, 36 (2019), p. 101021.
- [14] X. LIN, R. DONG, AND Z. LV, *Deep learning-based classification of raw hydroacoustic signal: A review*, Journal of Marine Science and Engineering, 11 (2022), p. 3.
- [15] Y. LIU, H. CHEN, AND B. WANG, *Doa estimation based on cnn for underwater acoustic array*, Applied Acoustics, 172 (2021), p. 107594.
- [16] G. QUINTANA-DÍAZ, P. MENA-RODRÍGUEZ, I. PÉREZ-ÁLVAREZ, E. JIMÉNEZ, B.-P. DORTA-NARANJO, S. ZAZO, M. PÉREZ, E. QUEVEDO, L. CARDONA, AND J. J. HERNÁNDEZ, *Underwater electromagnetic sensor networkspart i: Link characterization*, Sensors, 17 (2017), p. 189.
- [17] N. SAEED, A. CELIK, T. Y. AL-NAFFOURI, AND M.-S. ALOUINI, *Underwater optical wireless communications, networking, and localization: A survey*, Ad Hoc Networks, 94 (2019), p. 101935.
- [18] G. SCHIRRIPIA SPAGNOLO, L. COZZELLA, AND F. LECCESE, *Underwater optical wireless communications: Overview*, Sensors, 20 (2020), p. 2261.
- [19] B. SHIHADA, O. AMIN, C. BAINBRIDGE, S. JARDAK, O. ALKHAZRAGI, T. K. NG, B. OOI, M. BERUMEN, AND M.-S. ALOUINI, *Aqua-fi: Delivering internet underwater using wireless optical networks*, IEEE Communications Magazine, 58 (2020), pp. 84–89.
- [20] M. TAHIR, I. ALI, P. YAN, M. R. JAFRI, J. ZEXIN, AND D. XIAOQIANG, *Exploiting w. ellison model for seawater communication at gigahertz frequencies based on world ocean atlas data*, ETRI Journal, 42 (2020), pp. 575–584.
- [21] J. WANG, J. SHEN, W. SHI, G. QIAO, S. WU, AND X. WANG, *A novel energy-efficient contention-based mac protocol used for oa-uwsn*, Sensors, 19 (2019), p. 183.
- [22] M. M. WANG AND J. ZHANG, *Machine-Type Communication for Maritime Internet-of-Things*, Springer, 2021.
- [23] J. WOSOWEI AND C. SHASTRY, *Underwater wireless sensor networks: Applications and challenges in offshore operations*, Int. J. Curr. Adv. Res, 10 (2021), pp. 23729–23733.
- [24] H. WU, Q. SONG, AND G. JIN, *Underwater acoustic signal analysis: Preprocessing and classification by deep learning.*, Neural Network World, (2020).
- [25] X. WU, W. XUE, AND X. SHU, *Design and implementation of underwater wireless electromagnetic communication system*, in AIP Conference Proceedings, vol. 1864, AIP Publishing LLC, 2017, p. 020022.
- [26] T. XIA, M. M. WANG, J. ZHANG, AND L. WANG, *Maritime internet of things: Challenges and solutions*, IEEE Wireless Communications, 27 (2020), pp. 188–196.
- [27] T. YANG, J. CHEN, AND N. ZHANG, *Ai-empowered maritime internet of things: A parallel-network-driven approach*, IEEE Network, 34 (2020), pp. 54–59.
- [28] Q. ZHAO, Z. PENG, AND X. HONG, *A named data networking architecture implementation to internet of underwater things*, in Proceedings of the 14th International Conference on Underwater Networks & Systems, 2019, pp. 1–8.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Jan 6, 2024

Accepted: Feb 9, 2024



DESIGN OF FINANCIAL DATA ANALYSIS AND DECISION SUPPORT SYSTEM BASED ON BIG DATA

SUFANG ZHENG*

Abstract. A cutting-edge Decision Support System (DSS) utilizing Deep Reinforcement Learning (DRL) for improved financial data analysis is the primary focus of the proposed research. In light of the prospering difficulties presented by large information in the monetary space, our creative methodology outfits the force of DRL to foster a powerful and versatile framework. By flawlessly incorporating DRL into the DSS structure, we mean to improve the framework's capacity to break down huge and complex monetary datasets. This DSS not only provides financial professionals with intelligent decision-making support but also real-time insights into market trends and patterns. The collaboration between enormous information investigation and DRL works with a dynamic and responsive framework equipped for adjusting to the quickly developing financial scene. Our exploration adds to the headway of choice by tending to the particular requests of monetary information, consequently enabling clients with ideal and informed dynamic abilities. The proposed DRL-based DSS addresses a change in perspective in monetary information examination, offering a versatile and effective answer for exploring the intricacies of enormous information in the financial area. This examination holds huge potential for changing dynamic cycles, advancing monetary security, and at last adding to the progression of the more extensive monetary industry.

Key words: DRL, decision support system, financial data analysis, big data, intelligent decision making, adaptive technology

1. Introduction. In the current financial landscape, intelligent decision support systems are in high demand. These frameworks act as priceless instruments for monetary experts, offering experiences, examination, and dynamic help [4]. Regardless, standard financial decisions and genuinely strong organizations experience basic limitations in really managing the intricacies of present day money related data. Continuous handling of immense datasets, adjusting to quickly moving economic situations, and removing helpful data from multi-faceted monetary examples are snags [6, 7]. Because of these limitations, standard frameworks are unable to be deft and responsive, which results in subpar dynamic results and hampers the ability of financial experts to investigate the powerful concept of the financial world [12].

As we analyse the continuous creative part, existing solutions for money related decisions assist with showing drawbacks that warrant thought. Conventional progressions fight to keep awake with the surprising improvement of colossal data in the financial region, habitually provoking lethargy issues and compromised logical accuracy [8, 9]. The weaknesses to these advancements feature the pressing precondition for novel systems that can beat the challenges presented by the consistently expanding volume and intricacy of monetary information. Significant learning emerges as a promising street to address the lack of ordinary money-related decisions and sincerely strong organisations [10]. Its ability to subsequently acquire depictions from data benefits from more exact and nuanced examination. Monetary choice emotionally supportive networks can possibly turn out to be stronger and versatile thanks to profound learning procedures' capacity to remove includes and perceive designs. This impact in context opens approaches to the extra present day and strong procedures for unravelling complex money related data, empowering prevalent powerful cycles.

By clearly analysing the limitations and restrictions of the existing techniques here we proposed the novel approach of Deep Reinforcement Learning based Decision Support System (DRL-DSS) [1, 11, 5]. This combination means bridling the qualities of both profound learning and support, figuring out how to make a framework that not only explores the difficulties of large amounts of monetary information but also adjusts progressively to changing economic situations. By coordinating DRL into the choice help structure, our exploration looks to provide a versatile, responsive, and smart arrangement, engaging monetary experts with upgraded dynamic capacities notwithstanding the developing monetary scene.

*School of Business, Zhengzhou University of Industrial Technology, Xinzheng, 451100, China, (sufangzhengfinan@outlook.com)

The motivation behind our research on the "Design of a Financial Data Analysis and Decision Support System Based on Big Data" stems from the increasingly complex and voluminous nature of financial data in the global economy. The explosion of big data in the financial sector presents both a significant challenge and a golden opportunity for financial institutions. Traditional economic analysis tools and methodologies are becoming inadequate to process, analyse, and derive meaningful insights from this vast amount of data efficiently. As a result, there is a pressing need for innovative solutions that can handle the complexity and scale of financial data while providing actionable insights to support decision-making processes.

Integrating Deep Reinforcement Learning (DRL) with a Decision Support System (DSS) represents a pioneering approach to tackling the challenges posed by big data in finance. DRL, with its ability to learn optimal actions through trial and error by interacting with a dynamic environment, offers a powerful tool for analysing financial datasets that are large, complex, and constantly changing. By leveraging DRL, our proposed DSS aims to not only process and analyze big financial datasets more efficiently but also adapt to new data and market conditions in real-time, providing financial professionals with timely and relevant insights.

The main contribution of the paper as follows

1. Proposed a novel approach of Deep Reinforcement Learning Decision Support System (DRL-DSS) that significantly enhances the decision support capabilities in the financial domain.
2. This research contributes to addressing the challenges posed by big data in financial systems.
3. The proposed DRL-DSS includes a DRL-DQN-based technique in the decision support framework; this intelligent technique enables the system to learn and adapt to market conditions autonomously and offers a better path in financial decision-making that goes beyond the capabilities of traditional systems.
4. This proposed efficacy is proved with valid experiments.

2. Related Work.

2.1. Decision support system based discussions. This study [14] investigates how big data in financial decision-making affects information asymmetry, principal-agent relationships, and risk management. It examines how big data could enhance predictions, increase the relevance of decisions, provide companies a competitive edge, and promote flexible decision-making. The research emphasises the value of big data in merging business and finance and the necessity of a robust information infrastructure for this type of integration through examination of practical instances of its application in corporate financial decisions. According to the study, big data is crucial for removing obstacles between business and finance, expediting the decision-making process, and raising company value. This study [5] highlights flaws, including inefficiency and intelligence deficiency, in the present financial decision support systems, and investigates how artificial intelligence might be integrated to improve them. Using the X business as a real-world example, it suggests cutting-edge, clever computerised technology to help with financial decision-making. Based on a questionnaire survey, the results demonstrate that the AI-enhanced system offers higher intelligence, timeliness, and accuracy in financial decision-making while lowering costs and simplifying the integration of management and financial accounting.

The goal [3] of this study is to help businesses make better financial decisions by utilizing the current data boom. It emphasizes how crucial timely data analysis and intelligent systems are to the efficient management of resources, time, and choices in order to maximize profitability. The article highlights the applicability of edge computing and criticizes conventional data management in businesses as being out of date. It then suggests an information-based financial management system. The system meets performance metrics requirements with high efficiency, responsiveness, and CPU usage after extensive testing and modifications. The whale algorithm's integration also demonstrates excellent energy and computational resource management. The system's potential to enhance enterprise financial management is highlighted in the study's conclusion, which also calls for more improvement.

This study [16] shows how big data may improve financial analysis and decision-making in businesses, replacing more conventional approaches that depend on human resources. The study highlights notable enhancements in decision-making accuracy and efficiency through the use of an intelligent financial decision support system that incorporates big data web crawler technology and ETL procedures, as demonstrated in a case study involving J Group. The system's ability to analyse data in real time and provide financial insights represents a significant advancement in corporate financial management and decision-making. This work [15] fills a need in basic medical care, particularly in distant areas, by creating a ground-breaking artificial intelligence system that

can speak with patients on its own using voice recognition and synthesis. The system works as a virtual doctor. The AI system's promise to provide precise probability forecasts for patient diagnosis is demonstrated by its ability to foresee type 2 diabetes mellitus utilizing non-invasive sensors and deep neural networks. The study also examines the level of acceptance of artificial intelligence among young people in the healthcare industry, including details on the possible long-term effects of this kind of technology. Artificial intelligence has been used more and more in the legal industry since the 1980s to handle an increase in the number of cases. This paper introduces "TaSbeeb," a deep learning-based JDSS that can retrieve religious texts and judicial reasoning for use in Saudi courts. It is divided into three stages: using stacked DL models to handle unbalanced classification, semi-automated judicial text annotation, and a judicial language model for information retrieval. With its excellent accuracy and F-scores, TaSbeeb represents a substantial improvement in the efficiency and accuracy of Arabic JDSS and holds the potential for wider implementation in judicial systems.

3. Methodology.

3.1. Proposed Overview. The suggested idea for the DRL-DSS coordinates both DRL-DQN and TRPO calculations to break down and settle on choices in the complex and information-serious field of money is shown in Figure 3.1. At first, the framework gathers an exhaustive cluster of monetary information, including stocks, securities, market lists, macroeconomic pointers, and ongoing worldwide monetary news. This information is carefully preprocessed to guarantee quality and importance, utilizing procedures like standardization, sound decrease, and element choice. The DRL-DQN algorithm is used in the first phase. This includes the utilization of a profound Q-organization that is prepared to comprehend and expect market patterns and ways of behaving. The DQN is intended to deal with the tremendous time-series information, perceiving examples and gaining from verifiable patterns to make informed expectations about future market developments. This organization is advanced for security and effectiveness, utilizing procedures, for example, experience replay and fixed Q-focuses, to improve learning. Parallely, the TRPO calculation is used to advance the dynamic approach of the framework. TRPO, known for its viability in overseeing strategy slope techniques in support learning, offers a more steady and hearty way to deal with learning strategies. It guarantees that the updates to the strategy are made inside a trust locale, forestalling uncommon arrangement changes that could prompt shaky preparation or lacklustre showing. This is especially vital in the unstable and variable monetary business sectors.

The framework orchestrates the forecasts and experiences acquired from the DRL-DQN with the approach choices made through TRPO. DQN provides deep insight into the data, and TRPO ensures that the decision-making process is continuously refined and optimized for the best possible outcomes. To address the difficulties of enormous information in finance, high level information, the board, and dimensionality decrease procedures are carried out. These procedures help in refining the most effective data from the tremendous datasets, guaranteeing that the DRL-DQN and TRPO calculations are working with the most applicable and huge information highlights. Far-reaching back-testing and forward-testing instruments structure a vital piece of the procedure. The situation is thoroughly tried against verifiable information and simulated conditions to approve its viability and flexibility in various economic situations. The DRL-DSS's robustness and dependability in real-world financial scenarios are assured by this extensive testing. At long last, an accentuation is put on making an interpretable and easy-to-understand interface for the DRL-DSS. Users are able to make well-informed decisions based on the system's insights and recommendations as a result of this, guaranteeing that the intricate workings of the DRL-DQN and TRPO algorithms are presented in an approachable manner. This approach expects to make the DRL-DSS a strong and natural device for monetary examination and dynamics in the space of enormous information.

3.2. Proposed DRL-DSS Approach in Financial big data.

3.2.1. DRL-DQN based Financial Decision Making. DQN is a useful tool for making well-informed judgments in the financial domain when used in conjunction with a DRL-DSS. To control the current status of the market it analyses a lot of financial data, including stock prices and market movements. Then, by assessing the anticipated advantages of different financial transactions, like purchasing or selling stocks, it makes predictions about their likelihood to be successful. The DQN regularly analyzes market behaviour to refine its approach to making decisions by identifying what works and what doesn't. Because of this, it is a

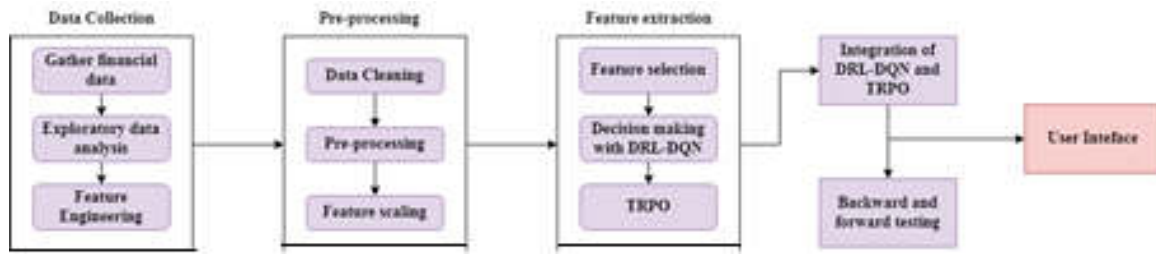


Fig. 3.1: Proposed DRL-DSS Architecture

priceless tool for financial trading investment advice and automated decision-making, assisting in maximizing returns and lowering risks. Some of the DRL-based decision support systems are discussed in the study [2, 13].

The system is designed to collect only the data necessary for the specific purposes for which it is processed, adhering to the GDPR principles of data minimization and purpose limitation. It incorporates robust consent management mechanisms to ensure that data is collected and processed only after obtaining explicit consent from individuals, in line with GDPR requirements. To protect personal data, the system employs data anonymization techniques and encryption to ensure that individual identities cannot be traced. This approach safeguards personal information against unauthorized access and data breaches.

Algorithm 15 DRL-DQN based Financial Decision Making

- 1: **Input:** n_{min} (minimum experience replay size), n_{max} (maximum experience), Δ (learning rate), α (exploration rate), γ (discount factor), δ (random action probability)
 - 2: **Output:** Financial decision making strategy
 - 3: Initialize replay memory rpm to capacity n (to store diverse market scenarios)
 - 4: Initialize action value function Q with random weights θ
 - 5: Initialize target action value function \widehat{Q} with weights θ^-
 - 6: For episode = 1, M do –number of episodes representing different market conditions
 - 7: Initialize sequence $S_1 = \{X_1\}$ and preprocessed sequence $\sigma_1 = \sigma_1(S_1)$
 - 8: For $t = 1, T$ do (decision points of financial data)
 - 9: with probability δ select a random action A_t (explore the action space)
 - 10: otherwise select $A_t = \max_A Q(S_t, A, \theta)$ (exploit the learned strategy)
 - 11: Execute action A_t and observe reward R_t and next state X_{t+1} –based on financial outcome
 - 12: Set $S_{t+1} = S_t, A_t, X_{t+1}$ and preprocess $\sigma_{t+1} = \sigma(S_{t+1})$
 - 13: Store transition $(\sigma_t, A_t, R_t, \sigma_{t+1})$ in replay memory rpm
 - 14: Sample random minibatch of transitions $(\sigma_J, A_J, R_J, \sigma_{J+1})$ from replay memory rpm
 - 15: Set $Y_j = \left\{ R_J + \gamma \max_{A'} \widehat{Q}(S_{J+1}, A', \theta^-) \right\}$ for non terminal states or $Y_J = R_J$ for terminal states
 - 16: Perform gradient descent step on $(Y_J - Q(S_J, A_J, \theta))^2$ with respect to the network Parameters ϑ
 - 17: Every C steps reset $\widehat{Q} = Q$
 - 18: End inner loop
 - 19: End outer loop
 - Online financial decision making phase**
 - 20: Load the parameters ϑ ;
 - 21: Calculate action-value Q for the current financial state $(S_t, A; \theta)$
 - 22: Output $A_t = \operatorname{argmax} Q(S_t, A; \theta)$ based on the learned policy
-

The algorithm explains how the DQN in a DRL framework is used to make financial judgments. The first step is to set up a replay memory system, which is a memory system that stores a variety of market scenarios. Two different kinds of action value functions are first allocated random weights when this memory is first used. After that, the algorithm runs through a number of episodes, each of which represents a distinct set of market

circumstances. Every episode begins with the financial data series being set up, and at each time step, decisions are made iteratively. To balance the discovery of novel methods with the exploitation of proven lucrative ones, a combination of randomly picked actions and those chosen based on the greatest anticipated value from the Q-function are used in the decision-making process. The algorithm learns from the actual results of its judgments by watching the reward that results from its actions and the subsequent status of the market. The replay memory contains these events. It updates its knowledge by sampling a batch of these events on a regular basis. It does this by adjusting the network parameters through a process known as gradient descent, which improves future predictions. In order to prevent the algorithm from deviating too far from its most recent known effective method, this update occurs in cycles. The algorithm proceeds to an online decision-making stage upon the completion of the iterative learning cycles. Here, based on its forecasts and acquired information, it utilizes the factors it has learnt to assess the present financial situation and choose the optimal course of action, such as purchasing or selling assets. Because of this procedure, the algorithm becomes a dynamic instrument for financial decision-making that is always learning from and adjusting to the shifting conditions of the financial market.

LearnFlex utilises a cloud-based infrastructure for dynamic scaling based on real-time demand. This ensures that server capacity can be quickly adjusted to handle spikes in user access, particularly during peak times like the start of new courses or examination periods. The system employs advanced load-balancing techniques to distribute traffic evenly across servers, preventing any single server from becoming a bottleneck. This enhances performance and ensures a smooth and responsive experience for all users.

The DSS ensures transparency in trading activities and decision-making processes by maintaining detailed logs and reports. This aligns with MiFID II's requirements for transparency and accurate reporting to regulators. By employing advanced analytics, the system helps identify and mitigate market abuse and ensure fair trading practices, thereby supporting the integrity of financial markets as envisaged by MiFID II. The system is designed to analyse a wide range of data to ensure that trades are executed at the best possible terms for clients following MiFID II's best execution requirements.

3.2.2. TRPO (Trust Region Policy Optimization). TRPO is a critical component of the proposed DRL-DSS that helps the system make dependable and efficient financial judgments. TRPO is an advanced algorithm that works to gradually enhance the system's decision-making policy without bringing about abrupt or unstable modifications. Its primary goal is to gradually adjust the system's approach so that it may pick up new information and experiences without deviating too much from its prior understanding. This is especially crucial in the irregular and turbulent world of finance, where it's necessary to strike a balance between trying out novel tactics and upholding a certain standard of dependability and consistency. Through the use of TRPO, the DRL-DSS is able to minimize the danger of large policy swings that might result in subpar performance or unanticipated losses while also learning gradually and making increasingly educated judgments. This allows the system to adapt to the constantly shifting financial markets.

Algorithm 16 Trust Region Policy Optimization

```

Initialize  $\pi_0$  suitable for financial market scenarios
for  $do i = 0, 1, 2, \dots$  Until convergence do
    Compute all advantage values  $a_{\pi_i}(S, A)$ 
    Solve the constrained optimization problem to update the policy
    Update the policy using  $\pi_{i+1} = \operatorname{argmin}_{\pi} [L_{\pi_i}(\pi) + \left( \frac{2\epsilon\gamma}{(1-\gamma)^2} \right) D_{KL}^{max}(\pi_i, \pi)]$ 
    where  $\epsilon = \max_S \max_A |A_{\pi}(S, A)|$ 
    And  $L_{\pi_i}(\pi) = \mu(\pi_i) + \sum_S P_{\pi_i}(S) \sum_A \pi(A|S) A_{\pi_i}(S|A)$ 
    Repeat the steps 2-4 until the policy converges
end for

```

Enhancing decision-making methods in a stable and regulated manner is the goal of the improved TRPO for a DRL-DSS in financial situations. First, a policy, represented by π_0 , is initialized. This establishes the starting point for financial choices such as purchasing, disposing of, or retaining assets. After that, the algorithm refines this strategy iteratively. The advantage values for the current policy are initially determined by the

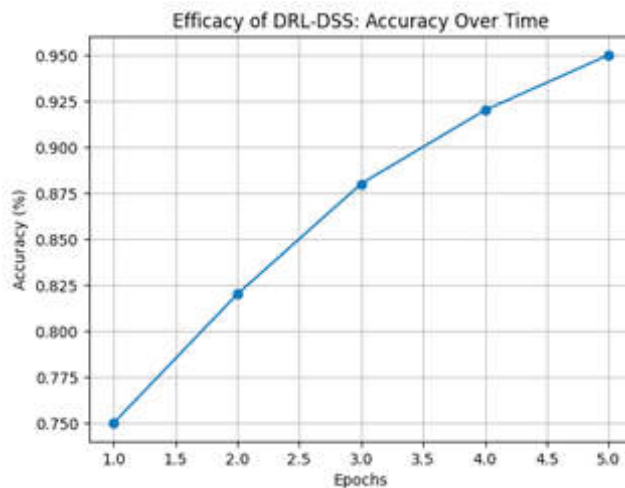


Fig. 4.1: Accuracy

algorithm in each iteration. When deciding whether to purchase or sell stocks, for example, these numbers indicate how much better or worse a certain action is in comparison to the typical action in that particular financial situation. This is a critical stage in determining how successful the market's present decisions are. The method then resolves an optimization issue using constraints. The goal of this stage is to identify a new policy that marginally outperforms the existing one. Here, stability must be maintained by ensuring that the new policy doesn't stray too far from the previous one. This is especially crucial in the banking industry because of how irregular and unsettled the markets can be. A precise formula that strikes a compromise between making improvements and adhering closely to the prior version of the policy is used to update it. It restricts the amount of change to prevent significant deviations and takes into account how much the projected return on financial activities would increase under the new policy. Eventually, this procedure is carried out again until the policy converges, or reaches a point at which it no longer changes significantly. This suggests that the system has discovered a reliable and efficient method for deciding how to allocate funds. The financial industry, where reliability and performance are crucial, is a perfect fit for the TRPO because of its emphasis on steady development.

4. Results and Experiments. We move further with the evaluation of the suggested innovative DRL-DSS based on the study [16].

4.1. Evaluation Criteria. The performance of the suggested DRL-DSS across several epochs demonstrates its remarkable efficacy in Figure 4.1. The accuracy increases clearly and consistently throughout the course of five epochs, rising from 0.75 in the first epoch to 0.95 by the fifth. This pattern suggests that the system is picking up new skills and becoming more adept at what it does. It is especially remarkable that the last epoch reached a high degree of precision, at 0.95. It indicates that 95% of the time the DRL-DSS makes accurate judgments or predictions, which is a powerful testament to its efficacy and capabilities. Furthermore, the notable increase in accuracy in just five epochs demonstrates the system's capacity for quick learning. This is an essential characteristic in the dynamic and fast-paced field of financial data analysis, where the capacity to swiftly adjust to new knowledge and market developments is priceless. In addition, the system's dependability is demonstrated by the continuously high accuracy rates throughout all epochs. A consistent performance like this indicates that users may rely on the DRL-DSS to provide reliable and accurate forecasts or choices, which confirms the tool's appropriateness for financial analysis and decision-making.

The accuracy and recall values of the proposed DRL-DSS may be evaluated throughout a sequence of epochs, which correspond to discrete time intervals in the systems learning and operating phases, as shown in Figure 4.2. The accuracy values show a progressive rise over the period of five epochs, peaking at 0.92 and

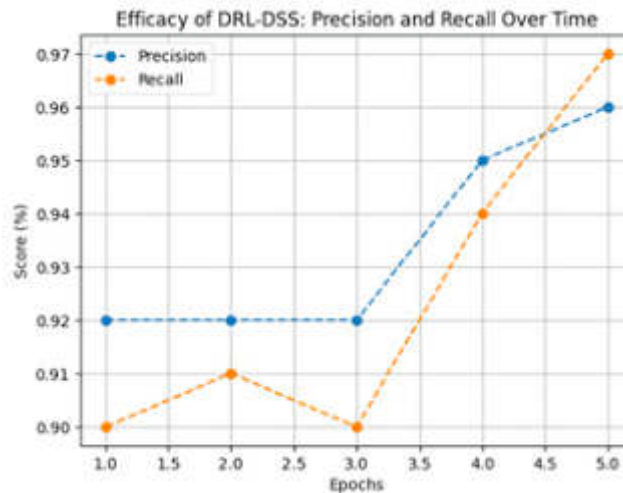


Fig. 4.2: Precision and recall

eventually reaching 0.96 by the fifth epoch. The system's capacity to correctly identify relevant cases without overextending to irrelevant ones is demonstrated by its consistency and increasing trend in precision. This is an important feature in financial decision-making, where accuracy in forecasts or classifications is critical. In a similar vein, the recall values show how well the system recognizes all relevant occurrences; they begin at 0.90 and increase to 0.97 by the fifth epoch. The increase in recall indicates that the system is getting better at catching all possible opportunities or threats as it develops, which is a crucial attribute in dynamic financial contexts where it can be expensive to overlook important information. High recall and accuracy numbers combined over all epochs show the efficacy of the DRL-DSS. As demonstrated by the precision and recall figures, the system not only keeps up a high degree of prediction accuracy but also makes sure that all important data points are identified. This balance is especially important in financial contexts, where decision-making results can be greatly impacted by the accuracy of forecasts as well as the completeness of information gathered.

An excellent way to assess the effectiveness of the suggested DRL-DSS is to examine its F1-Score values across a number of epochs present in Figure 4.3. An impartial assessment of the accuracy and completeness of the system in forming judgments or predictions is given by the F1-Score, which is a harmonic mean of precision and recall. The F1-Score values, in this case, throughout the course of epochs are 0.89, 0.90, 0.92, 0.92, and 0.96, respectively. We can clearly observe an improving trend in these ratings as they develop. The DRL-DSS exhibits a steady increase in its decision-making efficacy, beginning with a comparatively high score of 0.89 in the first epoch and ending with a score of 0.96 by the fifth. This increasing trend shows the system's strong initial performance as well as its capacity to successfully learn and adapt over time. In particular, the system's final F1-Score of 0.96 indicates a high degree of accuracy and dependability in its predictions, suggesting that it covers a wide variety of pertinent cases and makes the right selections the majority of the time. Such performance is suggestive of a system that can reduce false positives and negatives in addition to being proficient at accurately detecting and acting upon relevant data items. This is particularly important since errors can have a big cost in the dynamic and intricate field of financial data analysis.

5. Conclusion. In conclusion, the DRL-DSS that has been suggested has shown to be incredibly effective when it comes to financial data processing. The system's strong capacity to make comprehensive and precise financial decisions is demonstrated by the steady increase in key performance indicators over time, especially the F1-Score. The F1-Score values of the DRL-DSS have been growing consistently from 0.89 to 0.96, demonstrating the model's ability to cover a wide range of relevant financial scenarios in addition to producing accurate forecasts. The F1-Score captures this balance between recall and accuracy, which is critical in the volatile and

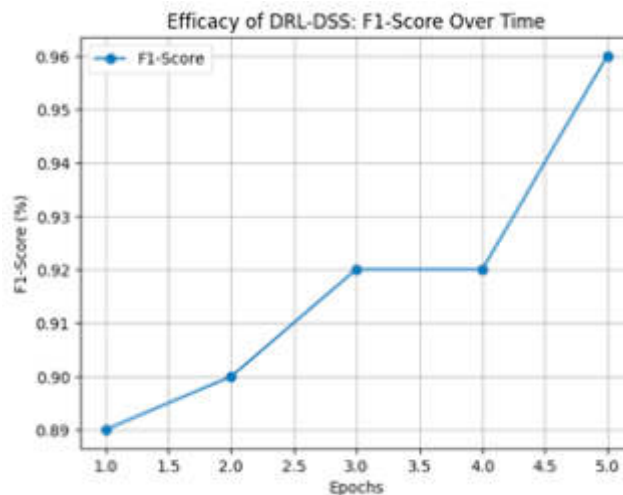


Fig. 4.3: Efficacy of DRL-DSS: F1-Score over time

complicated field of finance where decisions can have far-reaching effects. The system is a very dependable and useful tool for financial experts because of its capacity to adapt and learn over time, as shown by the rising trend in its performance indicators. It offers evidence of the potential benefits of incorporating cutting-edge machine learning methods into decision support systems, such as deep reinforcement learning. The development of intelligent financial decision-making tools has advanced significantly with the DRL-DSS's ability to navigate the complexities of financial data analysis. This study establishes a standard for the creation of complex, data-driven decision support systems in the banking industry and beyond, and it opens the door for future developments. Future developments will likely focus on refining and advancing DRL models to enhance their predictive accuracy, efficiency, and scalability. This includes exploring cutting-edge neural network architectures, reinforcement learning strategies, and algorithmic improvements to better handle the nuances of financial data and decision-making processes.

REFERENCES

- [1] Y. ANSARI, S. YASMIN, S. NAZ, H. ZAFFAR, Z. ALI, J. MOON, AND S. RHO, *A deep reinforcement learning-based decision support system for automated stock market trading*, IEEE Access, 10 (2022), pp. 127469–127501.
- [2] C. CAPUTO AND M.-A. CARDIN, *Analyzing real options and flexibility in engineering systems design using decision rules and deep reinforcement learning*, Journal of Mechanical Design, 144 (2022), p. 021705.
- [3] G. CHENG, *Intelligent financial data analysis and decision management based on edge computing*, Journal of Sensors, 2022 (2022).
- [4] F. O. FEDIN, O. V. TRUBIENKO, AND S. V. CHISKIDOV, *Assessment of intelligent decision support systems effectiveness in technological processes of big data processing*, in 2019 International Russian Automation Conference (RusAutoCon), IEEE, 2019, pp. 1–6.
- [5] T. JIA, C. WANG, Z. TIAN, B. WANG, AND F. TIAN, *Design of digital and intelligent financial decision support system based on artificial intelligence*, Computational Intelligence and Neuroscience, 2022 (2022).
- [6] D. JUNG, V. TRAN TUAN, D. QUOC TRAN, M. PARK, AND S. PARK, *Conceptual framework of an intelligent decision support system for smart city disaster management*, Applied Sciences, 10 (2020), p. 666.
- [7] K. KOURTIT AND P. NIJKAMP, *Big data dashboards as smart decision support tools for i-cities—an experiment on stockholm, Land use policy*, 71 (2018), pp. 24–35.
- [8] M. KRAUS AND S. FEUERRIEGEL, *Decision support from financial disclosures with deep neural networks and transfer learning*, Decision Support Systems, 104 (2017), pp. 38–48.
- [9] D. T. S. KUMAR, *Data mining based marketing decision support system using hybrid machine learning algorithm*, Journal of Artificial Intelligence and Capsule Networks, 2 (2020), pp. 185–193.
- [10] K. LAL, V. K. R. BALLAMUDI, AND U. R. THADURI, *Exploiting the potential of artificial intelligence in decision support systems*, ABC Journal of Advanced Research, 7 (2018), pp. 131–138.
- [11] S. LI AND T. WU, *Deep reinforcement learning-based decision support system for transportation infrastructure management*

- under hurricane events*, Structural Safety, 99 (2022), p. 102254.
- [12] M. W. MOREIRA, J. J. RODRIGUES, V. KOROTAEV, J. AL-MUHTADI, AND N. KUMAR, *A comprehensive review on smart decision support systems for health care*, IEEE Systems Journal, 13 (2019), pp. 3536–3545.
- [13] X. QIU, X. TAN, Q. LI, S. CHEN, Y. RU, AND Y. JIN, *A latent batch-constrained deep reinforcement learning approach for precision dosing clinical decision support*, Knowledge-based systems, 237 (2022), p. 107689.
- [14] S. REN, *Optimization of enterprise financial management and decision-making systems based on big data*, Journal of Mathematics, 2022 (2022), pp. 1–11.
- [15] S. SPÄNIG, A. EMBERGER-KLEIN, J.-P. SOWA, A. CANBAY, K. MENRAD, AND D. HEIDER, *The virtual doctor: an interactive clinical-decision-support system based on deep learning for non-invasive prediction of diabetes*, Artificial intelligence in medicine, 100 (2019), p. 101706.
- [16] D. TONG AND G. TIAN, *Intelligent financial decision support system based on big data*, Journal of Intelligent Systems, 32 (2023), p. 20220320.

Edited by: Rajanikanth Aluvalu

Special issue on: Evolutionary Computing for AI-Driven Security and Privacy:
Advancing the state-of-the-art applications

Received: Jan 6, 2024

Accepted: Feb 9, 2024



RESEARCH ON THE CONSTRUCTION OF INTELLIGENT SYSTEM OF LANDSCAPE IN SCIENCE AND INNOVATION PARK OF SMART CITY – BASED ON THE CONCEPT OF SMART GARDEN DESIGN

KE XIE*

Abstract. With the advancement of urbanization in China, there is a growing demand for intelligent landscape management in urban science and innovation parks. To improve landscape management efficiency in science and technology innovation parks, this study adjusted the Inception module and activation function of the GoogLeNet algorithm, designed an improved GoogLeNet algorithm, and based on this, constructed a multi clue science and technology innovation park landscape vegetation intelligent recognition system. This is also the main contribution of this study. The experimental results show that the average image recognition accuracy of the Improve GoogLe Net+MM model designed from the study is 92.95%, which is higher than all the comparison models. Additionally, the computation time is significantly lower than the comparison models when the computation data volume is larger. The model designed in this study has a certain application value to improve the performance of the landscape intelligent management system within smart cities' science and technology parks.

Key words: Smart city; Landscape; Intelligent management; Plant recognition; GoogLe Net

1. Introduction . After entering the 21st century, artificial intelligence technology has developed by leaps and bounds, especially image recognition technology. This technology is extensively utilized in security, e-commerce, and various other industries [10]. Due to the demand for smart city construction, the landscape plant management mode at urban science and innovation park needs to become more intelligent in order to reduce the cost of manual management and improve management efficiency, and lay the foundation for subsequent intelligent plant watering and intelligent fertilization to remove pests. Therefore, the application of image recognition technology in artificial intelligence to this field is a potential research direction [13]. Intelligent recognition of landscape vegetation is a commonly utilized aspect within this area. However, issues such as inadequate recognition accuracy or speed frequently arise. Therefore, this study aims to enhance insufficient recognition accuracy and low recognition efficiency of image recognition models based on AI technology. The study will adjust the activation function and Inception module of the GoogLeNet neural network algorithm in artificial intelligence technology. Subsequently, a multi clue integration model for intelligent recognition of landscape plants in science and technology innovation parks will be designed with the improved GoogLeNet algorithm as the core. This enables GoogLeNet to obtain the possibility of improving recognition image accuracy while minimizing network parameters. This is the main academic contribution reflected in this study. In order to verify the application effect of the designed model, several comparative plant recognition experiments were arranged and carried out, in which not only the effects of using different dataset processing methods and model training methods on the final image recognition performance of the model are compared, but also the recognition accuracy of different plant recognition algorithm models built with the same dataset processing methods and model construction methods are horizontally compared.

2. Related works. Patel H et al. proposed an improved convolutional neural network called “Depth-FuseNet” [12]. This algorithm is used to fuse thermal and visible images, and extract data features from the two images to generate high-latitude features. The experiment found that this method has a recognition accuracy of 6.74 percentage points higher than the VGG16 algorithm in constructing recognition systems on the test set [12].

The research team of Hwang used an improved convolutional neural network to build a model for detecting neovascularization and age-related macula of patients. The experimental outcomes demonstrated that the

*School of Culture and Tourism, Chengdu Polytechnic, Chengdu, 610041, China (xie_ke0041@outlook.com)

approach effectively enhances clinical diagnosis efficiency [4].

The paper [7] concluded that deep learning-based detection of optic disc abnormalities in color fundus photographs is mainly limited to the field of glaucoma. However, numerous systemic and neurological diseases that pose life-threatening risks can present as deviations in the optic disc. Therefore, the authors trained a color fundus photo optic disc abnormality detection model using migration learning for ResNet-152 neural network, and the test results showed that the algorithm model has a significant advantage over the comparison algorithm with 90% recognition sensitivity and 69% specificity [7].

The paper [2] developed a gangue recognition method using a convolutional neural network based on an auditory model. According to the experimental results, the method achieved a high recognition accuracy of 99.5%. In addition, the method offers significant noise immunity in comparison to frequently utilized recognition methods across different noise conditions [2]. Maior C S believes that with the rapid spread of SARS coronavirus type 2 around the world, the scientific community has spent a lot of energy to better understand the characteristics of the virus and the possible methods of prevention, diagnosis and treatment of COVID-19. Therefore, the author proposes an improved convolutional neural network to aid in diagnosing COVID-19. The test results show that the model achieves a balance accuracy of 87.7% when predicting one of the three categories (“no discovery”, “COVID-19” and “pneumonia”), and a specific balance accuracy of 97.0% in predicting the “COVID-19” category [9].

The research team of [1] proposed an improved convolutional neural network to improve the recognition accuracy of protein images based on mass spectrometry. According to experimental results, the neural network significantly improves the recognition accuracy of protein images [1].

The paper [6] constructed an improved convolutional neural network model utilizing a two-way long-short memory network and attentional architecture to identify protein-specific spot recognition. The test results showed that the recognition performance of the algorithm is better and the computational speed is faster than the traditional convolutional neural network algorithm [6].

The author of [5] designed a convolutional neural network that combined the Gaussian statistical method. The test results show that the recognition effect of the algorithm for quantum microscopic images is significantly higher than the traditional deep learning algorithm. However, it has lower computational efficiency, being 24.2% slower than the recognition model built by the GoogLe Net algorithm and 21.9% slower than the recognition model built by the VGG16 algorithm [5].

In summary, although a lot of algorithm improvement studies have been conducted by previous people to improve the image recognition efficiency of the convolutional neural network, most of the studies failed to reduce the computational time consuming of the algorithm and improve the computational efficiency under the premise of improving the recognition accuracy of the algorithm. And identifying plants in science and technology parks is a significant and complex task that demands an intelligent system with higher efficiency. Therefore, this research seeks to improve the recognition speed of the recognition algorithm under the premise of improving its accuracy.

3. Design of intelligent plant identification system based on GoogLe Net algorithm for smart city science and innovation park.

3.1. Improved GoogLe Net algorithm design considering plant image features. Intelligent plant recognition belongs to an image recognition task, and the use of artificial intelligence (AI) algorithms is widespread in this area. The GoogLe Net neural network in artificial intelligence technology is an algorithm with good performance and fast computation speed that has been made public in recent years [15]. In addition, considering the aesthetics of the landscape and the intelligent and precise management requirements of the urban science and innovation park [3]. Figure 3.1 illustrates the typical network structure of GoogLe Net, which has demonstrated high recognition accuracy. The structure shown in Figure 3.1 has been proven to have high recognition accuracy, and the model is reasonable in terms of computation and training time, and computation consumption resources [11]. The GoogLe Net algorithm incorporates the Inception module, which expands the network width and reduces the computational effort by using asymmetric convolution. Specifically, it adds a 1×1 network in front of the 5×5 and 3×3 convolutional layers to reduce the dimensionality of the data.

The network input is a $224 \times 224 \times 3$ RGB image, and preprocessing requires subtracting the mean data from the training set's RGB channels for each pixel. That is to subtract the mean data of the three primary color

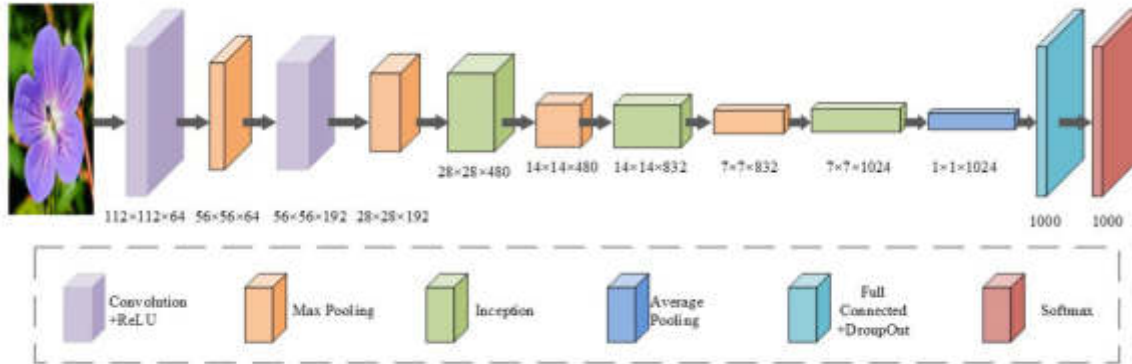


Fig. 3.1: Typical structure hierarchy of GoogLe Net

channels obtained from the training set for each pixel. And to reduce the error caused by the phenomenon of “gradient vanishing”, here it sets the convolutional layer in the Inception module to use ReLU as the activation function. Specifically, $5 * 5$. $3 * 3$ convolutional layers are used to reduce the depth of the convolutional layer through a $1 * 1$ filter. After maximizing the pooling layer, a $1 * 1$ filter is also used to reduce the depth of the convolutional layer. After reducing the size or achieving maximum pooling, ReLU activation processing is required. The input layer of the GoogLeNet neural network is connected to a regular convolutional layer and a maximum pooling layer, followed by a reduced size convolutional layer and a regular pooling layer. Afterwards, a separate convolutional layer should not be designed. A 7-step average pooling layer of $7 * 1$ should be set after the Inception module to reduce feature extraction errors caused by neighborhood size limitations during the convolution calculation process. After completing the data dimensionality reduction, the data will be input into the subsequent Dropout layer, with an output ratio of generally 60%. It will be followed by a fully connected layer containing 1024 neurons, which still uses the ReLU activation function. Finally, the data will be input into a softmax function classifier, and the predicted category data of the corresponding size will be output according to the usage requirements. After the design is completed, the Inception module structure is shown in Figure 3.2. As shown in Figure 3.2, compared to the simple Inception module, the Inception module in (b) subgraph has added a $1 * 1$ convolutional filter, which can achieve the effect of changing the data dimension without changing the original feature structure, meeting the proportion invariance of features.

Considering the limited variety of cultivated landscape plants in the science park, especially the dataset vegetation types used to test the performance of the algorithm in this study are only 12, the number of neurons in the last fully connected layer of the GoogLe Net algorithm is also reduced to 12. Considering that the system needs to cope with the classification task, the cross entropy function with L_1 regular term is used here to construct the loss function and the adaptive gradient method is used to estimate the network parameters, and the loss function is calculated as shown in equation (3.1).

$$loss = -\frac{1}{N} \sum_{n=1}^N \sum_{m=1}^M [p_m^n \log p_m^{\hat{n}} + (1 - p_m^n) \log (1 - p_m^{\hat{n}})] + \frac{1}{2} \lambda \sum_{n=1}^N \sum_{l=1}^L [\|W_l\|_1 + \|b_l\|_1] \quad (3.1)$$

In equation (3.1), M represents the number of images and plant species in each training batch, p_m^n and $p_m^{\hat{n}}$ represent the true probability that the n image belongs to them category and the probability that it is predicted to be the category, W_l and b_l represent the weight coefficients and intercept coefficient matrix of the l layer, respectively. λ The former in this study is initially set to 5×10^{-4} according to the industry experience. L represents the regularization coefficients and the number of layers containing parameters in the network, respectively. The architecture of the GoogLeNet algorithm has been established, and detailed improvement methods for the GoogLe Net algorithm used for plant recognition tasks will continue to be designed.

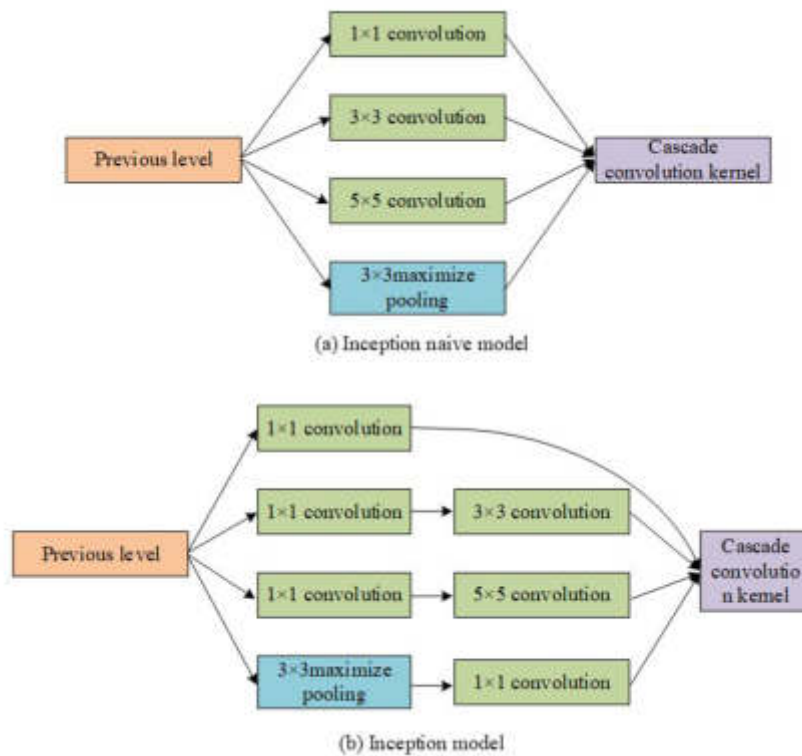


Fig. 3.2: Schematic diagram of the calculation structure of the designed Inception module

In summary, a series of structural improvements have been made to GoogLeNet to solve the problem of gradient vanishing. The GoogLeNet algorithm introduces the Inception module as its initial module to increase the network width and reduce computational complexity. This module adopts an asymmetric convolution method, which reduces the dimensionality of the data by adding a 1x1 network before the 5x5 and 3x3 convolutional layers. In the activation module, the convolutional layer in the Inception module uses ReLU as the activation function, while the 5x5 and 3x3 convolutional layers use a 1x1 filter to reduce the depth of the convolutional layer. After maximizing the pooling layer, a 1x1 filter is also used to reduce the depth of the convolutional layer. By introducing the Inception module and increasing the depth and width of the network, the network can better capture the complex features and structural information of landscape vegetation. This helps to identify different types of vegetation, their morphology, growth status, and other details, thereby improving the cognitive ability of vegetation. By using asymmetric convolution and adding a 1x1 network before the 5x5 and 3x3 convolutional layers to reduce the dimensionality of the data, the computational load can be effectively reduced. This enables the network to process large-scale image data more efficiently, improving the speed and accuracy of vegetation recognition. And ReLU, as an activation function, has nonlinear characteristics, which helps the network better learn and represent the complex features of vegetation, and can help the network better capture the nonlinear relationship of vegetation, improving the accuracy and robustness of vegetation recognition.

3.2. Design of intelligent plant recognition system based on hybrid improved GoogLe Net algorithm. . On the one hand, the deeper the hierarchy, the more training data are needed to train the neural network before it may be able to play a good application effect. However, the image data used for training the plant recognition system at the Smart City Science and Technology Park may be extremely limited, so it

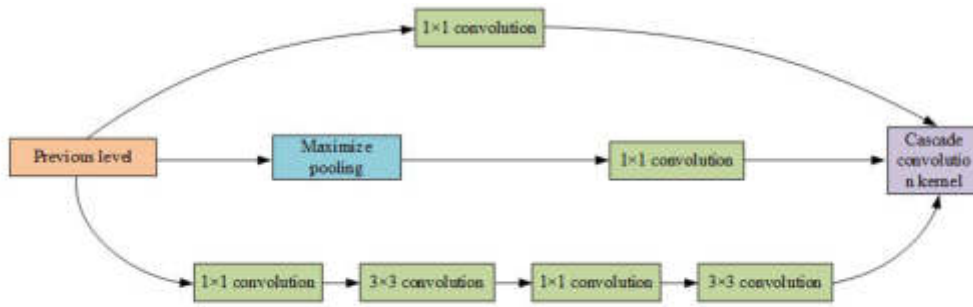


Fig. 3.3: Improved Inception module computational structure

is necessary to build a convolutional neural network with fewer network parameters to ensure its adaptability and good computational performance. Therefore, this time, various improvements are made to the classical GoogLe Net neural network algorithm to meet the demands of the intelligent plant recognition system in the science and innovation park.

In this study, the main purpose of achieving the overall number of parameters reduced while maintaining the computational accuracy of the GoogLe Net neural network is from the perspective of improving the Inception module. In the design idea of Visual Geometry Group Net (VGGNet) neural network, two 3×3 convolutional layers can achieve the same receptive field as a 5×5 kernel while reducing the number of parameters by at least 20% compared to the latter. In the process of improving the Region-Convolutional Neural Networks (R-CNN) target detection network to Fast R-CNN, the structure of reusing the convolutional layer information is used in it in order to reduce the training time of the algorithm. Specifically, the R-CNN algorithm utilizes the convolutional structure to create a feature map for each proposed region individually. While Fast R-CNN is also a convolutional feature map in the original image, resulting in all proposed boxes being formed in the same feature map output following convolutional computation. Based on the concepts of reusing the convolutional feature layer and convolutional kernel substitution, an improved Inception module is now designed by combining the two computational branches of 3×3 and 5×5 size convolutional kernels in the original Inception module; The input module first uses a 1×1 size convolutional kernel for dimensionality reduction processing, followed by an input 3×3 size convolutional block; The output information for this convolutional layer has two branches, one of which will directly use the output information as one of the outputs of the Inception module; The other branch will pass through convolutional layers of size 1×1 and 3×3 in order to reduce the output to one of the final module outputs. Figure 3.3 displays the enhanced computational hierarchy of the Inception module.

The activation function plays a crucial role in determining the final computational performance of a deep neural network algorithm. A nonlinear activation function can effectively ensure the network's fitting ability. In general, a good activation function needs to be monotonic, non-saturated, low computational complexity, with few parameters, non-linear, and differentiable everywhere. However, the ReLU or Swish activation functions used in traditional neural networks suffer from neuron "necrosis" and large computational effort, respectively. The latter is contrary to the original purpose of the research design to improve the Inception module. Therefore, in order to improve the computational accuracy and reduce the computational complexity of the GoogLe Net neural network, H_Swish function is chosen as the activation function in the algorithm, and its calculation method is shown in equation (3.2).

$$H_Swish(x) = x \cdot ReLU(x + 3)/6 \quad (3.2)$$

In equation (3.2), x is the independent variable of the input H_Swish function. After improving and replacing the Inception module and the activation function of the GoogLe Net neural network respectively, the other structures in the neural network and the way the parameters are updated are left unchanged.

Given the limited amount of training and testing image data utilized in this study, it is necessary to use the migration learning technique to ensure high computational accuracy of the algorithm under such conditions.

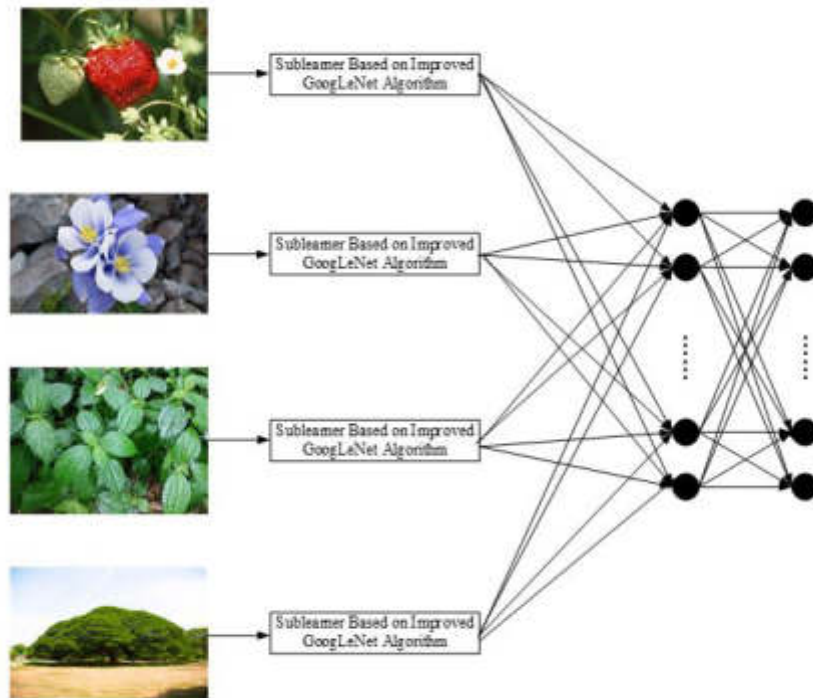


Fig. 3.4: Computational structure of plant multi-cue recognition model based on convolutional neural network algorithm

In other words, the study borrowed more complete and rich training data to train a neural network with more reasonable and mature parameters. Specifically, with reference to the general steps of migration learning, this study chose to use a richer image dataset to train the improved GoogLe Net neural network. Specifically, the last two fully connected layers in the network structure were modified to adjust the number of fully connected neurons to the number of desired classification categories, and a richer image dataset was used to train the neural network. After training, the neural network was trained again by applying the originally matched dataset to fine-tune some of the parameters to make the network more adaptable to the plant recognition needs of the science and technology park.

In plant recognition tasks, the information on recognition features that a single plant organ can provide is often more limited, especially different plant species in the same subject may have organs with similar appearance. For instance, while the flower shape and color of plum blossom and cherry blossom are relatively similar, but the leaf structure and the shape of the entire plant of plum blossom tree and cherry blossom tree differ considerably. It can be seen that the utilizing the multi-cue model for developing a plant recognition system can further improve the plant recognition ability of the system and alleviate the recognition error problem caused by the over-similarity of individual organs or local structures of plants to a certain extent. The computational layout of the plant multi-cue model based on the convolutional neural network algorithm is shown in Figure 3.4. As shown in Figure 3.4, the convolutional neural network-based plant multi-cue feature recognition model consists of two key components. Firstly, it requires training a sub-recognition model for each major recognition organ of the plant for a single organ, and the sub-recognition model in this study consists of a modified GoogLe Net neural network. Secondly, it involves efficiently merging the independent single-organ recognition models.

It is evident that each input for the plant multi-cue recognition model constructed in this study, is the final

computational output of a related single-organ recognition model. This facilitates the encompassing model in acquiring an expanded insight into the plants, that are supposed to be recognized, in distinct dimensions. However, whether the final recognition results after integration can be more accurate than individual models depends mainly on the integration method of each single model. This study refers to the sublearners integration method of random forest algorithm, and uses the weighted summation of each single organ classifier to form the final classification results. The percentage of the sublearner's contribution in the final prediction is determined by the prediction category score of the single-organ prediction model. The detailed design of the integration approach is presented in this section. Since most plants can be identified by four aspects: leaves, fruits, flowers, and whole plants, only four sub-learners are also correspondingly set in the multi-cue plant identification model designed in this study, assuming that the model identification accuracy of each sub-learner is A_1 , A_2 , A_3 , and A_4 , and can be described by equation (3.3).

$$A = \{A_1, A_2, A_3, A_4\} \quad (3.3)$$

In equation (3.3), A represents the sublearners recognition accuracy matrix, then the integrated output weight of each sublearners can be described by equation (3.4),

$$q_i = \frac{A_i}{\sum_i^N A_i} \quad (3.4)$$

Where A_i represents the classification accuracy of the i th plant organ recognition model on the test set, and N represents the number of single-organ plant recognition sublearners. It should be noted that the inputs of each sublearners are the corresponding plant part images of the plant parts, and each sublearner operates independently of the others. The sublearners' scores weighted according to the weights of equation (3.5) will be used as the initial values of the prediction scores for each category of the multi-cue model S_i^* ,

$$S_i^* = \frac{q_i S_i}{\sum_{i=1}^4 q_i \cdot S_i} \quad (3.5)$$

Where P_j is the score of the final predicted category by the i th sub learner, which needs to satisfy the $0 < S_i < 1$ relationship, and S_i^* the initial score of the i th sub learner input to the integrated model. If the same category scores appear in the sub learners, it means that multiple sub learners predict the same plant category and they need to be combined. That is, if $P_i = P_j$, and $P_i, P_j \in \{1, 2, \dots, 12\}$, $i < j$, the relationship of equations (3.6) and (3.7) exists.

$$S_i^* = S_i^* + S_j^* \quad (3.6)$$

$$P_j = 0 \quad (3.7)$$

In Eq. (3.7), P_j represents the predicted plant category output by the j sub learners. The integrated learner then counts the scores of each category label of the current image to be recognized and outputs the category with the highest score as the predicted category, as shown in Eq. (3.8).

$$\hat{y} = P_i \quad (i = \arg \max(S_i^*), i = 1, 2, 3, 4) \quad (3.8)$$

Finally, the experiment presents the calculation formula for the test indicators. This study shares two indicators: accuracy and calculation time. The latter is obtained through a computer-based timer and does not require further processing. Equation (3.9) demonstrates the calculation method for accuracy.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.9)$$

In equation (3.9), TP and TN respectively represent the number of cases accurately judged as positive and negative, while FP and FN respectively represent the number of cases wrongly judged as positive and wrongly judged as negative.

Table 4.1: Statistics on the number of various types of images in the PlantCLEF2016 dataset

Number	Category ID	Blade pictures	Fruits pictures	Flower pictures	Pictures of the whole plant
#01	309265	198	96	62	119
#02	007574	120	24	37	255
#03	106245	27	23	91	122
#04	012807	41	29	24	98
#05	016730	32	61	465	135
#06	041186	25	25	28	164
#07	309347	71	51	67	230
#08	012513	115	38	38	241
#09	251599	45	117	82	256
#10	228025	30	23	33	40
#11	309295	359	27	81	269
#12	133595	162	80	50	293

4. Plant intelligent identification system performance analysis.

4.1. Performance analysis experimental program development. The following computational experiments are designed to validate the performance of the improved GoogLe Net neural network algorithm-based plant intelligent recognition system for science and technology parks designed in this study. The datasets used in the experiments are ImageNet, which is a large image database containing at least 14,000K in size and more than 20,000 categories, and PlantCLEF2016, which assisted in the initial training of the model during the migration learning phase. The latter dataset was also used for fine-tuning and performance testing of the model post-initial training. Table 4.1 displays the key components of the PlantCLEF2016 dataset.

Before using the dataset, it also needs to be preprocessed. To avoid the model parameters biased to certain categories, the image quality is enhanced by using perspective transformation, affine transformation, etc. to expand a smaller number of image types and by using contrast enhancement, color dithering, etc. To compare the computational effects of the improved algorithms, the VGG16 algorithm, the classical Google Net algorithm, and the Faster-RCNN algorithm, which are commonly used in the industry and have good performance, are selected as the comparison methods to build the plant recognition system. The hyperparameters for each neural network algorithm were determined using the dichotomous method of multiple debugging with common tuning parameters in the industry. The parameters were not duplicated. To comprehensively compare the performance of the algorithms designed in this study, a comparative experimental scheme is designed as shown in Table 4.2. It is important to note that "MM" in Table 4.2 represents the abbreviation for the multi-cue model.

The reason for choosing this method of "expand+drop" for preprocessing is to solve the imbalance problem in the dataset and enhance the quality of training data. By expanding the dataset, the model can be exposed to more diverse images, which helps improve its generalization ability for unseen data. Discarding images with insufficient pixel ratio helps prevent the model from bias towards certain categories and ensures that relevant features are focused during training. The use of "expand+drop" preprocessing techniques can enhance the performance of the model by providing more diverse and balanced datasets. This may lead to better generalization ability and higher accuracy during training and testing stages. By focusing on relevant features and avoiding bias towards specific categories, the model becomes more robust and effective, and can identify different plant organs or species.

Transfer learning can utilize knowledge gained from a task to improve the learning of related tasks. In this case, the source task involves model training on a large dataset such as ImageNet, which contains various images from multiple categories. By pre training on ImageNet, the model learned rich and universal features that can be transferred to the target task of plant recognition. The use of transfer learning can significantly improve the performance of models, especially when applied to tasks with limited labeled data. By using the weight initialization model learned in ImageNet, the model learns more basic features from images, which can accelerate convergence speed in target task training. This typically leads to faster training and better

Table 4.2: Comparison experimental protocol display table

Experiment number	Algorithm solutions	Program explanation	Purpose of the experiment
#01	Improve GoogLe Net+MM+New Learning	Direct training and testing with PlantCLEF2016	Exploring whether the use of migration learning is beneficial for improving recognition system performance
	Improve GoogLe Net+MM+Migration Learning	Initial training with ImageNet, fine-tuning and testing with PlantCLEF2016	
#02	Improvements to GoogLe Net+MM	Multi-cue model by improved GoogLe Net	Exploring the effect of different model organization methods on recognition performance
	Improving GoogLe Net+ Flower Single Organ	Improving the GoogLe Net algorithm by training with only flower images	
	Improving GoogLe Net+ fruit single organ	Training a single model using only fruit images	
	Improvement of Google Net+ leaf single organ	Training a single model using only leaf images	
	Improved GoogLe Net+ whole single organ	Training a single model with only the whole image	
	Improved GoogLe Net+ hybrid recognition	Training a single model with all images	
#03	Improvements to GoogLe Net+MM	/	Comparing the recognition performance of multi-cue models composed of different algorithms
	VGG16+MM	Multi-cue model construction using the VGG16 algorithm	
	GoogLe Net+MM	Similar to the previous algorithm scheme	
	Faster-RCNN+MM		
#04	Consistent with #03	/	Compare the calculation time of each model

Table 4.3: Experimental working environment and hyperparameter setting results

Type	Number	Name	Values and Setting Results
Hardware environment	#11	Host processor	Intel Core i7-6800K
	#12	Random Access Memory Specifications	6GB
	#13	Read Only Memory Specifications	1024GB
Software environment	#21	Operating system	Windows 10 Professional Edition
	#22	Programming language	Python
	#23	Database software	MySQL
Parameter settings	#31	Learning rate	0.0001
	#32	Maximum number of iterations	800
	#33	Does the hidden layer have offset items	Yes
	#34	Parameter initialization method	Random Initialization
	#35	Number of training samples in a single batch	64

generalization performance, resulting in higher accuracy and better overall performance.

The environmental settings and hyperparameter settings for this experiment are shown in Table 4.3. The environmental settings are categorized as either hardware or software. The hyperparameters are obtained by conducting multiple experiments within the conventional range or conventional setting method to select the optimal value. The remaining parameter settings were determined during the model design phase.

4.2. Analysis of experimental results. The horizontal axis of Figure 4.1 displays various strategies for preprocessing data sets, including processing method 1, processing method 2, and processing method 3, which correspond to “no expansion + no discard”, “no expansion + discard”, and “expansion + discard” in that order. The processing method “expand + discard” is also depicted. “Expansion” represents the use of image processing techniques to expand the number of images in a relatively small number of categories in the dataset, and “discard” represents the deletion of images in the dataset where the percentage of pixels of the

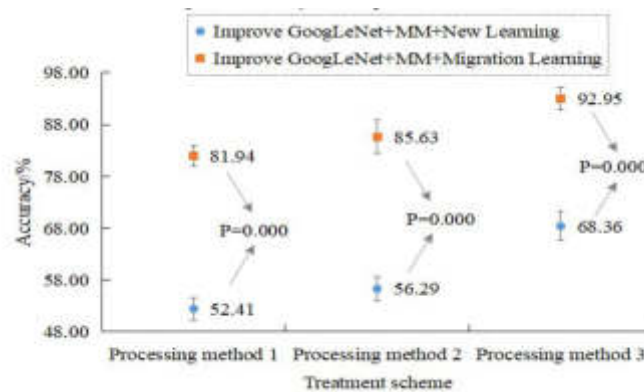


Fig. 4.1: Comparison of recognition accuracy of improved GoogLe Net under new learning and migration learning models

target plant organ is too small. The vertical axis of Figure 4.1 represents the image recognition accuracy of each model, and the different color labels represent different model training schemes. It is noted that to reduce the effect of experimental random error, each experimental scheme was repeated 20 times. The metrics of the type of measurement were presented in the form of mean \pm standard deviation. The difference between groups was verified using the T difference significance test, with the level of significance set at 0.05. Observing Figure 4.1, it can be seen that the accuracy of the training scheme with the fused migratory learning is higher under the condition of using the same dataset pre-processing scheme. For example, the recognition accuracy of the migration learning scheme for treatment schemes 1 to 3 is 29.53, 29.34, and 24.59 percentage points higher than that of the brand-new learning scheme, respectively. From the perspective of processing schemes, the model using scheme 3 achieves the highest accuracy in comparison to the other conditions.

Considering the results of the study in Figure 4.2, the datasets of all model schemes of the subsequent experiments need to be pre-processed in the way of processing scheme 3, and all of them are trained with the migration learning method. The results of Experiment #02 were used to generate Figure 6, where the horizontal axis represents the model scenarios formed by different organization methods, which are explained in Table 4.2, and the vertical axis represents the image recognition accuracy of each model. As we can see in Figure 4.2, the recognition accuracy of the “Improve GoogLe Net+MM” organization scheme is the highest, with a mean value of 92.95%, while the accuracy of the models trained with whole plant images or mixed all images is lower, with a mean value of 66.38% and 71.42%, respectively. It indicates that the multi-cue model combining each major plant organ and the overall image, outperforms a single model in terms of recognition performance.

Considering the results of the study in Figure 4.2, it is necessary for all algorithms in the following experiments to create multi-cue models in order to take part in the study. The results of Experiment #03 were used to create Figure 4.2, where the horizontal axis represents the recognition models built based on each algorithm, the left vertical axis represents the recognition accuracy, and the right vertical axis represents the difference between the recognition accuracy of each algorithm model and the “Improve GoogLe Net+MM” model designed in this study, in %. As can be seen in Figure 4.3, the recognition accuracy of the “Improve GoogLe Net+MM” model designed in this study is still significantly higher than the other comparable models. The “Faster-RCNN+MM” model follows with the second-highest accuracy rate.

The average accuracy of the MM model is 90.51%, which is only 2.44 percentage points lower than that of the previous model. It shows that the recognition performance of the algorithm model designed in this study is already better than the Faster-RCNN model which has a better recognition effect in the market.

The results of Experiment #04 were used to create Figure 4.3. The graph displays the amount of time spent on computing (vertical axis) versus the number of images to be identified (horizontal axis). These images were sourced from the ImageNet dataset. As we can see in Figure 4.4, the “Improve GoogLe Net+MM” model

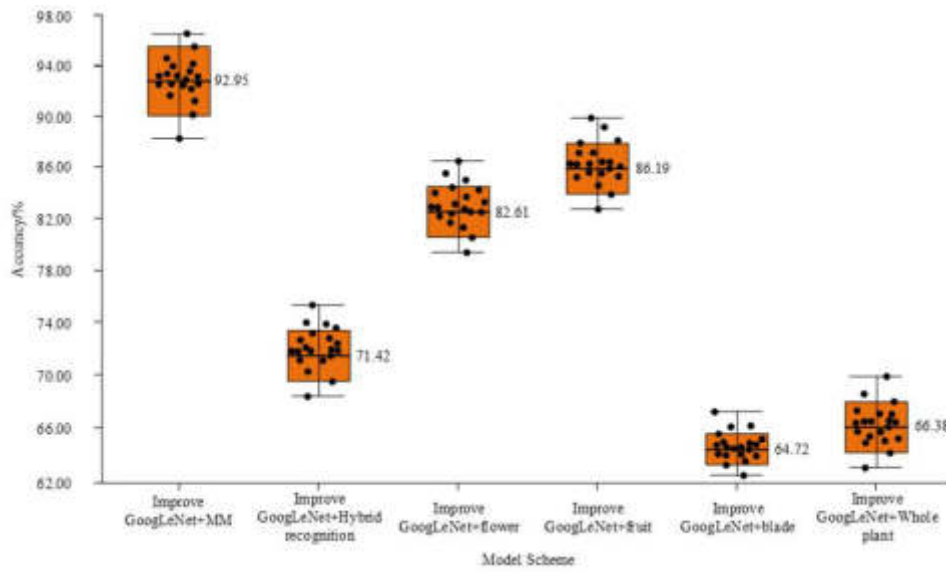


Fig. 4.2: Comparison of recognition accuracy of different model organization methods

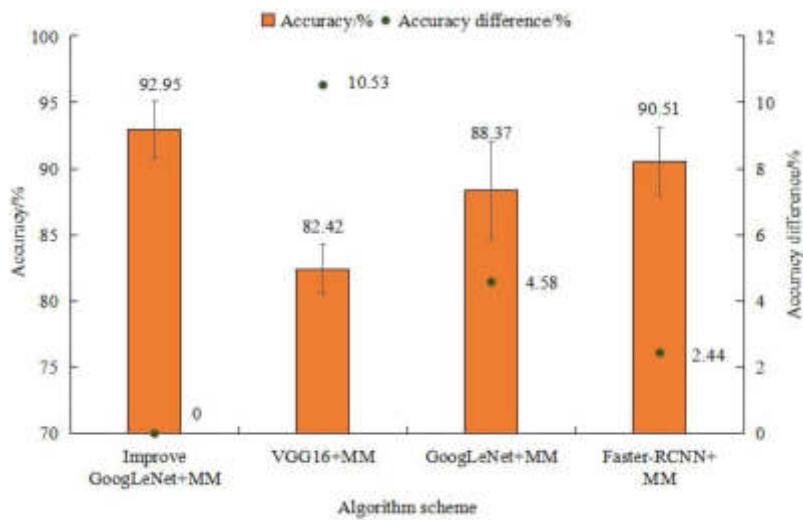


Fig. 4.3: Comparison of recognition accuracy of each algorithm model

is lower than the other three comparison models as the number of images to be recognized increases, while the computation time of the model built by the unimproved GoogLe Net algorithm is higher, but the computation time of the “VGG16+MM” model is lower than the other three comparison models when the sample size is larger. The “VGG16+MM” model requires the longest computation time of 37.86 seconds.

To further analyze the reliability and application value of this design model, 80 landscape plant experts from both domestic and foreign sources are now invited to conduct an evaluation experiment. Obtain a total of 762 real landscape plant images from a domestic science and technology innovation park, and use this design model and comparison model for landscape plant recognition. And have experts rate the recognition results of

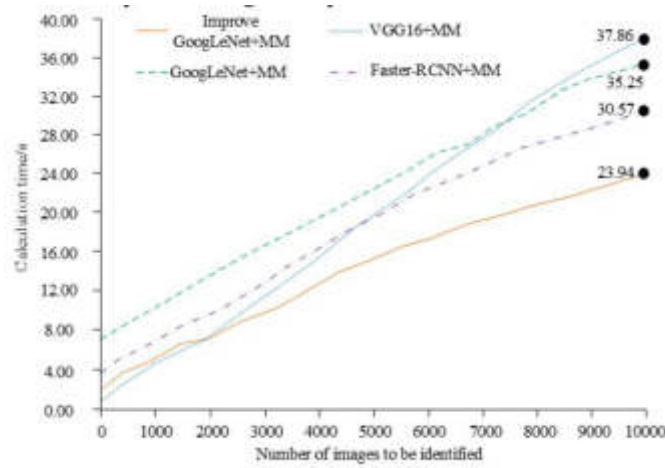


Fig. 4.4: Comparison of computational time consumption of each algorithm model

Table 4.4: Statistics of Evaluation Experiment Scoring Results

Identification	Average value	Standard deviation	Maximum value	Minimum value
Improve GoogLeNet+MM	9.18	0.43	9.77	8.81
GoogLeNet+MM	8.24	0.52	9.05	7.63
VGG16+MM	7.65	0.64	6.89	8.30
Faster-RCNN+MM	8.69	1.09	9.62	6.84

each model on a 10 point scale, and the scoring results are shown in Table 4.4. The scoring results, shown in Table 4.4, indicate that the overall average of the Improved GoogLeNet+MM recognition model designed in this study is the highest, with the most stable scoring results. The overall rating data for the Faster RCNN+MM recognition model is only lower than the former, but the stability of the rating results is the worst because the latter has the largest standard deviation of 1.09 among all models.

From the results of the above data experiments and evaluation experiments, it can be seen that the model proposed in this study exhibits favorable application outcomes. The communication between the experimental members and the expert group members in the evaluation experiment found that the expert team believes that the recognition model designed in this study better solves the problem of insufficient accuracy in identifying scarce categories in traditional landscape plant recognition models within science and technology innovation parks. This is mainly because the model designed in this study uses transfer learning to assist in training the model, enabling the model to obtain more effective local feature information that exists in different plant images.

Finally, the study will separately analyze the limitations of the design model. Firstly, the biggest limitation of this study is the inability to deploy the designed model into the plant management system for testing its effectiveness in a more realistic usage environment. Secondly, the evaluation experiment part of this study only invited experts to participate, and did not test the evaluation and application attitude of ordinary people to this model. These shortcomings will be addressed and improved in subsequent research.

5. Conclusion. This study focuses on the recognition accuracy and speed issues of traditional landscape vegetation intelligent recognition models. The GoogLeNet algorithm is improved and a landscape plant recognition system for science and technology innovation parks is designed. The experimental results show that the model’s recognition accuracy is significantly improved by pre-processing the training data set with “expand+drop” and training the model with transfer learning. The recognition model constructed with multiple

cues achieved an average classification accuracy of 92.95%, which surpasses that of all single models significantly. Compared with other models based on different algorithms but built in the same way, the recognition accuracy of the “Improve GoogLe Net+MM” model designed in this study is 92.95%, which is higher than all the comparison models. The “Faster-RCNN +MM” model follows with an average accuracy of 90.51%. Moreover, the computation time of the “Improve GoogLe Net + MM” model is lower than other models when processing larger data. The research data show that the improved GoogLe Net algorithm can improve the accuracy of plant recognition. However, due to limited research energy, this study was unable to combine the designed recognition system with the plant intelligent management system to analyze its application value. Subsequent research can develop real-time monitoring and feedback systems to continuously monitor the health status of plants and provide timely feedback. Real time data is fed back to the factory’s intelligent management system for real-time decision-making and adjustment through sensor networks, cameras, or other IoT devices. Utilize machine learning and intelligent algorithms to deeply integrate plant recognition systems and factory intelligent management systems. By continuously learning and optimizing, the system can adapt to different environments and plant species, and provide more accurate predictions and recommendations.

REFERENCES

- [1] A. R. Basharat, X. Ning, and X. Liu, “EnvCNN: A Convolutional Neural Network Model for Evaluating Isotopic Envelopes in Top-Down Mass-Spectral Deconvolution,” *Analytical Chemistry*, vol. 92, no. 11, pp. 7778–7785, 2020.
- [2] X. Chen, S. Wang, H. Liu, et al., “Coal Gangue Recognition using Multichannel Auditory Spectrogram of Hydraulic Support Sound in Convolutional Neural Network,” *Measurement Science and Technology*, vol. 33, no. 1, pp. 015107–015116, 2022.
- [3] J. Y. Choi and B. Lee, “Ensemble of Deep Convolutional Neural Networks with Gabor Face Representations for Face Recognition,” *IEEE Transactions on Image Processing*, vol. 29, no. 3, pp. 3270–3281, 2020.
- [4] D. J. Hwang, S. Choi, J. Ko, et al., “Distinguishing Retinal Angiomatic Proliferation from Polypoidal Choroidal Vasculopathy with a Deep Neural Network Based on Optical Coherence Tomography,” *Scientific Reports*, vol. 11, no. 1, pp. 1–12, 2021.
- [5] O. Ken, J. Liu, and H. Zoltán, “TNG: effect of baryonic processes on weak lensing with IllustrisTNG simulations,” *Monthly Notices of the Royal Astronomical Society*, vol. 502, no. 4, pp. 5593–5602, 2021.
- [6] Z. Li, J. Fang, S. Wang, et al., “Adapt-Kcr: a novel deep learning framework for accurate prediction of lysine crotonylation sites based on learning embedding features and attention architecture,” *Briefings in Bioinformatics*, vol. 23, no. 2, pp. 1467–1469, 2022.
- [7] T. Liu, J. Wei, H. Zhu, et al., “Detection of Optic Disc Abnormalities in Color Fundus Photographs Using Deep Learning,” *Journal of Neuro-Ophthalmology*, vol. 41, no. 3, pp. 368–374, 2021.
- [8] B. Lla, C. Sc, Xz. D, et al., “Deep Convolutional Neural Network for Accurate Segmentation and Quantification of White Matter Hyperintensities,” *Neurocomputing*, vol. 384, no. 7, pp. 231–242, 2020.
- [9] C. S. Maior, J. M. Santana, I. D. Lins, et al., “Convolutional Neural Network Model Based on Radiological Images to Support COVID-19 Diagnosis: Evaluating Database Biases,” *PLoS ONE*, vol. 16, no. 3, pp. e0247839, 2021.
- [10] N. B. Marya, P. D. Powers, L. Fujii-Lau, et al., “Application of Artificial Intelligence using a Novel EUS-based Convolutional Neural Network Model to Identify and Distinguish Benign and Malignant Hepatic Masses,” *Gastrointestinal Endoscopy*, vol. 93, no. 5, pp. 1121–1130, 2020.
- [11] K. B. Meena and V. Tyagi, “Distinguishing Computer-Generated Images from Photographic Images using Two-Stream Convolutional Neural Network,” *Applied Soft Computing*, vol. 100, no. 4, pp. 107025–107033, 2021.
- [12] H. Patel and K. P. Upla, “DepthFuseNet: An Approach for Fusion of Thermal and Visible Images using a Convolutional Neural Network,” *Optical Engineering*, vol. 60, no. 1, pp. 013104–013132, 2021.
- [13] X. Tan, K. Li, J. Zhang, et al., “Automatic Model for Cervical Cancer Screening Based on Convolutional Neural Network: A Retrospective, Multicohort, Multicenter Study,” *Cancer Cell International*, vol. 21, no. 1, pp. 1–10, 2021.
- [14] Y. Wan, X. Wang, Q. Chen, et al., “A Disease Category Feature Database Construction Method of Brain Image based on Deep Convolutional Neural Network,” *PLoS ONE*, vol. 15, no. 6, pp. e0232791, 2020.
- [15] B. Yu, L. Xu, J. Peng, et al., “Global Chlorophyll-a Concentration Estimation from Moderate Resolution Imaging Spectroradiometer using Convolutional Neural Networks,” *Journal of Applied Remote Sensing*, vol. 14, no. 3, pp. 15–23, 2020.

Edited by: Zhengyi Chai

Special issue on: Data-Driven Optimization Algorithms for Sustainable and Smart City

Received: Nov 16, 2024

Accepted: Jun 26, 2024



PERFORMANCE ANALYSIS OF SMART CITY LANDSCAPE DESIGN AND PLANNING BASED ON THE INTERNET OF THINGS

CHAO KANG*, YANTING HE† AND JINGJING XU‡

Abstract. With the continuous development of the IoT technology, the concept of smart city has gradually become one of the key elements of urban planning and design. The purpose of this study is to explore the smart urban landscape design and planning based on the IoT, and to provide new perspectives and methods for the future urban development through in-depth research and analysis of related issues. First, this study reviews the relevance of smart city and landscape design, highlighting the importance of IoT technology in urban planning. Secondly, through case analysis and field trips, the successful experiences and challenges of smart cities that have implemented the IoT technology are analysed. In terms of design and planning, this study proposes a landscape design framework integrating IoT technology, emphasizing the interaction between urban landscape and information technology to promote sustainable urban development. Finally, this study summarizes the relevant findings, and prospects the future development trend of smart urban landscape design planning, to provide useful guidance and inspiration for urban planners and designers. Through this study, theoretical support and practical experience can be provided for building a more intelligent and liveable urban environment.

Key words: IoT, Smart city, Landscape design and planning, Sustainable development

1. Introduction. The Internet of Things (IoT) refers to the use of information sensing devices to connect any object to a network according to agreed protocols. Objects exchange and communicate information through information dissemination media to achieve intelligent recognition, positioning, tracking, supervision and other functions. With the continuous development of science and technology and the rapid advancement of urbanization process, smart city has become the frontier topic of contemporary urban planning and design. Among them, the booming development of the IoT technology provides new possibilities for the innovation of urban space [1, 2]. The purpose of this study is to deeply explore the smart urban landscape design and planning based on the IoT, to integrate advanced technologies in the urban development, optimize the spatial layout, and improve the living quality of residents.

Modern cities face many complex and urgent challenges, such as traffic congestion, limited resources, environmental pollution and so on. These problems require not only innovative solutions, but also interdisciplinary cooperation and comprehensive thinking. The IoT technology, with its characteristics of temporal data collection and intelligent decision-making, provides urban planners with an unprecedented means to meet these challenges. Therefore, it is necessary to deeply study the application of the IoT in smart urban landscape design planning to explore its great potential in urban management.

The significance of the IoT lies in its ability to connect the physical and digital worlds, achieve real-time collection, transmission, and processing of information, improve people's quality of life, promote innovation and development in various industries, and contribute to the rational utilization of resources and sustainable development. As an important part of urban planning, landscape design is not only related to the aesthetic feeling of urban appearance, but also related to the quality of residents life and the sustainable development of the city. Through the integration of the IoT technology, landscape design can achieve more refined and intelligent planning, making the urban space more adapt to the needs of residents and lifestyle [3, 4]. In the face of the growing trend of urbanization and complex urban challenges, wisdom urban landscape design planning how to better use of the IoT technology, to improve the urban sustainability, liability, and intelligence,

*Fujian CIECC Planning & Design Research Group Co., Ltd, Information Research Center, Zhangzhou 363000, China

†North Minzu University, School of Civil Engineering, Department of civil engineering, Yinchuan, 750000, China (HeYanting999@163.com).

‡Zhangzhou Engineering Consulting Center Co., Ltd., Consulting Office, Zhangzhou 363000, China.

and how to balance in the digital age technology innovation and cultural heritage, to create a humanistic care of urban landscape is facing the problem of [5, 6].

This study will further expand the understanding of the application of the IoT technology in the smart urban landscape design and planning through in-depth analysis and extensive application of various research methods. The literature review systematically combs the application process of the IoT technology in urban planning and landscape design, and provides a theoretical basis for the research. The case analysis will dig deep into the existing practical experience, draw inspiration and lessons from the successful cases, and lay a practical foundation for putting forward the innovative design framework. Expert interviews will bring together in-depth insights from urban planning experts, landscape architects, and IoT professionals to provide a more comprehensive understanding of the needs and challenges in different areas.

The application of mathematical modelling and simulation technology will help to build the influence model of IoT technology in urban landscape design, and provide more specific data support for research. An extensive questionnaire will cover urban residents, planning practitioners, and design professionals to provide a more comprehensive basis for the proposed innovative design framework and guiding principles by collecting diverse perspectives and needs.

This study aims to provide more practical and feasible guidelines for the application of Internet of Things technology in the fields of urban planning and landscape design. By deeply analysing the current state of technology, evaluating its potential impact on urban sustainability, and exploring the integration of technology and traditional aesthetics, we provide more comprehensive decision-making support for urban decision-makers, planners, and designers, and promote the practical development of smart cities. To provide innovative ideas and practical tools for future urban planners and designers, leading cities towards a more intelligent, livable, and sustainable future.

2. Overview of Smart City Landscape Design.

2.1. Overview of smart city development. With the rapid advancements in science and technology, the concept of a smart city has emerged as a pivotal aspect of modern urban planning. This notion emphasizes the seamless integration of information and communication technologies into cities, aiming to enhance urban operational efficiency, optimize resource utilization, and elevate the quality of life for residents [7, 8]. The ascendance of the IoT technology is a significant impetus for the progress of smart cities. Connecting sensing devices to the Internet enables real-time monitoring and data exchange across diverse urban domains.

Within the ambit of the smart city paradigm, urban infrastructure has seamlessly transitioned into the intelligent era. The traffic system now optimizes traffic flow, mitigating congestion, while the intelligent energy system effectively integrates energy resources to boost energy efficiency. Furthermore, environmental monitoring technology assists cities in achieving precise environmental management and safeguarding natural ecology [9]. Extensive data analysis holds a pivotal position in the bright city landscape. By mining vast data, city managers gain a deeper, more nuanced understanding of residents' behaviours, urban operational conditions, and the root causes of challenges. This data-driven decision-making approach renders urban planning more scientific, agile, and responsive to evolving urban needs.

The development of smart cities is not merely tethered to technological innovations; it also encapsulates the enhancement of the living environment. Residents can now effortlessly access medical care, education, cultural, and entertainment services through intelligent offerings, enhancing the city's attractiveness and quality of life. Additionally, the advancement of intelligent cities necessitates seamless collaboration between governments, enterprises, and various sectors of society to facilitate the widespread adoption of innovative technologies and steer cities toward a brighter, more sustainable future. Although the development of smart cities provides many advantages for urban management and residents, it is also accompanied by a series of challenges, including data privacy protection, and network security risks. Fig. 2.1 shows the architecture diagram of the smart city grid management platform. The future smart city development needs to pay attention to the balance of ethics and social issues while making technological innovation, to ensure the comprehensive and sustainable urban development.

2.2. Application of IoT technology in urban planning and Landscape design. With the continuous evolution of the IoT technology, its application in urban planning and landscape design presents a promising

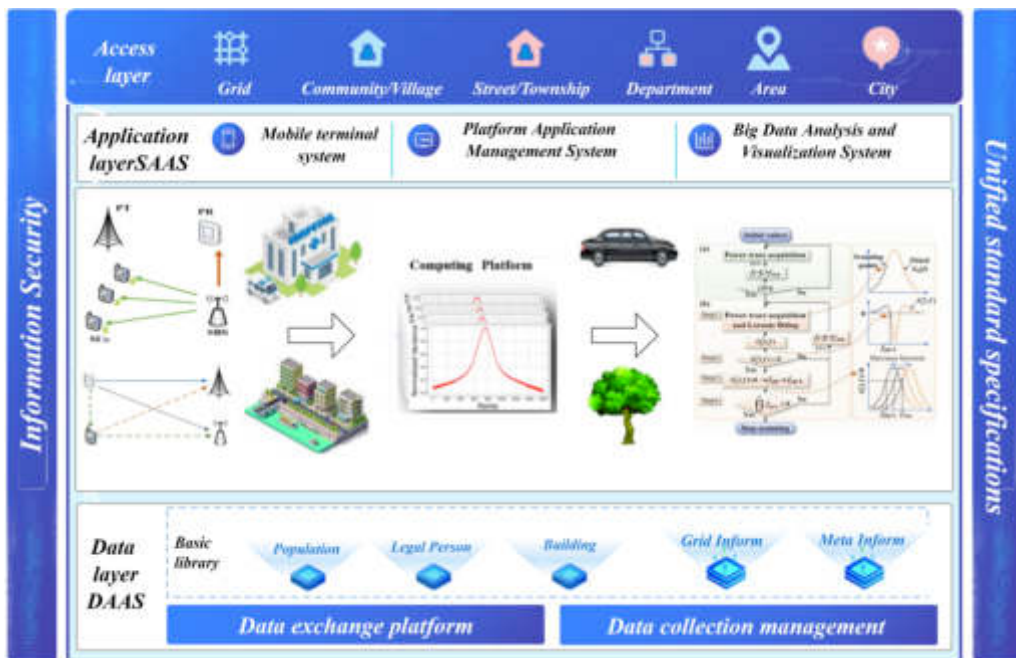


Fig. 2.1: Deep learning applied to short-Term traffic models

prospect [10, 11].

First, the real-time data monitoring of the IoT provides an accurate information foundation for urban planning. Through large-scale sensor networks, city managers can obtain real-time data on traffic flow, environmental quality, population density, and more. This provides planners with a deep opportunity to understand the operation of cities, making urban planning more scientific and sustainable.

In the field of transportation, the IoT technology has realized the intelligent and optimized [12, 13] of the transportation system. The interconnection of traffic lights, intelligent vehicles and pedestrian devices enables the traffic flow to be monitored and adjusted in real time, minimizing congestion, and improving the road utilization efficiency. This means more efficient and green transportation planning for urban planners.

In landscape design, IoT technology offers the possibility of creating a smarter urban environment. Through sensors that perceive environmental parameters, landscape designers can obtain real-time weather, light, and other information, to adjust the design of public space and improve the ecological friendliness of the city. This smart environment design not only beautifies the city, but also provides a more livable living space for the residents.

The IoT technology also promotes the intelligence of buildings. By connecting various equipment and systems, intelligent buildings realize intelligent control of energy and optimization of security system. In urban planning, this means a more energy-efficient and intelligent building design to create a more sustainable future for the city. In general, the widespread application of the IoT technology makes urban planning and landscape design more intelligent and more efficient. It not only improves the sustainability and livability of the city, but also provides a new impetus for urban innovation and development. This trend will further drive cities to be smart and sustainable.

3. Innovative framework of smart urban landscape design and planning.

3.1. Principle of the IoT technology. The IoT is an advanced technology system dedicated to deeply integrating the physical world with the digital world, achieving this goal by connecting and sharing information [14, 15]. Its basic principle covers several key aspects, among which sensing technology is one of the foundations, including various sensors and detectors, used to collect data of temperature, humidity, light, motion, and other



Fig. 3.1: Full view of the IoT technology

parameters, to transform physical phenomena into digital signals.

Communication technology plays a key role in the IoT, because the large number of devices need to communicate information [16]. Wireless communication technologies, such as Wi-Fi, Bluetooth, Zigbee, and cellular networks, enable devices to deliver data in real time and connect to the Internet. Data processing and storage is another key link. Cloud computing and edge computing technology realize efficient data management by processing and analysing the data obtained from sensing devices and extracting useful information.

Fig. 3.1 shows the full view of the IoT technology. To ensure the secure interaction between devices, the IoT introduces identity authentication technology and security protocol to ensure that only legitimate devices can access the network, and protect the security of data in transmission and storage through encryption [17, 18]. Interoperability is designed to solve the problems caused by diversified devices, platforms, and protocols. Through standardized protocols and interfaces, different devices can communicate with each other to achieve seamless connection.

Considering that many IoT devices rely on limited energy sources, energy management becomes a critical principle. Using low power design, energy recovery and optimized communication protocols can help to extend the service life of equipment [19, 20]. Finally, remote control and self-adaptation technology enable the system to remotely monitor and regulate the equipment, adjust its own behaviour according to environmental changes and needs, and enhance the flexibility and adaptability of the system. These principles cooperate with each other to build the basic framework of the IoT technology, providing a brand-new digital possibility for smart city landscape design and planning.

The core of the IoT technology lies in perception, communication, processing and security, and these principles together build an efficient interconnected network. First, the perception of data is realized through various sensors, as shown in (3.1):

$$Data_{sensed} = Sensor(Physical_world) \tag{3.1}$$

These sensing devices convert information from the physical world into digital signals. Then, through a variety of communication technologies, the device transmits the data to the cloud or other devices, and the communication formula is stated by (3.2):

$$Data_{transmitted} = Communication_Tech(Data_{sensed}) \quad (3.2)$$

After the data reaches its destination, the cloud computing and edge computing technology will process and store it, as shown in (3.3):

$$Processed_Data = Data_Processing(Data_{transmitted}) \quad (3.3)$$

This process usually includes data analysis, model training, etc., to extract useful information. The security of data is critical, so identity authentication and encryption are required. The security (3.4) states:

$$Secure_Data = Security_Protocol(Processed_Data) \quad (3.4)$$

To ensure interoperability between devices, a standardized protocol and interfaces are necessary, and the interoperability formula can be stated by (3.5):

$$Interoperability = Standard_Protocols_and_Interfaces \quad (3.5)$$

Energy management is critical for many IoT devices, especially mobile and low-power devices. The energy management formula is shown in the (3.6):

$$Energy_Management = Low_Power_Design + Energy_Harvesting \quad (3.6)$$

Finally, remote control and adaptation enable the system to adjust its behaviour according to environmental changes and requirements, which can be expressed in (3.7):

$$Adaptation = Adaptive_Algorithms(Changes, Demands) \quad (3.7)$$

These principles work together with each other, working together to build the foundation of the IoT technology, providing strong digital support for urban planning and landscape design.

3.2. Smart city landscape design process based on the IoT. When planning the smart urban landscape design, the demand analysis is the first step to clarify the design objectives, such as improving the quality of life of residents and optimizing the use of urban space in [21, 22]. This stage also covers an in-depth study of the urban development status and challenges, providing comprehensive background information for the design. Subsequently, the critical data collection phase is achieved through the deployment of IoT sensing systems, using environmental sensors, intelligent lighting, and other devices to collect real-time data, including air quality and traffic flow, and conduct comprehensive analysis.

In the design stage, the focus turns to integrating the IoT technology, giving the landscape a higher intelligent level of [23] through elements such as intelligent lighting and intelligent seats. Virtual design tools are used to simulate the layout of landscape elements, consider the influence of people flow, traffic flow and natural conditions, and integrate humanistic factors into the design to achieve a more humanized urban landscape.

Fig. 3.2 shows the design scheme of the IoT scenario architecture. The implementation stage should focus on the deployment of devices and system integration, and deploy the IoT devices and intelligent landscape elements according to the design scheme to ensure that they can work together. Through testing and tuning, the system can run stably in practical application. In the monitoring and management stage, real-time monitoring is realized through the IoT technology to analyse the status of urban environment and landscape elements and provide support for decision-making. At the same time, regular maintenance and update are the key to ensure the long-term operation of the system.

Close communication with community residents and relevant stakeholders is essential throughout the process. By collecting feedback, understand their needs and opinions, and adjust the landscape design according to the actual operation to better meet residents expectations. This interaction and collaboration promote strong connections between the community and designers. Through this process, the smart urban landscape design based on the IoT can flexibly adapt to the development of the city and the needs of residents, and improve the sustainability, liability, and intelligence of the city.

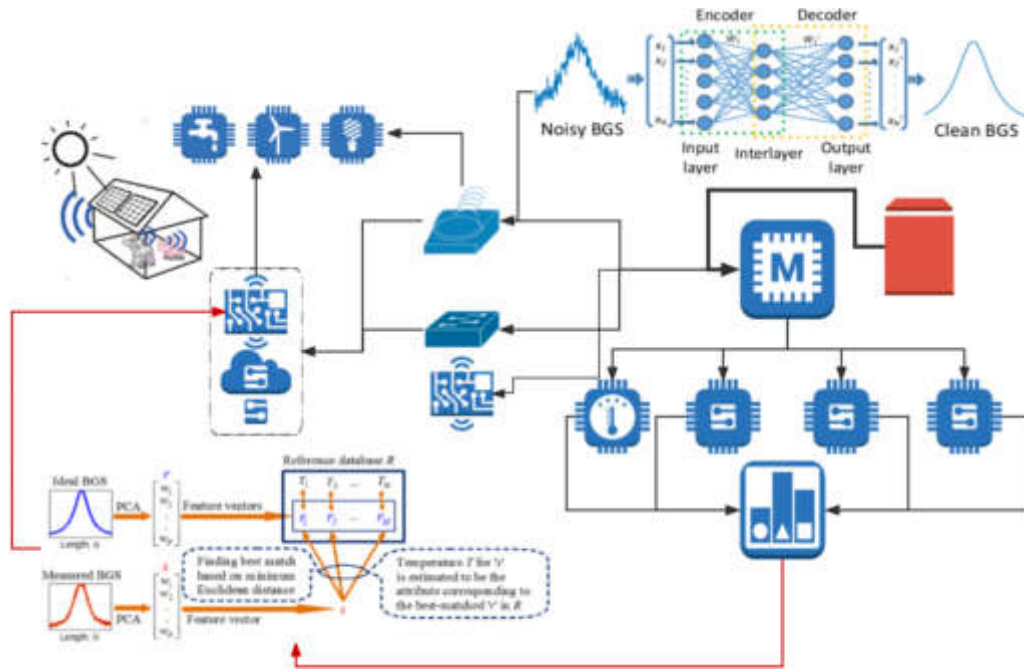


Fig. 3.2: Architecture design scheme of the IoT scenario

3.3. Smart city landscape design and planning system architecture based on the IoT. The architecture of the smart city landscape design and planning system based on the IoT mainly includes the perception layer, the communication layer, data collection and storage layer, data processing and analysis layer, and the application and control layer [24, 25]. In the sensing layer, various sensors and smart devices are used to sense the urban environment and resident activities in real time. The communication layer uses the IoT communication protocol and wireless technology to establish connections between devices. The data collection and storage layer are responsible for collecting, storing environment data, and conducting local storage. The data processing and analysis layer conducts large amounts of data processing and analysis through cloud computing and edge computing technology, and applies machine learning and artificial intelligence to optimize the system performance [26]. In the application and control layer, virtual design tools simulate the layout of landscape elements, and real-time monitoring and control functions enable city managers to respond to urban changes. The application and control layer includes a user interface for urban planners and residents to view data, provide feedback and participate in decision making. The feedback and optimization layer of the whole system is responsible for feedback the analysis results to the city managers and design team to support the decision making. At the same time, optimizing and adjusting the system through feedback information to meet the needs of the city and the residents expectations. Fig. 3.3 shows the application framework design of the IoT. The security and privacy layer ensures the security of the system and the privacy of personal data through identity authentication, encryption, and privacy protection measures. Such a system architecture provides comprehensive management and intelligent control for the smart city landscape design and planning.

4. Results and discussion.

4.1. Evaluation indicators. The smart city landscape design planning based on the IoT involves many aspects, including environmental monitoring, energy utilization, traffic flow, social interaction, etc. The evaluation indicators involve the following formula [27, 28]:

In (4.1) is the environmental monitoring index, where n is the number of the environmental parameters

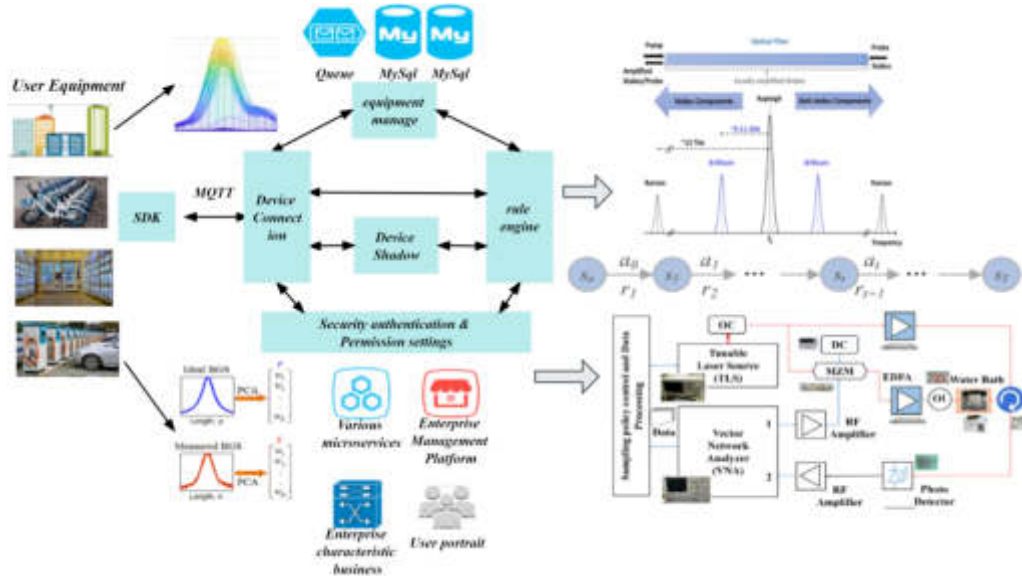


Fig. 3.3: Application frameworks design of the IoT

monitored, including air quality, water quality, noise, etc.

$$EMI = \frac{\sum_{i=1}^n Parameter}{n} \tag{4.1}$$

In (4.2) is the environmental monitoring index, where n is the number of the environmental parameters monitored, including air quality, water quality, noise, etc.

$$SEEE = \frac{LightingEfficiency}{TotalEnergyConsumption} \times 100\% \tag{4.2}$$

The traffic flow optimization index is shown in (4.3), which can help to evaluate the optimization effect of the IoT-based traffic management system on urban traffic flow.

$$TFOI = \frac{OptimizedTrafficFlow}{UnoptimizedTrafficFlow} \times 100\% \tag{4.3}$$

The social interaction activity assessment is conducted as shown in (4.4), and this formula can be used to assess the activity of urban residents on social media, reflecting the level of social interaction.

$$SIAA = \frac{SocialMediaInteractionCount}{TotalCityResidents} \times 100\% \tag{4.4}$$

In (4.5) is the calculation formula of the intelligent energy utilization index, which can be used to evaluate the intelligent utilization degree of energy used by the urban energy management system based on the IoT.

$$ESUI = \frac{SmartEnergyUtilization}{TraditionalEnergyUtilization} \times 100\% \tag{4.5}$$

4.2. Analysis of the experimental results. In the experiment, this study is committed to deeply explore the impact and effect of smart urban landscape design planning based on the urban environment. Through detailed data acquisition and careful analysis of results, a comprehensive understanding of the advantages and

potential problems of the design scheme. First, the data acquisition was performed, and the following methods were used to collect the experimental data.

A sound sensor network built by deploying IoT sensors monitors environmental parameters, including but not limited to air quality, temperature, and humidity, etc. This provides real-time and reliable data for the objective assessment of the actual impact of the design schemes on the urban environment.

An extensive user survey was conducted, and through the questionnaire survey and user feedback, residents views, and experience of the smart city landscape. This combination of qualitative and quantitative approaches provides the basis for an in-depth understanding of key information such as citizen engagement and satisfaction.

Through a detailed analysis of energy use, the actual impact of the design scheme on energy efficiency is evaluated [29, 30]. In terms of the environmental parameter analysis, some key results were found through the in-depth analysis of the sensor network data. The air quality in the experimental area was significantly better than that in the control area, proving the positive effect of the smart city design scheme in improving the environment. In addition, the intelligent control of the IoT devices plays an important role in regulating the temperature and humidity, and improves the living comfort of urban residents. Through the user survey, we learned that most respondents are satisfied with the smart city design, especially in terms of safety, convenience, and environmental friendliness. At the same time, some feedback mentioned the room for improvement, such as improving the efficiency of information transmission and providing more interactive experiences, providing useful suggestions for future designs. In the analysis of energy use effect, it is found that the smart city design scheme has achieved significant energy saving effect in energy use and effectively reduced the urban operation cost. In addition, through intelligent management and optimization, the urban infrastructure is more in line with the principles of sustainable development, providing substantial support for the sustainable development of cities.

Fig. 4.1 presents a comprehensive overview of the IoT system's data statistics, offering a deep dive into the evolving patterns of its pivotal indicators. Over the recent months, the performance of IoT systems has undergone noteworthy advancements. Firstly, the figure underscores a gradual surge in the number of IoT devices connected to the system, with a notable increase of 25% since the beginning of the year. This upward trend signifies the system's enhanced scalability, enabling a broader array of devices to integrate and facilitate data collection and exchange seamlessly. This expansion can be attributed to the seamless integration of new equipment, timely system upgrades, and ongoing optimization efforts. Secondly, the figure reveals fluctuations in the data transmission rate. Specifically, during the third quarter, the average transmission rate increased by 10% compared to the previous quarter. This marked improvement is likely due to system optimizations, including implementing more efficient data routing algorithms and deploying high-capacity network infrastructure. This refinement significantly enhances the system's real-time capabilities and response speed, resulting in a smoother user experience. The data statistics reflect the system's stability and reliability. Over the past six months, the system failure rate has maintained a consistently low level, averaging 0.5% per month. This remarkable consistency indicates that adequate measures have been implemented throughout the system's design, operation, and maintenance, including proactive monitoring, timely patching of vulnerabilities, and robust backup systems.

Fig. 4.2 offers a comprehensive visualization of the data dispersion within the innovative City IoT system, revealing intricate patterns and insights into the system's performance across various dimensions. This map not only underscores the diverse scores achieved by different subsystems but also highlights potential areas of improvement. Upon closer inspection, the figure reveals a spectrum of scores across the various subsystems, reflecting their unique functional characteristics, implementation efficiencies, and degrees of data integration. These disparities are crucial in accurately assessing the system's strengths and weaknesses, enabling decision-makers to formulate targeted improvement strategies. Outliers within the data dispersion map indicate potential system issues or bottlenecks. These outliers, which may arise from equipment malfunctions, network hiccups, or security vulnerabilities, are red flags for urgent attention. Prompt identification and remediation of these outliers are vital in enhancing the system's usability and stability, ensuring seamless operation under all circumstances.

Fig. 4.3 presents the accuracy curve of the innovative city model, a graphical representation that offers a deep dive into the models performance across various conditions. This analysis provides a snapshot of the models capabilities and identifies potential areas for improvement. The curves upward trajectory initially reflects the models learning process during the training and validation phases. As the model encounters and processes

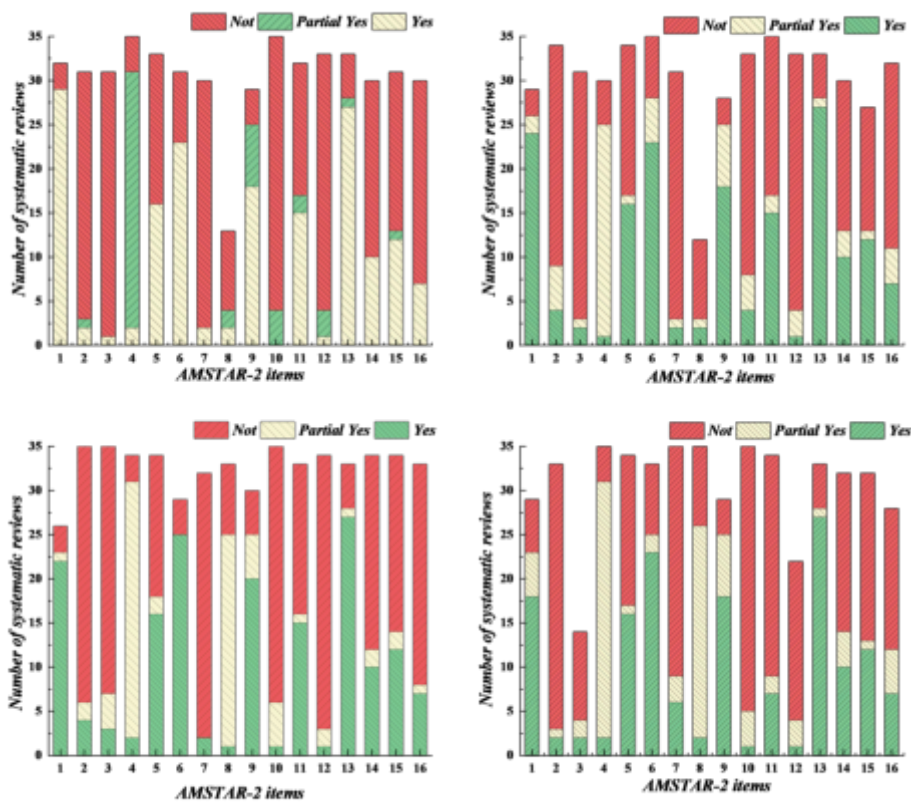


Fig. 4.1: Data statistics of the IoT system

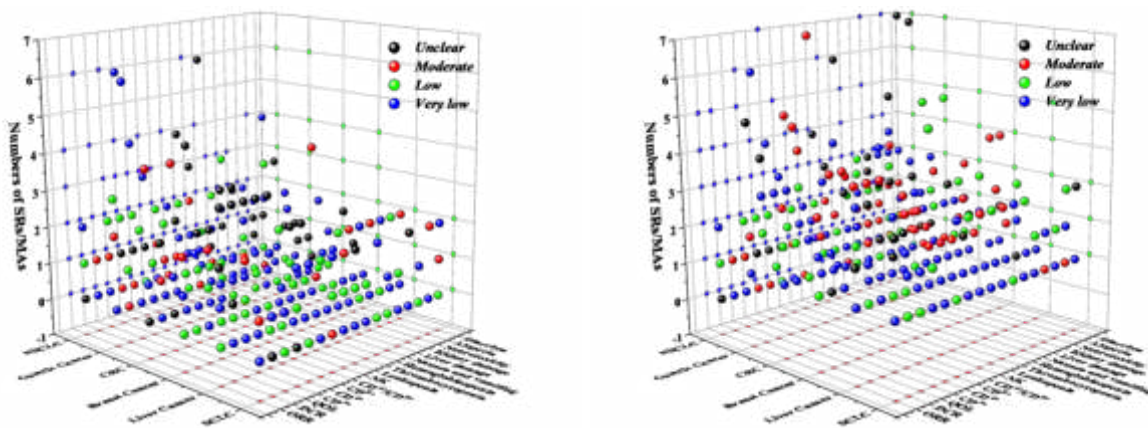


Fig. 4.2: Scale Data Dispersion Map of Smart City IoT System

more data, it gradually learns to recognize patterns and improve its predictive accuracy. The steepness of this initial rise indicates the models learning speed, which is crucial for its ability to adapt quickly to new data and

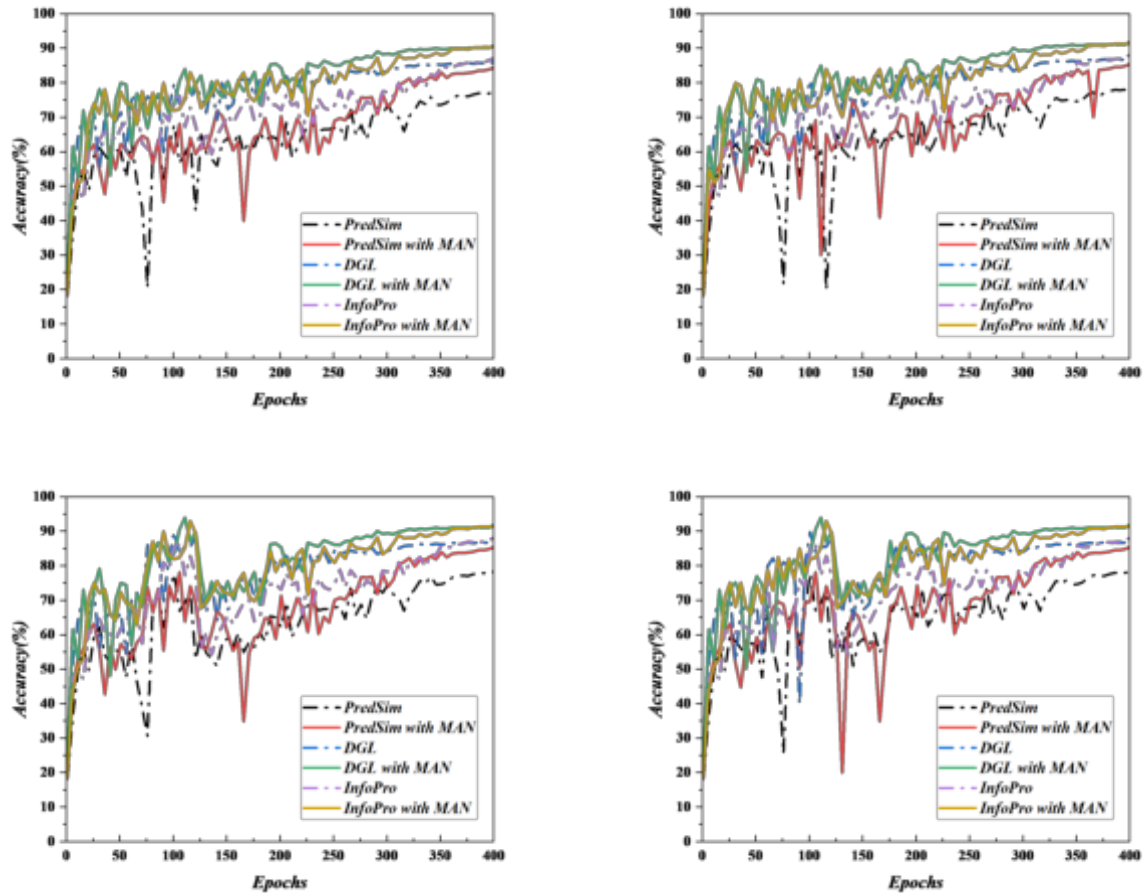


Fig. 4.3: Accuracy Curve of The Smart City Model

scenarios. Once the curve stabilizes, it indicates that the model has reached a level of maturity. The model has demonstrated robust performance on task-specific predictions without overfitting the training data. This stability ensures the model can generalize well to unseen data, a crucial requirement for real-world applications.

However, if the curve experiences a sudden drop during the validation phase, it suggests that the model may be over-fitting in certain aspects. This overfitting could be due to various factors, such as an inappropriate choice of hyper-parameters or an insufficiently diverse training dataset. In such cases, further analysis and adjustment of the model are necessary to improve its generalization performance. Moreover, the fluctuations in the curve provide insights into the models robustness under different conditions. The model's ability to handle various scenarios is paramount in the context of smart cities, where the urban environment is constantly evolving and complex. By analysing these fluctuations, we can understand how the model performs under different challenges and identify potential weaknesses that must be addressed. By leveraging the insights gained from Fig. 4.3, decision-makers and model developers can optimize the intelligent city model effectively. This optimization process involves adjusting hyper-parameters, refining the training dataset, or exploring alternative model architectures to enhance the models performance. The ultimate goal is to ensure that the model functions robustly and efficiently in practical applications, meeting the demands of urban management and optimization.

5. Conclusions. Through the research and analysis of the smart urban landscape design and planning based on the IoT, this study draws the following conclusions from multiple dimensions:

First, in terms of environmental parameter analysis, the air quality of the IoT sensor network and monitoring the environmental parameters is significantly better than that of the experimental area, highlighting the remarkable effect of the smart city design scheme in improving the urban environmental quality. The control of temperature and humidity by intelligent control improves the living comfort of urban residents and provides substantial support for the construction of liveable cities. The user survey results show that residents are generally satisfied with the smart city landscape design, especially in terms of safety, convenience, and environmental friendliness. However, some users make suggestions for improvements, such as improving information transmission efficiency and increased interactive experience, providing a useful reference for future design and planning.

In terms of the energy use effect, the analysis shows that the smart city design scheme has achieved significant energy saving effect in the energy use, and effectively reduced the urban operation cost. This not only reflects the positive contribution of the design scheme to urban sustainable development, but also provides reference and inspiration for similar urban planning. The smart urban landscape design and planning based on the IoT shows obvious advantages in improving environmental quality, improving residents satisfaction, and realizing energy benefits. In future research, we will pay more attention to field research, collect more real and comprehensive data, optimize data processing methods, and improve the accuracy and universality of results.

REFERENCES

- [1] KANG L, *Street architecture landscape design based on Wireless Internet of Things and GIS system*, *Microprocessors and Microsystems*, 2021, 80, 103362.
- [2] GAO C, WANG F, HU X, AND ET AL, *Research on Sustainable Design of Smart Cities Based on the Internet of Things and Ecosystems*, *Sustainability*, 2023, 15(8), 6546.
- [3] JIA A, *Intelligent Garden planning and design based on agricultural internet of things*, *Complexity*, 2021, pp. 1–10.
- [4] JIANG D, *The construction of smart city information system based on the Internet of Things and cloud computing*, *Computer Communications*, 2020, 150, pp. 158–166.
- [5] XUN Y, AND REN G, *Smart Garden Planning and Design Based on the Agricultural Internet of Things*, *Computational Intelligence and Neuroscience*, 2022
- [6] LIU, YU, AND ET AL, *Cognitive digital twins for freight parking management in last mile delivery under smart cities paradigm*, *Computers in Industry*, 2023, 153, 104022.
- [7] YU L W, ZHANG L, AND GONG Z, *An Optimization Model for Landscape Planning and Environmental Design of Smart Cities Based on Big Data Analysis*, *Scientific Programming*, 2022
- [8] LI J, AND WANG Y, *Characteristic analysis and integration method of urban planning data based on GIS of internet of things*, *Sustainable Computing: Informatics and Systems*, 2022, 36, 100801.
- [9] MA X, AND XUE H, *Intelligent smart city parking facility layout optimization based on intelligent IoT analysis*, *Computer Communications*, 2020, 153, pp. 145–151.
- [10] WADE K, VRBKA J, ZHURAVLEVA N A, AND ET AL, *Sustainable governance networks and urban Internet of Things systems in big data-driven smart cities*, *Geopolitics, History, and International Relations*, 2021, 13(1), pp. 64–74.
- [11] XU S, HOU Y, AND MAO L, *Application Analysis of the Ecological Economics Model of Parallel Accumulation Sorting and Dynamic Internet of Things in the Construction of Ecological Smart City*, *Wireless Communications and Mobile Computing*, 2022
- [12] YUANMENG Z, JIE C, HONG Z, AND ET AL, *A deep learning traffic flow prediction framework based on multi-channel graph convolution*, *Transportation Planning and Technology*, 2021, 44(8)
- [13] YANG Z, JIANJUN L, FAQIRI H, AND ET AL, *Green internet of things and big data application in smart cities development*, *Complexity*, 2021, pp. 1–15.
- [14] SINGH D K, SOBTI R, JAIN A, AND ET AL, *LoRa based intelligent soil and weather condition monitoring with internet of things for precision agriculture in smart cities*, *IET Communications*, 2022, 16(5), pp. 604–618.
- [15] CHEN C, ZIYE L, SHAOHUA W, AND ET AL, *Traffic Flow Prediction Based on Deep Learning in Internet of Vehicles*, *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS*, 2021, 22(6)
- [16] PRAKOSO, VELANDANI, AND ET AL, *Research Trends, Topics, and Insights on Network Security and the Internet of Things in Smart Cities*, *Jurnal Studi Ilmu Pemerintahan*, 2023, 4(3), pp. 191–206.
- [17] EL ABD N M, *Smart monitoring solution through internet of things utilization to achieve resilient preservation*, *Ain Shams Engineering Journal*, 2023, 14(6), 102176.
- [18] CHUI K T, ORDÓÑEZ DE PABLOS P, SHEN C, AND ET AL, *Towards Sustainable Smart City via Resilient Internet of Things, Resilience in a Digital Age: Global Challenges in Organisations and Society*, 2022, pp. 117–135.
- [19] YAZDINEJAD A, PARIZI R M, DEGHANTANHA A, AND ET AL, *Enabling drones in the internet of things with decentralized blockchain-based security*, *IEEE Internet of Things Journal*, 2020, 8(8), pp. 6406–6415.
- [20] PANDHARIPANDE A, AND THIJSSSEN P, *Connected Street lighting infrastructure for smart city applications*, *IEEE Internet of Things Magazine*, 2019, 2(2), pp. 32–36.

- [21] RANGARAJAN, SARATHKUMAR, AND TAHSIEN AL-QURAISHI, *Navigating the Future of the Internet of Things: Emerging Trends and Transformative Applications*, Babylonian Journal of Internet of Things, 2023, pp. 8–12.
- [22] LV X, AND LI M, *Application and research of the intelligent management system based on internet of things technology in the era of big data*, Mobile Information Systems, 2021, pp. 1–6.
- [23] JIA A, AND XU C, *Smart city image landscape design based on wireless sensors*, Microprocessors and Microsystems, 2021, 83, 104022.
- [24] HUANG Y, PENG H, SOFI M, AND ET AL, *The city management based on smart information system using digital technologies in China*, IET Smart Cities, 2022, 4(3), pp. 160–174.
- [25] PRIYANKA E B, AND THANGAVEL S, *Influence of Internet of Things (IoT) In Association of Data Mining Towards the Development Smart Cities-A Review Analysis*, Journal of Engineering Science & Technology Review, 2020, 13(4).
- [26] HUANG W, ZHANG Y, AND ZENG W, *Development and application of digital twin technology for integrated regional energy systems in smart cities*, Sustainable Computing: Informatics and Systems, 2022, 36, 100781.
- [27] NITOSLAWSKI S A, GALLE N J, VAN DEN BOSCH C K, AND ET AL, *Smarter ecosystems for smarter cities? A review of trends, technologies, and turning points for smart urban forestry*, Sustainable Cities and Society, 2019, 51, 101770.
- [28] ROSA L, SILVA F, AND ANALIDE C, *Mobile networks, and Internet of Things infrastructures to characterize smart human mobility*, Smart Cities, 2021, 4(2), pp. 894–918.
- [29] KAMRUZZAMAN M M, *Key technologies, applications, and trends of internet of things for energy-efficient 6G wireless communication in smart cities*, Energies, 2022, 15(15), 5608.
- [30] SYED A S, SIERRA-SOSA D, KUMAR A, AND ET AL, *IoT in smart cities: A survey of technologies, practices, and challenges*, Smart Cities, 2021, 4(2), pp. 429–475.

Edited by: Zhengyi Chai

Special issue on: Data-Driven Optimization Algorithms for Sustainable and Smart City

Received: Dec 8, 2024

Accepted: May 6, 2024



RESEARCH ON DATA-DRIVEN URBAN INTELLIGENT MONITORING AND OLD CITY RECONSTRUCTION

YI WANG*

Abstract. As the global urbanization process is further accelerating, the number of urban people is also steadily increasing. With the continuous growth of the urban population, the use of various functional facilities in the city is gradually becoming saturated, which seriously restricts the development of the city. Therefore, real-time perception and prediction of the operational status of various functional urban facilities, as well as environmental safety monitoring in the city, are of great significance for improving the functionality and livability of the city. In complex urban environments, multi-source data is interrelated, and the noise reduction of multi-source data and the integration of this correlation still face great challenges. At the same time, how to apply urban intelligent monitoring in the reconstruction of old cities is rarely mentioned. Based on the above problems, this paper proposes a data noise reduction model based on a wavelet algorithm and combines it with the Macroscopic Fundamental Diagram (MFD) multi-source data fusion structure proposed in this paper to provide a theoretical basis for the construction of an urban intelligent monitoring model. Then, the author constructed a data-driven urban safety environment monitoring model and a multi-source data-based urban congestion monitoring model, which have good experimental results and certain practical value. At the end of the paper, the author briefly discussed the application of the smart city monitoring model in the reconstruction of old cities, hoping to provide certain guidance for further integration in the future.

Key words: Data-Driven, MFD construction, Multi-Source Data, Old city reconstruction, Urban intelligent monitoring

1. Introduction. Urban road networks are an essential component of urban infrastructure that directly affects people's everyday life. Urban road networks are under rising load strain due to the city's population and vehicle count, which frequently causes traffic jams and accidents that seriously impair urban development. Therefore, real-time perception and prediction of the operational status of urban road networks, as well as the identification of high-risk areas for traffic accidents in urban geographical space, are of paramount importance for enhancing the operational efficiency and traffic safety of urban road networks. This paper will take this as the research goal to solve the intelligent detection and recognition of urban road network.

The selection of traffic state evaluation indicators exhibits considerable variability across different scenarios. Recognizing the importance of data fusion between these diverse indicators, the concept of data fusion emerged in the late 20th century, driven by the rapid development of information technology [10]. Notably, Choi and Chung's [1] 2001 study on travel time using fusion coil sensor and GPS data laid the foundation for subsequent traffic flow data fusion. Yang Zhaosheng et al. [2] were pioneers in China, utilizing data fusion to meet information accuracy requirements in the ATMS subsystem of intelligent transportation systems. Henry [3] and others emphasized the real-time dynamics and data quality requirements for data fusion, playing a pivotal role in advancing the application of this technology in transportation. Xu Tao et al. [4] integrated urban road information, eliminating "information islands" and enhancing traffic flow parameters through Bayesian estimation fusion, ultimately identifying road states with fuzzy logic. Ding Yue et al. [5] proposed a multi-source relational data fusion framework, comprising pattern matching, entity alignment, and entity fusion, facilitating rapid pattern matching and providing a unified data view for analysis. Addressing the challenge of identifying traffic network conditions, common indicators include average speed, travel time, and travel delay. Ehrlich categorized traffic status indicators into time and distance, concluding that time-based indicators more accurately reflect travelers' perceptions [6]. Washburn used video data to calculate traffic flow density, evaluating road traffic conditions based on subjective traveler experiences [7]. Taylor et al. established a congestion coefficient model, incorporating GIS technology to build a congestion information system [8]. Kerner employed floating car data to establish a threshold model for evaluating congested sections [9]. The average

*Eurasia Art and Design School, Xian Eurasia University, XiAn, Shaanxi Province, 710065, China (Yi1Wang@outlook.com)

speed of a road section, known for simplicity and high reliability, stands out as one of the most commonly used indicators for traffic operation status evaluation [10,11]. Jiang Tao applied pattern recognition to classify traffic network states and predict future states based on current traffic conditions [12]. Sun Ya used data mining to extract new traffic status information from extensive data, achieving successful classification results through real-time traffic flow data collection [13]. Wang Meihong proposed a method to calculate regional traffic congestion correlation based on spatio-temporal association rules [14]. Scholars have increasingly focused on the Macroscopic Fundamental Diagram (MFD) method for discriminating road network traffic operation status. Xu et al. derived optimal cumulative intervals and corresponding average traffic flow density states, categorizing traffic operation into free flow, optimal accumulation, and congestion [72]. Liu and Xu proposed using the standard deviation of the number of vehicles to supplement traffic flow or vehicle accumulation in traffic operation state division [73]. Daganzo and Gayah identified branch points in road networks, leading to multi-valued and unpredictable MFDs when exceeding certain densities, categorizing networks into stable and unstable states [74]. Haddad Geroliminis explained the generation of branch points, theoretically deducing equilibrium points under a dual system and proposing calculated and unstable boundaries [75]. Aboudolas et al. analyzed traffic ladder diagrams for urban road networks, differentiating between unsaturated, incomplete saturated, supersaturated, and deadlock states, recommending different control strategies for each [76]. With an increasing number of vehicles equipped with location devices, a considerable amount of vehicle Global Positioning System (GPS) data has emerged, leading to the gradual rise of research on road traffic state discrimination based on floating car GPS data. For example, based on massive GPS data analysis of residents' daily travel patterns, the regularities of road traffic operational status can be identified, including the recognition of urban hotspots for travel [20][21]. Lin [22] proposed a combination of spectral clustering techniques and branches based on MFD constructed from floating vehicle data Road network state identification method based on vector machine algorithm.

Urban road traffic is a complex network system; its state is challenging to summarize by a single parameter and often needs to combine a variety of traffic flow parameters and system methods to identify. However, the current research is usually carried out for a single parameter, and only some are carried out for multi-source data fusion. At the same time, noise processing of multi-source data is also a big challenge, and there are few scholars in this area of algorithm research. This paper mainly focuses on the research of multi-source data fusion and data noise reduction algorithm, and builds an intelligent monitoring model of urban road network based on this. In response to these issues, this paper proposes a data denoising model based on the wavelet algorithm and combines it with our proposed MFD multi-source data fusion construction, providing a theoretical foundation for building an intelligent urban traffic monitoring model. We then create a multi-source data-based urban traffic congestion monitoring model and a data-driven urban traffic safety environment monitoring model. Experimental results demonstrate that these models have good performance and practical value. In conclusion, we briefly discuss the application of the urban traffic intelligent monitoring model in the renovation of old city road networks, aiming to provide some guidance for future research.

2. Research on data processing in complex urban environment.

2.1. Research on data noise reduction model. In the field of traffic detection, more and more high-speed and high-precision sensors are being used due to the growing complexity of transportation networks. Traffic detector fault identification and prompt data rectification are critical in order to prevent sensor malfunctions that might jeopardize system safety and result in financial losses [23]. Taking loop detectors as an example, a single intersection may require the use of dozens or even more loop detectors to obtain more accurate traffic parameters such as traffic flow, vehicle speed, and occupancy. However, due to the variability of the natural environment and the inconsistency of the hardware and software attributes of front-end sensors, the rapid growth of data leads to a simultaneous increase in random noise and faulty data. Some fault information is masked by noise, and the distribution of faults becomes more multiscale. This renders traditional fault diagnosis methods based on analytical models inadequate to meet the safety requirements of current transportation systems. Therefore, how to use the accumulated large amount of offline traffic data, extract data features, analyze their intrinsic patterns, monitor online traffic data in real-time and effectively, and accomplish fault detection and diagnosis has become one of the hot topics of concern for many experts and scholars. Wavelet analysis is used to obtain the multiscale properties of so-called fault information, which might occur in many frequency bands.

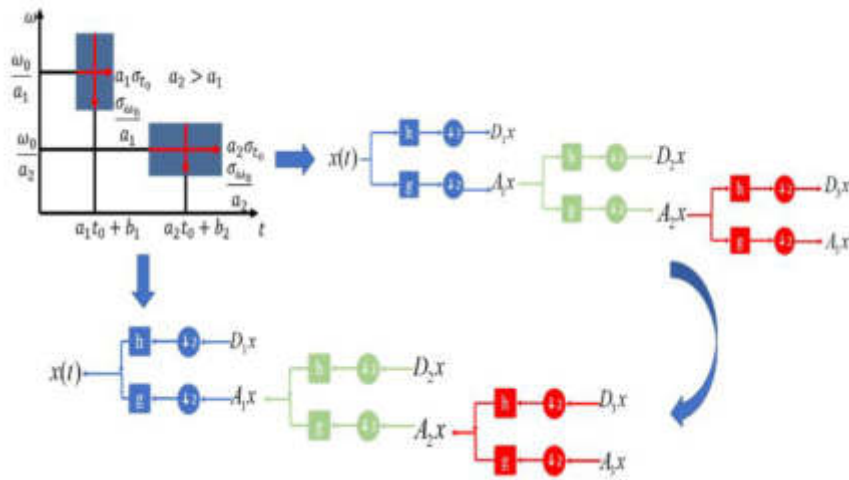


Fig. 2.1: Time-frequency coordinates of wavelet transform and Mallat algorithm

Let $f(t)$ be a finite energy signal; the discrete wavelet transform of this signal can be defined as:

$$(W_{\psi} f)(a, b) = \langle f, \psi_{a,b} \rangle = |a_0|^{-m/2} \int_{-\infty}^{+\infty} f(t) \bar{\psi}(t) dt \tag{2.1}$$

Where $\psi_{a,b}$ is called the generating or basis function of the wavelet transform,

$$\psi_{a,b}(t) = a^{-1/2} \psi\left(\frac{t-b}{a}\right), a > 0, b \in R \tag{2.2}$$

The displacement factor is denoted by 'b' in equation (2.1), while the scale factor is represented by 'a'. As illustrated in Figure 1, a time-frequency coordinate system is established for wavelet transformation. In wavelet transformation, the position of the time window is solely influenced by the displacement factor. Therefore, as the scale factor increases, the time window widens, the frequency window narrows, and the center of the frequency window shifts towards the low-frequency direction. Conversely, with a narrower time window, the frequency window widens, and the center of the frequency window moves towards the high-frequency direction. The essence of wavelet transformation lies in manipulating these two factors to construct a combination that can represent any signal in space [24]. By utilizing the scale factor, it is possible to perform a tower-like decomposition of a specific signal in space, as depicted in Figure 2.1, following the classical Mallat algorithm [25]. This algorithm provides a computational method for wavelet decomposition and reconstruction, simplifying the overall wavelet calculations.

There is an impulse response function in the Mallat algorithm: $h(n)$. As a result, the following definitions for the wavelet and scale functions are given:

$$\begin{cases} \varphi(t) = \sum_n h(n) \varphi(2t - n) \\ \psi(t) = \sum_n g(n) \psi(2t - n) \end{cases} \tag{2.3}$$

In formula (2.3), $g(n) = (-1)^{1-n} h(1-n)$, The signal $x(t)$ is decomposed by Mallat, and the scale is set to $j(j \geq 1)$. Approximate signal and detailed signal obtained by decomposition are respectively:

$$\begin{cases} A_j x(t) = \langle x(t), \varphi_{j,k}(t) \rangle = 2^{-j/2} \int x(t) \varphi(2^{-j}t - 2k) dt \\ D_j x(t) = \langle x(t), \psi_{j,k}(t) \rangle = 2^{-j/2} \int x(t) \psi(2^{-j}t - 2k) dt \end{cases} \tag{2.4}$$

It can be found in equation (2.4) that the process of decomposing signal $x(t)$ is the process of decomposing it step by step from scale j to $j+1$. That is, the process from low resolution to high resolution. It ultimately broke

down into two signals: an approximation signal, D_jx , and a detailed signal, A_jx , both at a high frequency [26]:

$$\begin{cases} A_{j+1}x = \sum_k h(k - 2n)A_jx \\ D_{j+1}x = \sum_k g(k - 2n)A_jx \end{cases} \quad j \geq 1 \tag{2.5}$$

Equation (2.6) is the Mallat wavelet reconstruction formula:

$$x = \sum_{k=1}^k h(n - 2k)A_{j+1} + \sum_{k=1}^k g(n - 2k)D_{j+1} \quad j \geq 1 \tag{2.6}$$

After a finite energy signal undergoes wavelet transformation, it is decomposed into a set of detail signals and approximation signals. Each sample point of every signal has its wavelet decomposition coefficient $\omega_{j,k}$. When the signal contains noise, the noise is also decomposed along with the host signal. Therefore, theoretically, the noise may be removed if the wavelet decomposition coefficients of the noise are found and processed, and then all signals are submitted to wavelet reconstruction. The fundamental idea of wavelet threshold denoising is to target certain characteristics of noise in the signal, as well as the differences between noise and signal obtained after wavelet decomposition. A critical threshold is set, and the wavelet decomposition coefficients obtained after decomposition are compared with this threshold. If the result is less than the threshold, the coefficient is considered to belong to a normal signal and is retained without processing. This portion of wavelet coefficients is left untouched [27]. Conversely, if the result is greater than the threshold, the coefficient is considered to be from noise and needs to be zeroed out or processed through a specific threshold function. This yields an estimate for this portion of wavelet coefficients to replace the original ones. Once all wavelet coefficients are processed, wavelet reconstruction is performed to achieve the denoising effect. The key to wavelet threshold denoising lies in finding an appropriate threshold function, also known as a threshold rule. Hard, soft, and semi-soft threshold functions are the three primary categories of conventional wavelet threshold functions [28]. This study suggests enhancements to the previously described semi-soft threshold denoising approach, based on substantial testing and empirical analysis:

Let the high frequency signal be $W_{a_{j,k}}$, then The formula for estimating the noise standard deviation is as follows:

$$\sigma_j = \frac{1}{0.6745} \times \frac{1}{N} \sum_{K=1}^N |W_{a_{j,k}}|, 1 \leq j \leq J \tag{2.7}$$

Since the real original signal's SNR varies, the threshold's setting must also be adjusted to reflect the current circumstances. The following is the unified threshold formula found in the body of available literature [29]:

$$\lambda_{1,j} = \sigma_j \sqrt{2 \log(N)} \tag{2.8}$$

The high-frequency signal coefficients in the J group are produced once the signal has been broken down to the scale of J. Each group's wavelet coefficients are sorted in absolute value from small to big, yielding the following vector:

$$P = [W_{a_{j,n}}], 1 \leq n \leq N \tag{2.9}$$

The evaluation vector under the JTH wavelet coefficient is computed using this vector: $R = [r_n], 1 = n = N$. Where:

$$r_n = \sum_{k=1}^n W_{a_{j,n}} + (N - i)W_{a_{j,n}} + (N - 2n)\sigma_j^2 \tag{2.10}$$

After sorting the evaluation vector's interruption value from big to small and using the lowest value as the approximation error, the associated wavelet coefficient $W_{a_{j,m}}$ is discovered. The wavelet coefficient is used to determine the threshold value of the J-layer wavelet decomposition in the following manner:

$$\lambda_{a,j} = \sqrt{CD_{\min}} \tag{2.11}$$

The J-layer wavelet decomposition's threshold selection function is as follows:

$$\lambda_j = \begin{cases} \lambda_{1,j} & , (P_{a,j} - \sigma_j^2 < \rho_{N,j}) \\ \min(\lambda_{1,j}, \lambda_{a,j}) & , (P_{a,j} - \sigma_j^2 \geq \rho_{N,j}) \end{cases} \quad (2.12)$$

In this case, $\rho_{N,j}$ represents the wavelet coefficient vector's minimum energy level, and $P_{a,j}$ is the average value of the wavelet coefficient's absolute value. The following is the calculating formula:

$$P_{a,j} = \frac{1}{N} \sum_{k=1}^N W a_{j,k} \quad (2.13)$$

It is necessary to restore the signal to its original state since the calculated wavelet coefficient, or wavelet coefficient, is believed to be the result of noise. As a result, the original wavelet coefficient value is substituted for the estimated wavelet coefficient value through a series of computations, and wavelet reconstruction is ultimately used to accomplish the goal of noise reduction. The J-layer wavelet high-frequency signal's noise intensity is reflected by the introduction of a coefficient $\Gamma(\sigma_j)$ representing noise intensity. The following is the calculating formula:

$$\Gamma(\sigma_j) = \sqrt{\sigma_j/A_j} \quad (2.14)$$

In equation (2.14), A_j represents the amplitude of the high-frequency partial coefficient of the J-layer wavelet. The calculation formula of wavelet coefficient estimate is given:

$$w_{j,k} = \begin{cases} w_{j,k} - \Gamma(\sigma_j) \times \lambda_j, & w_{j,k} > \lambda_j \\ w_{j,k} + \Gamma(\sigma_j) \times \lambda_j, & w_{j,k} < -\lambda_j \\ 0, & -\lambda_j \leq w_{j,k} \leq \lambda_j \end{cases} \quad (2.15)$$

Here are the specific actions to do in order to enhance the wavelet threshold denoising algorithm: The high-pass filter h and low-pass filter g are configured, and the original signal $x(t)$ is discretized. The wavelet is disassembled by the J layer. The wavelet coefficients and the wavelet detail signal amplitude A_j of the layer decomposition are obtained. Based on the precise signal coefficients of each layer, the noise standard deviation σ_j and noise intensity coefficient $\Gamma(\sigma_j)$ of each layer are computed. Each layer signal's unified threshold, $\lambda_{1,j}$, is computed. The computation involves determining the adaptive threshold $\lambda_{a,j}$ for each layer signal, the lowest energy level $\rho_{N,j}$ of the layer's wavelet coefficients, and the average value $P_{a,j}$ of the absolute value of the layer's wavelet coefficients. Equation (2.12) is used to compute each layer's wavelet threshold. The wavelet coefficients are adjusted to complete the threshold denoising on this scale, and finally, the wavelet reconstruction is carried out according to the Figure 2.1.

Based on this as the main idea, simulation experiments are conducted on existing data. The results are shown in Figure 2.2. The studies demonstrate that, under the assumption of maintaining the majority of the fault information, the enhanced wavelet threshold denoising described in this study can suppress and eliminate noise better. Furthermore, the denoising findings are more appropriate for diagnosing data faults in the future.

2.2. MDF fusion construction for complex data scenarios. In the preceding section, a method model for data denoising was introduced. However, in complex data scenarios, it is often necessary to perform multi-source data fusion rather than just data denoising. This subsection will further discuss data processing based on this requirement. The construction of the MFD for road networks is the foundation of estimating traffic capacity and discerning the traffic operational status of road networks in this paper, based on the MFD method.

Addressing the potential issue of inconsistent MFD parameter estimation results in the process of constructing MFD based on multiple data sources, this paper proposes a framework and method for MFD fusion construction based on data reliability under the context of multi-source data, as illustrated in Figure 2.3.

The average traffic flow and average traffic flow density of the road network may be estimated using vehicle trajectory data and sectional detection traffic flow data. Several research have employed diverse data

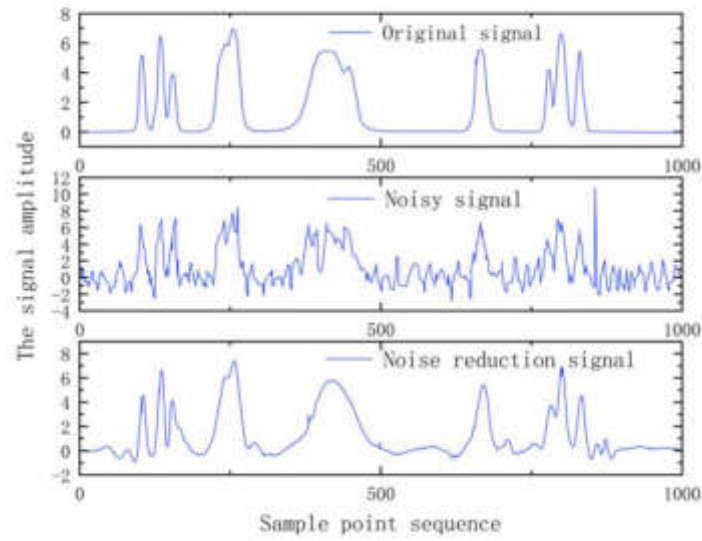


Fig. 2.2: Signal noise reduction effect diagram

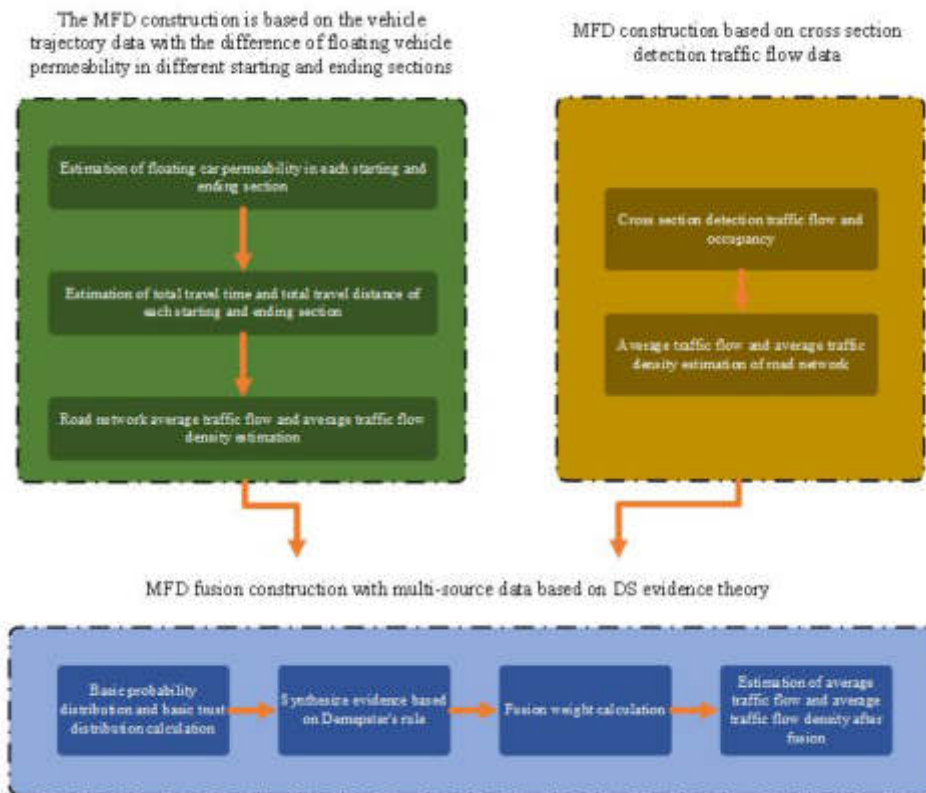


Fig. 2.3: MFD fusion construction framework based on multi-source data

sources to estimate different characteristics of MFD traffic operation states in order to get reliable estimates of the parameters of different traffic operation states in the MFD. The average traffic flow density and average traffic flow under various data sources are estimated using different assumptions, though. When estimating the average traffic flow density using vehicle trajectory data, it is common to assume that there are enough floating vehicle samples and that the penetration rates are distributed uniformly over space. On the other hand, the assumption that the traffic flow at the section represents the traffic flow of the road segment is typically used when estimating the average traffic flow based on sectional detection traffic flow data. As a result, outcomes of model parameter estimate under various data sources may differ. Equations (2.16) through (2.17) illustrate the fundamental foundation for the MFD fusion building in this study, taking the validity of data sources into consideration.

$$q(t) = \alpha(t) \times q^{\mathcal{F}}(t) + [1 - \alpha(t)] \times q^{\mathcal{M}}(t) \tag{2.16}$$

$$k(t) = \beta(t) \times k^{\mathcal{F}}(t) + [1 - \beta(t)] \times k^{\mathcal{M}}(t) \tag{2.17}$$

Among them, $q(t)$ represents the average road network traffic flow estimated by multi-source data fusion during period t (vehs/h), $q^{\mathcal{F}}(t)$ represents the average road network traffic flow based on vehicle track data during period t (vehs/h), the average traffic density on a road network, measured in vehicles per kilometer for a certain time t is represented by the symbol $k^{\mathcal{F}}(t)$, $q^{\mathcal{M}}(t)$ represents the road network average traffic flow (vehs/h) based on the traffic flow data detected in section during the period t , $k^{\mathcal{M}}(t)$ represents the road network average traffic flow density (vehs/km) based on cross section detection traffic flow data during period t , the average traffic flow density (vehs/km) for the road network during period t , as determined via multi-source data fusion, is denoted by $k(t)$. $\alpha(t)$ represents the floating vehicle data weight in the average road network traffic flow fusion estimation during period t ; $[1-\alpha(t)]$ represents the road during period t The weight of traffic flow data detected by section in the fusion estimation of network average traffic flow, $\beta(t)$ represents the weight of floating vehicle data in the fusion estimation of network average traffic flow density during the t period, $[1-\beta(t)]$ represents the weight of traffic flow data detected by section in the fusion estimation of network average traffic flow density during the t period.

The process of figuring out the fusion weights of two different kinds of data sources is the basis of the fusion construction. This research aims to ascertain the fusion weights of diverse data sources based on the mean and variance distributions (reliability) of the average road network traffic flow density and average traffic flow across different periods of multiple days, utilizing the Dempster-Shafer(DS) evidence theory [30]. In order to illustrate the determination procedure, the following uses the floating vehicle data weight $\alpha(t)$ in the fusion estimate of road network traffic flow characteristics during t period as an example.

First, as indicated by Equation (2.18), the recognition framework of the DS evidence inference model $\Theta(t)$ is built based on two types of data sources:

$$\Theta(t) = \{q^{\mathcal{M}}(t), q^{\mathcal{F}}(t)\} \tag{2.18}$$

where $q^{\mathcal{M}}(t)$ and $q^{\mathcal{F}}(t)$ respectively represent the average traffic flow of the network in the t period based on the traffic flow data detected by the cross section and the vehicle track data.

Given that $q^{\mathcal{M}}(t)$ and $q^{\mathcal{F}}(t)$ originate from distinct data sources, it is possible to regard them as mutually exclusive. Equation (2.19) therefore displays the power set of all elements in

$$??(t) : \begin{cases} 2^{\theta(t)_0} = \{\emptyset, X_1(t), X_2(t), X_3(t)\} \\ X_1(t) = \{q^{\mathcal{M}}(t)\} \\ X_2(t) = \{q^{\mathcal{F}}(t)\} \\ X_3(t) = \{q^{\mathcal{M}}(t) \cup q^{\mathcal{F}}(t)\} \end{cases} \tag{2.19}$$

Among them, \emptyset as the empty set, lots the $\|\theta(t)\|_{-0}$ for $??(t)$ L - 0 norm, $X_1(t)$? $X_2(t)$? $X_3(t)$ for $??$ loophole (t) of the collection, $X_1(t)$ said the decision to cross section under the detection of traffic flow data of road network traffic flow (vehs/h) on average, $X_2(t)$ means that the decision is the average traffic flow of the

road network under vehicle track data (vehs/h) $X_3(t)$ is an uncertain decision, indicating that it is impossible to distinguish which decision is $X_1(t)$ or $X_2(t)$. $q^M(t)$ and $q^F(t)$ have the same meaning as equation (2.18).

Second, two key ideas are included in DS evidence theory: The demspster evidence synthesis rule and fundamental trust distribution, which is also referred to as proof distribution or evidence function. A quantitative measure of the level of support and evidence for each choice is the basic trust distribution. The rule of evidence synthesis is a thorough examination that considers several pieces of data in relation to each choice. Make $s:2^{\|\theta(t)\|_0} \rightarrow [0, 1]$, said s is a basic trust distribution $2^{\|\theta(t)\|_0} \rightarrow [0, 1]$ mapping, it must meet the following requirements:

$$\begin{cases} s_i(\emptyset) = 0 \\ \sum_{X_j(t) \subset 2^{\Theta(t)_0}} S_i(X_j(t)) = 1 \end{cases} \quad i = 1, 2; j = 1, 2, 3 \tag{2.20}$$

where $S_i(X_j(t))$ represents the degree of support for decision $X_j(t)$ by the evidence provided by the i -th data source, and its value is the basic trust assignment value of decision $X_j(t)$, the basic trust value of the empty set is 0, and the sum of the trust values of other subsets is 1. In this example, i stands for the evidence provided by the i -th data source, j for the JTH decision choice.

In this work, the basic probability allocation function of choice $p_i(X_j(t))$ under each data source determines the fundamental trust allocation:

$$s_i(X_j(t)) = p_i(X_j(t)) / \sum_{j=1}^3 p_i(X_j(t)), i = 1, 2 \tag{2.21}$$

In each data source, the fundamental probability distribution function of the decision $X_j(t)$ is represented by $p_i(X_j(t))$.

Hypothesis $p_i(X_j(t))$ meet the $\varphi_i(v) \sim N(\mu_i(t), \sigma_i^2(t))$, the $\mu_i(t)$ and $\sigma_i^2(t)$ in the first class I average traffic flow data source under the historical data network $f_i(v)$ in t time mean and variance ($i = 1, 2$), $p_i(X_j(t))$ can be calculated according to equation (2.22).

$$\begin{cases} p_i(\bar{\emptyset}) = 0 \\ p_i(X_1(t)) = \int_{\mu_1(t)-1/2}^{\mu_1(t)+1/2} \frac{1}{\sqrt{2\pi\sigma_i(t)}} \exp\left\{-\frac{1}{2}\left(\frac{v-\mu_i(t)}{\sigma_i(t)}\right)^2\right\} dv \\ p_i(X_2(t)) = \int_{\mu_2(t)-1/2}^{\mu_2(t)+1/2} \frac{1}{\sqrt{2\pi\sigma_i(t)}} \exp\left\{-\frac{1}{2}\left(\frac{v-\mu_i(t)}{\sigma_i(t)}\right)^2\right\} dv \\ p_i(X_3(t)) = \frac{1}{2\pi\sigma_i^2(t)} \left\{ \begin{aligned} &\int_{\mu_1(t)-1/2}^{\mu_1(t)+1/2} \exp\left\{-\frac{1}{2}\left(\frac{v-\mu_i(t)}{\sigma_i(t)}\right)^2\right\} dv \\ &\times \int_{\mu_2(t)-1/2}^{\mu_2(t)+1/2} \exp\left\{-\frac{1}{2}\left(\frac{v-\mu_i(t)}{\sigma_i(t)}\right)^2\right\} dv \end{aligned} \right\} \end{cases} \tag{2.22}$$

Based on the basic trust assignment $S_i(X_j(t))$ of decision $X_j(t)$, the Demspster evidence synthesis process orthogonal and processed the basic probability assignment of multiple evidences. The evidence synthesis method of the two data sources is shown in Equation (2.23) :

$$s(X) = \begin{cases} K_D * \sum_{X_j(t) \cap X_j(t) = X} s_1(X_j(t)) s_2(X_j(t)) & X \neq \emptyset \\ 0 & X = \emptyset \end{cases} \tag{2.23}$$

Among them, $j=1,2,3$, $s(X), X \subset 2^{\|\theta(t)\|_0}$ is the synthetic trust distribution, $X_j(t)$ means that the decision is made under the evidence provided by the JTH data source, and the following formula is used to determine the conflict coefficient between the evidence offered by various data sources:

$$1/K_D = 1 - \sum_{x_j(t) \cap X_j(t) = \emptyset} s_1(X_j(t)) s_2(X_j(t)) \tag{2.24}$$

The closer $1/K_D$ is to 0, the greater the degree of conflict between $q^M(t)$ and $q^F(t)$ evidence provided by different data sources. When $1/K_D=0$, the sum under Dempster's synthesis rule does not exist.

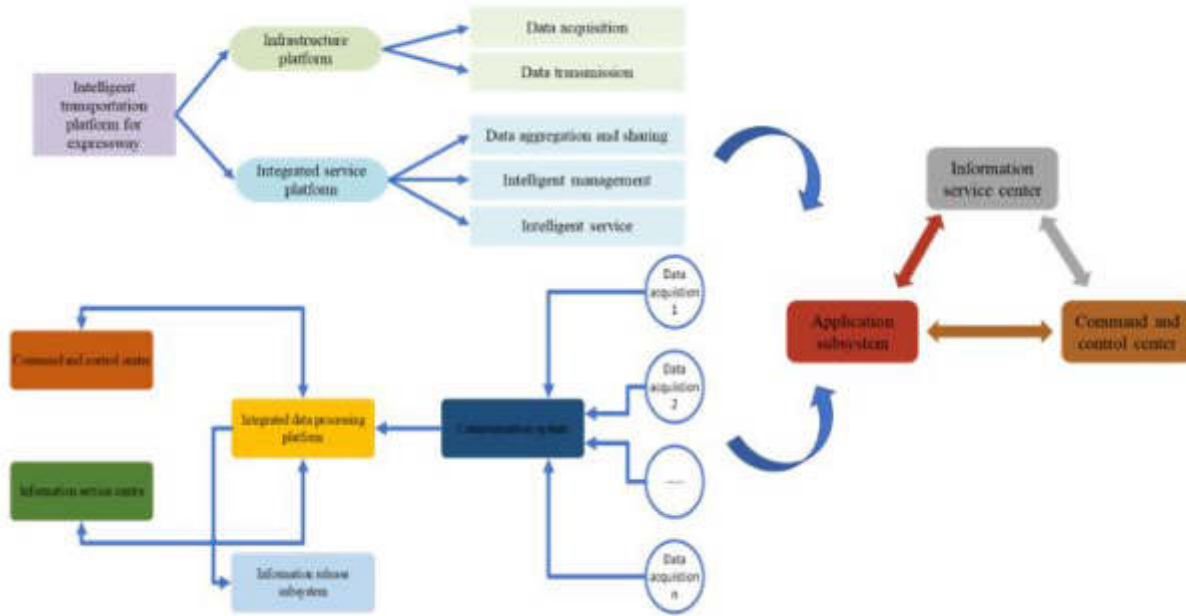


Fig. 3.1: ITS logical architecture diagram

The degree of disagreement between $q^M(t)$ and $q^F(t)$ evidence from various data sources increases with $1/K_D$'s proximity to 0. According to Dempster's synthesis rule, the total is null for $1/K_D=0$.

Equation (2.25) illustrates the approach used to calculate the $q^F(t)$ weights. Lastly, fusion weights are generated based on synthetic trust allocation.

$$\alpha(t) = \frac{s(X_2(t))}{s(X_1(t)) + s(X_2(t))} \tag{2.25}$$

Assuming that $\alpha(t)$ is equivalent to equation (2.16), the synthetic trust allocation for each choice is represented by $S(X_1(t)), S(X_2(t))$ and $S(X_3(t))$, taking into account the evidence supplied by each data source.

3. Urban intelligent monitoring. With the accelerated process of urbanization and the development of the automotive industry, the contradiction between the existing urban road capacity and the continuously growing traffic demand has become increasingly acute, leading to a growing prominence of traffic congestion. Utilizing Intelligent Transportation Systems (ITS) for controlling and guiding traffic flow to alleviate congestion and provide a smooth and orderly traffic environment is an important means vigorously developed and applied by various countries. The so-called Intelligent Transportation System is established on the basis of improved road infrastructure, integrating new-generation information, the Internet of Things (IoT), and computer technologies into traffic management. It aims to create an information-real-time, accurate, widely-serviced, all-encompassing, highly efficient, and high-quality transportation and management system [31]. In summary, ITS involves innovating traditional transportation systems through technology, creating a new type of transportation system that integrates informatization and intelligence [32]. The architecture of ITS is depicted in Figure 3.1. In this section, taking the complex urban road network as the research background, a combination of the two models mentioned earlier is performed. The goals of this include multi-source data-based urban traffic congestion monitoring and data-driven urban traffic safety environment monitoring.

3.1. Data-driven urban safety environment monitoring. A variety of factors are often considered while analyzing the traffic safety environment, such as the quantity (or frequency) of accidents, the nature of

the accidents (or types of vehicle collisions), and the severity of the accidents. The study of accident blackspots mostly uses accident frequency or quantity, with the usual objective being to pinpoint locations with a high accident rate. This kind of study often needs to pay more attention to including driver and vehicle conditions in the model and more attention to macroscopic traffic and road data. On the other hand, studies on accident types often concentrate on analyzing factors related to vehicles and roads, overlooking the impact of macroscopic traffic parameters and drivers. Therefore, a comprehensive analysis of accident severity that considers vehicles, drivers, macroscopic traffic, and road conditions can more comprehensively reflect the impact of accidents on the traffic safety environment. Thus, this paper selects accident severity as the research object.

Considering the comprehensiveness of data acquisition and the impact of the severity of injuries caused by accidents, a three-level classification method for accident severity is adopted, categorizing accidents into no-injury accidents, minor-injury accidents, and fatal or disabling accidents.

Various methods, including variable fusion, outlier detection and processing, are applied to restructure the relevant variables and data in the datasets. To ensure the timeliness and completeness of the data, this study uses accident and environmental data from recent years on Interstate 5 (I5) in Seattle, Washington. The data is collected from the National Highway Traffic Safety Administration's Fatality Analysis Reporting System (FARS) and the University of Washington's DRIVENet database. As the FARS dataset shares common labels such as TIME (time) and CASE NUMBER (accident number), the accident, vehicle, passenger, and environmental information in the information system table are fused based on these common labels and further combined with road information from the Washington University DRIVENet dataset, utilizing the MILEPOST (road mileage) label. The goal of restructuring the samples from multiple datasets is to make the samples suitable for model calculations. The initial complete sample labels include non-numeric symbols in textual expressions, mainly comprising some categorical variables. The restructuring of categorical variables involves label encoding and one-hot encoding.

On the basis of the above tests, in order to study the impact of different parameters on the traffic safety environment, the author introduced a factor importance index to evaluate the importance of different parameters. Considering that n -dimensional parameter vectors constitute the entire input space, the output vector Y corresponding to the first-stage partial derivative of the i th-dimensional variable x_i can explain the sensitivity of this variable to the output. According to the chain rule of calculus, equation (3.1) is obtained:

$$\frac{\partial Y}{\partial X_i} = \frac{\partial Y}{\partial \alpha} \frac{\partial \alpha}{\partial X_i} \quad (3.1)$$

?? is the calculated value from the network hidden layer to the output layer, and equation (3.1) can be rewritten as:

$$\frac{\partial Y}{\partial X_i} = \sum_{j=1}^L W_{k,j}^{(2)} \frac{\partial H_j}{\partial X_i} g(\alpha)' \quad (3.2)$$

J refers to the J TH node of the hidden layer, and $W_{k,j}^{(2)}$ refers to the weight from the hidden layer to the input layer. $g(\alpha)'$ is the activation equation from hidden layer to output layer.

Continuing to deduce according to the chain rule, Equations (3.3) and (3.4) can be obtained as follows.

$$\frac{\partial Y}{\partial x_i} = \sum_{j=1}^L W_{k,j}^{(2)} \frac{\partial H_j}{\partial \beta_j} \frac{\partial \beta_j}{\partial x_i} g(\alpha)' \quad (3.3)$$

$$\frac{\partial Y}{\partial x_i} = \sum_{j=1}^L W_{j,i}^{(1)} W_{k,j}^{(2)} f(\beta)' g(\alpha)' \quad (3.4)$$

$W_{j,i}^{(1)}$ refers to the connection weight from the input layer to the hidden layer, while $f(\beta)'$ is the activation equation from the input layer to the hidden layer.

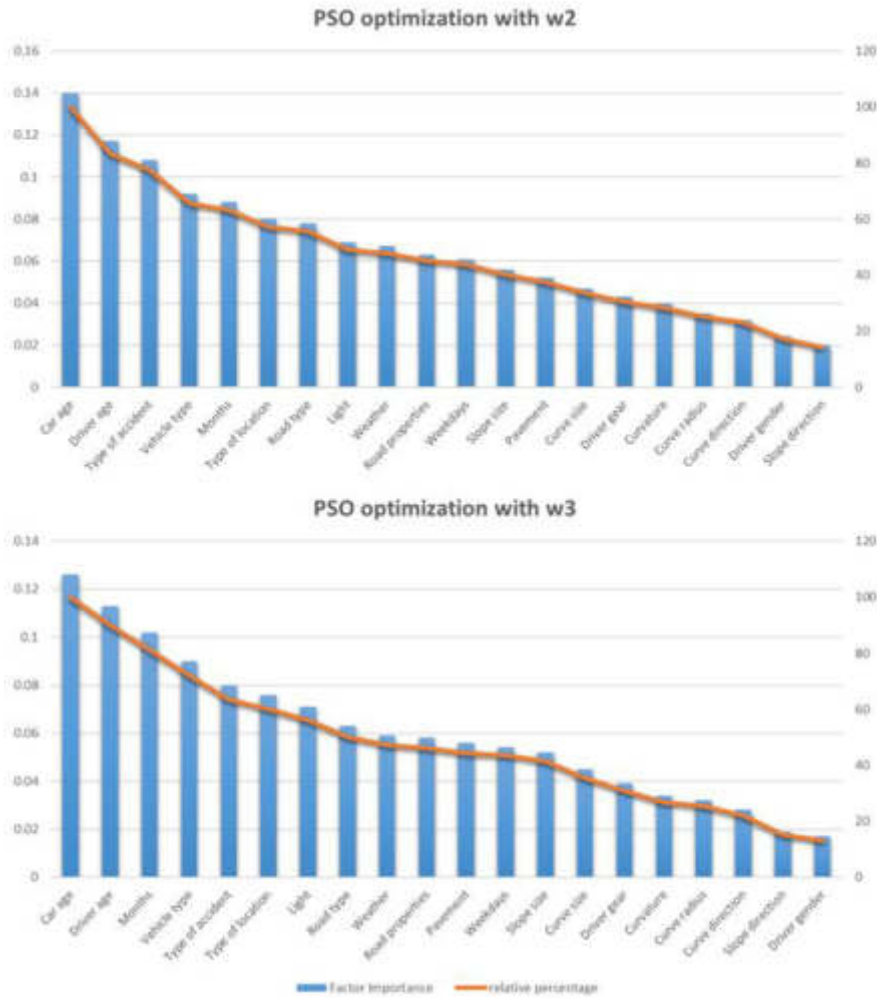


Fig. 3.2: Factor importance and relative percentage

The relevance R_i of a single indicator can be calculated as follows (3.5) :

$$R_i = \sum_{j=1}^L W_{j,i}^{(1)} W_{k,j}^{(2)} / \sum_i \sum_{j=1}^L W_{j,i}^{(1)} W_{k,j}^{(2)} \quad (3.5)$$

In the above formula, all weights W can be obtained from the results of network simulation. In order to eliminate the randomness of the simulation, the simulation is performed k times, and the factor importance is obtained as follows (3.6) :

$$E(R_i) = \frac{1}{K} \sum_{k=1}^K R_i^k \quad (3.6)$$

By combining the calculation results of Particle Swarm Optimization(PSO) optimization network with the calculation formula of factor importance, the importance and relative percentage of factors can be obtained in Table ?? below.

It is evident from Table ?? and Figure 3.2 that the factor importance rankings derived from the two PSO optimization approaches are almost exact. Eight categories of factors—vehicle age, driver age, accident type, month, accident location type, road function, and lighting conditions—have a relative importance of more than 50%. Vehicle age and driver age are the two most important factors influencing accident severity, which aligns with both our intuition and experience. The impact of road factors is limited, with road type being the only one with a high importance ranking, indicating that specific types of accidents tend to occur on roads with specific functions. Combining the example of accident statistics mentioned earlier, it can be observed that the wet and snowy weather and mountainous roads in Washington State significantly increase the number and severity of accidents during the winter.

3.2. Urban congestion monitoring based on multi-source data. The variation of traffic states is regarded as random because of the high level of complexity and uncertainty associated with the spatiotemporal

state of traffic systems. To deeply explore the correlations between various traffic data, it is necessary to integrate and analyze multiple data sources to extract the hidden operational patterns of traffic states buried within the data. This chapter, based on various traffic flow parameter data, employs the MFD for data fusion. It uses a Genetic Algorithm-Fuzzy C-Means (GA-FCM) algorithm to classify road traffic states. Road traffic states can then be predicted by combining the anticipated outcomes of traffic flow parameters. The effectiveness of this method in reflecting road traffic conditions is confirmed by experimental results.

The Genetic Algorithm (GA) is a model parameter optimization algorithm that relies on natural selection and genetic mechanisms. FCM is a commonly used fuzzy clustering algorithm for data classification and cluster analysis. The steps of optimizing the FCM model using GA are as follows: 1) Determine the fitness function: The fitness function, which evaluates the quality of each individual, is crucial for GA. In the FCM model, the fitness function can be selected as an evaluation index for clustering effects, such as clustering accuracy or clustering entropy. 2) Determine the encoding method: GA needs to encode each individual into chromosomes for genetic operations. In the FCM model, each individual can be encoded into a set of fuzzy cluster centers, where each chromosome contains multiple genes representing cluster centers. 3) Initialize the population: Randomly generate an initial population, with each individual representing a random set of cluster centers. 4) Selection operation: Based on the fitness function, select some excellent individuals as parents for the next generation. 5) Crossover operation: Perform crossover operations on parent individuals to generate new offspring individuals. In the FCM model, single-point crossover or multi-point crossover can be used. 6) Mutation operation: Introduce randomness to the offspring individuals through mutation operations to increase population diversity. In the FCM model, random perturbation or random replacement can be applied. 7) Evaluate fitness: Evaluate the fitness of the new generation, calculating the fitness value for each individual. 8) Repeat selection, crossover, mutation, and fitness evaluation operations until reaching the preset stopping conditions, such as reaching the maximum iteration times or convergence of fitness values. 9) Output the optimal solution: In the last generation of the population, select the individual with the highest fitness as the optimal solution, representing the optimal cluster centers.

Firstly, the GA-FCM clustering algorithm is applied to classify traffic flow parameters to obtain the optimal number of clusters and corresponding cluster centers for each state. Subsequently, a fuzzy clustering algorithm is utilized to partition the data, and based on the fuzzy cluster center results, different clusters representing traffic flow states are determined. These states are considered as prior knowledge for real-time traffic state identification. Finally, by calculating the membership degree of traffic flow data to each cluster center and the corresponding membership degree to different road traffic states, the traffic flow state at that moment can be determined. This process is accomplished using the membership degree function, selecting the traffic flow state with the maximum membership degree as the final identification result.

The selection range of the fuzzy factor is usually between 1 and 2. This parameter can suppress the influence of noise pollution by assigning larger weights to the membership function, reducing the impact of noise points on the FCM objective function during iterations. Starting from 1.0, experiments are conducted with a step size of 0.1, and the value of the fuzzy factor "m" is finally determined as 2. In the clustering analysis of traffic data, the number of clusters "C" plays a crucial role in the partitioning and identification of traffic states. To determine the optimal number of fuzzy partitions, the clustering validity function is commonly used as an evaluation criterion. This study employs the clustering validity function based on membership and squared membership weights to determine the number of clusters, as shown in Equation (3.7).

$$V_u = \frac{1}{2n} \left(\sum_{j=1}^n \sum_{i=1}^c u_{i,j}^2 + \sum_{j=1}^n \max_{i=1}^c u_i \right) \quad (3.7)$$

By computing when the number of clusters (C) is set to 4, the maximum fuzzy correlation value is achieved, indicating the highest effectiveness of data clustering. Therefore, in this experiment, we categorize the traffic states of road segments into four levels: smooth, moderately smooth, congested, and heavily congested. This categorization is based on the result obtained with a cluster number of 4.

The objective of the fuzzy mean clustering algorithm, optimized by a genetic algorithm in this study, is to

identify a set of cluster centers and membership degrees that minimize the objective function. To achieve this goal, an iterative optimization approach is employed, iteratively updating the cluster centers and membership degrees to reduce the value of the objective function progressively. The objective function in this paper is the fuzzy mean clustering algorithm's loss function, as indicated by the following equation, because the underlying methodology uses the Fuzzy C-Means (FCM) algorithm for traffic state partitioning.

$$fit = \min J_f = \min \sum_{j=1}^c \sum_{i=1}^n [u_j(d_i)]^m d_i - \nu_j \quad (3.8)$$

The selection of the number of iterations and the stopping error depends on the specific problem and dataset. Generally, a higher number of iterations leads to increased algorithm precision, but it also results in longer computation times. Therefore, it's advisable to choose an appropriate number of iterations based on the actual circumstances. The convergence of the algorithm can be observed to determine the number of iterations. If the algorithm has converged within a certain number of iterations, it can be stopped. The stopping error refers to halting iterations when the algorithm reaches a specific error range. The choice of the stopping error depends on the specific problem and dataset. If the dataset is noisy, a larger stopping error can be chosen to prevent overfitting.

Conversely, if the dataset is clean, a smaller stopping error can be chosen to enhance precision. In practical applications, the number of iterations and stopping errors often need to be considered together. Typically, one can start with a small stopping error and then determine the number of iterations based on the convergence of the algorithm. If the algorithm has converged within a certain number of iterations, it can be stopped. If the algorithm hasn't converged, the number of iterations can be increased until convergence or reaching the maximum iteration limit. Considering these factors, this experiment sets the maximum number of iterations to 100 and the iteration-stopping error threshold to 1e-5.

This study intends to conduct an instance analysis using two different levels of road segment data. Clustering analysis will be performed using historical traffic flow parameter data to obtain a road traffic state discrimination method based on multi-source data. The fuzzy partitioning process of road traffic states based on GA-FCM includes the following steps: extracting features such as flow, speed, and occupancy rate from the original data, constructing the objective function, utilizing genetic algorithms to find the optimal clustering centers, and finally outputting the clustering results of traffic states.

According to the above parameter Settings, the clustering centers of four traffic states of the main road obtained by GA-FCM clustering algorithm are represented by $V = \{\nu_1, \nu_2, \nu_3, \nu_4\}^T$, and the results of equation (3.9) and Figure 6 are obtained respectively.

$$V1 = \begin{bmatrix} 31.82 & 55.88 & 0.04721 \\ 85.17 & 49.74 & 0.08338 \\ 148.7 & 50.37 & 0.2233 \\ 109.9 & 18.3 & 0.1975 \end{bmatrix} \quad (3.9)$$

In this context, the three columns of the clustering centers represent flow, speed, and occupancy, respectively. Each row corresponds to a specific traffic state, namely, smooth-flowing, moderately smooth-flowing, congested, and heavily congested.

From the graph, it can be observed that there is a certain continuity in the value ranges between different traffic states. Under different traffic states, certain parameter ranges exhibit overlapping characteristics. As the traffic state transitions from smooth-flowing to congested, the speed gradually decreases while the flow and occupancy increase. In a heavily congested state, where vehicles move at a slower pace, the flow gradually decreases, aligning with the fundamental principles of traffic operation.

V2 represents the clustering centers for the four traffic states on the expressway, and the partition results

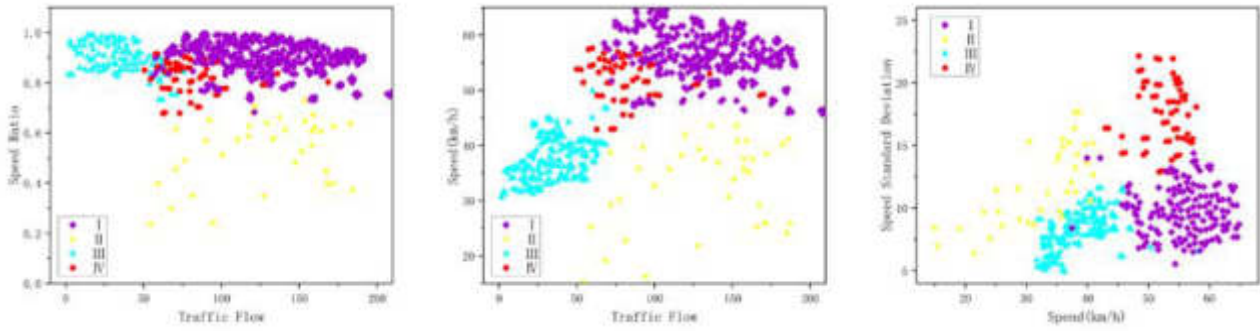


Fig. 3.3: Main Road Traffic Clustering Results

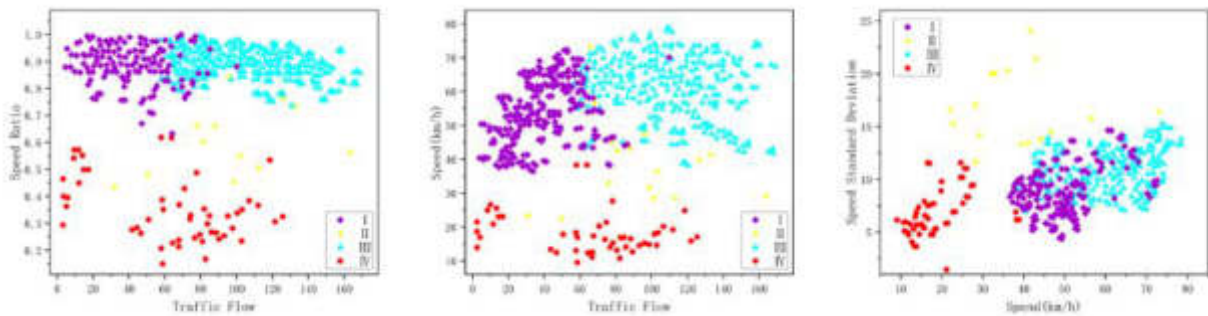


Fig. 3.4: Expressway Traffic Clustering Results

are illustrated in Equation (3.10) and Figure 3.4.

$$V2 = \begin{bmatrix} 25.41 & 50.3 & 0.04761 \\ 71.95 & 44.45 & 0.08646 \\ 119 & 45.5 & 0.223 \\ 87.87 & 15.91 & 0.2007 \end{bmatrix} \tag{3.10}$$

From the above figure, it is evident that, compared to arterial roads, expressways exhibit slightly higher traffic flow and speed. Expressways demonstrate relatively high traffic flow stability, with minimal speed differences among vehicles, resulting in a smoother traffic flow.

Different types of roads exhibit distinct characteristics in traffic flow parameters, necessitating reasonable planning and design to ensure the efficient operation and safety of the urban road network. Expressways typically have higher traffic volumes than arterial roads, as they handle a large number of passenger and freight vehicles. Arterial roads follow in terms of traffic volume, primarily accommodating internal traffic flows within the city, including commuting, commercial, and service-related traffic. Regarding speed, expressways usually have higher speeds than arterial roads due to their higher design speeds, wider lane widths, and absence of traffic signals, allowing vehicles to travel at higher speeds. Arterial roads have slightly lower speeds as traffic signals typically control them, and vehicles need to adhere to traffic rules and signal indications, resulting in relatively slower speeds.

Compared to the traditional FCM clustering algorithm, the GA-FCM clustering algorithm optimized by a genetic algorithm can more rapidly identify the optimal clustering partition. Additionally, both algorithms yield the same final value for the objective function, with the FCM algorithm consistently having a higher final

value than the GA-FCM algorithm. The graph also illustrates that GA-FCM exhibits faster convergence and better stability compared to the traditional FCM clustering algorithm.

3.3. Research on old city reconstruction. Urban traffic intelligent monitoring provides unprecedented opportunities for the transformation of urban road networks. With continuous technological advancements, traffic managers can utilize advanced monitoring technologies and data analysis tools to achieve more efficient and intelligent road network designs. This process not only improves traffic flow but also contributes to enhancing the quality of life for urban residents.

Firstly, intelligent monitoring systems provide cities with comprehensive traffic data. By deploying cameras, sensors, and other monitoring devices, traffic managers can obtain real-time information on vehicle flow, speed, congestion, and more. These data form the basis for a deep understanding of the city's traffic conditions. Leveraging advanced artificial intelligence algorithms, managers can extract valuable information from massive datasets and identify traffic patterns, peak periods, and potential bottleneck locations. With this information, cities can engage in more precise traffic planning and road network improvements.

The data analysis capabilities of intelligent monitoring systems empower traffic managers to identify bottlenecks and congestion points, allowing targeted optimization of intersections and adjustments to traffic signals. Through real-time traffic flow control, congestion can be effectively alleviated, and road operational efficiency improved. This not only helps relieve traffic pressure but also reduces carbon emissions, enhancing urban air quality.

Real-time traffic information dissemination systems extend the capabilities of intelligent monitoring systems by providing drivers and citizens with real-time information through mobile apps, digital signage, social media platforms, and more. This enables citizens to flexibly plan travel routes and choose alternative roads to avoid congestion. Additionally, it enhances the traffic awareness of participants, reducing the incidence of traffic accidents.

Data sharing and cross-departmental cooperation are crucial for achieving intelligent traffic management. Establishing data-sharing mechanisms among different traffic systems facilitates the collaborative operation of traffic systems. Through big data analysis, urban planners can better understand the city's traffic demands, providing a scientific basis for future road network transformations.

In addition to improving traffic efficiency, urban traffic intelligence monitoring should also focus on environmental sustainability. Combining traffic monitoring systems with environmental monitoring technologies allows cities to assess the impact of traffic on air quality and noise levels. Based on these assessments, corresponding measures can be taken, such as the establishment of green belts and the construction of sound barriers, to improve the living environment for urban residents.

In conclusion, road network transformation based on urban traffic intelligent monitoring represents a revolutionary attempt in urban traffic management. By fully leveraging advanced monitoring technologies and data analysis tools, cities can achieve a more intelligent, efficient, and environmentally friendly traffic system. This not only concerns the development of urban traffic but also affects the travel experience and quality of life for urban residents.

4. Conclusion. This paper makes a detailed discussion of urban intelligent monitoring, puts forward some algorithm models of intelligent city monitoring based on data-driven, and briefly discusses its application in the reconstruction and renewal of old cities. At the same time, it also puts forward new solutions for data noise reduction and multi-source data fusion in complex urban environments. Summarizing the relevant research in this paper, the following conclusions are drawn:

1. This paper proposes a data-driven signal denoising model using an improved wavelet threshold method. Three key points of wavelet threshold denoising are modified, and the detailed calculation formula and calculation flow of the denoising algorithm are given. Using the experimental data containing noise and fault to debug, the signal-to-noise ratio decreases significantly after noise reduction, and most types of fault information can still exist on multiple scales. Experiments show that the model can suppress and remove the noise well while retaining most fault information. The noise reduction results are more suitable for the late diagnosis of data faults and can be applied to constructing an urban intelligent monitoring model.

2. An MFD fusion structure is proposed to aim at the problem of multi-source data in urban traffic intelligence monitoring. An urban congestion monitoring model is constructed by combining the FCM model optimized based on a genetic algorithm. Compared with the traditional FCM clustering algorithm, the GA-FCM clustering algorithm optimized by the genetic algorithm can find the optimal cluster classification faster. In addition, the final value of the objective function of both algorithms is the same, and the final value of the objective function of the FCM algorithm is always higher than that of the GA-FCM algorithm.
3. The application of the monitoring model in the smart city in the reconstruction and renewal of the old city is discussed, with clear objectives, reasonable means and correct direction, which has certain guiding significance.

The research in this paper is a preliminary exploration of road network operation status and traffic capacity estimation after noise reduction of multi-source data using a wavelet algorithm. With the deepening of the research, the author feels the challenge of the research and still doubts the further application and development of the model in the reconstruction of old cities. Based on the current research, future research can be carried out from the following three aspects.

In the process of MFD fusion construction based on multi-source traffic perception data, limited by the acquisition of actual data, this paper only selected a small range of actual road networks to verify the method. In the future, when more realistic data are collected, the accuracy of the proposed method in the actual large-scale road network MFD fusion construction needs to be further verified. In the link of road network dynamic capacity estimation, this paper adopts the data-driven method, the premise of which requires the road network to have a relatively complete traffic operation state (unsaturated, saturated, and supersaturated). Therefore, its application may be limited to the capacity analysis of road networks such as central urban areas. In the future, theoretical methods based on analytical modeling can be further considered. A road network capacity estimation method driven by data and model is constructed. The application of this model in the reconstruction of the old city is only a superficial discussion in this paper, without further analysis and calculation, which can be studied in the future.

REFERENCES

- [1] Gao, X. & Wang, Y. Data fusion technology review. *Journal Of Computer Automatic Measurement And Control*, 706-709 (2002)
- [2] Yang, Z., Wang, S. & Ma, D. Review of basic traffic information fusion methods. *Highway Transportation Science And Technology*, 111-116 (2006)
- [3] El Faouzi, N., Leung, H. & Kurian, A. Data fusion in intelligent transportation systems: Progress and challengesA survey. *Information Fusion*. **12**, 4-10 (2011)
- [4] Xu, T., Yang, X. & Xu, A. Research on Data Fusion for Urban Road Traffic State Estimation. *Computer Engineering And Applications*. **47**, 218-221 (2011)
- [5] Ding, Y., Wang, J. & Lu, W. Fusion of multi-source relational data. *Science In China: Information Science*. **50**, 649-661 (2019)
- [6] D'Abadie, R. & Ehrlich, T. Contrasting Time-Based and Distance-Based Measures for Quantifying Traffic Congestion Levels: Analysis of New Jersey Counties. (2002)
- [7] Washburn, S. & Kirschner, D. Rural Freeway Level of Service Based on Traveler Perception. *Transportation Research Record Journal Of The Transportation Research Board*. **1988** pp. 31-37 (2006)
- [8] Taylor, M., Woolley, J. & Zito, R. Integration of the global positioning system and geographical information systems for traffic congestion studies. *Transportation Research Part C: Emerging Technologies*. **8**, 257-285 (2000)
- [9] Kerner, B., Demir, C. & Herrtwich, R. Traffic state detection with floating car data in road networks. *Proceedings. 2005 IEEE Intelligent Transportation Systems*. pp. 44-49 (2005)
- [10] Feng, X., Zhou, C. & Rong, J. Research on frequent traffic congestion on urban expressway based on speed characteristics. *Traffic Information And Safety*. **32**, 29-33 (2014)
- [11] Sun, C., Zhang, H. & Chen, X. Road traffic operation evaluation based on multi-source floating vehicle data fusion. *Journal Of Tongji University (Natural Science)*. **46**, 46-52 (2018)
- [12] Ren, J., Ou, X. & Zhang, Y. Research on traffic status pattern recognition. *Highway Transportation Science And Technology*, 63-67 (2003)
- [13] Sun, Y., Qian, H. & Ye, L. Application of data mining algorithm in traffic state quantification and recognition. *Journal Of Computer Applications*, 738-741 (2008)
- [14] Wang, M., Xie, D. & Zhao, X. A method for calculating regional traffic congestion correlation based on spatiotemporal association rules. (2013,12,24)

- [15] Xu, F., He, Z. & Sha, Z. Traffic state evaluation based on macroscopic fundamental diagram of urban road network. *Procedia-Social And Behavioral Sciences*. **96** pp. 480-489 (2013)
- [16] Shuqing, L. & Jianmin, X. Urban traffic state analysis based on the macroscopic fundamental diagrams of the variability of vehicle densities. *2016 12th World Congress On Intelligent Control And Automation (WCICA)*. pp. 1010-1015 (2016)
- [17] Daganzo, C., Gayah, V. & Gonzales, E. Macroscopic relations of urban traffic variables: Bifurcations, multivaluedness and instability. *Transportation Research Part B: Methodological*. **45**, 278-288 (2011)
- [18] Haddad, J. & Geroliminis, N. On the stability of traffic perimeter control in two-region urban cities. *Transportation Research Part B: Methodological*. **46**, 1159-1176 (2012)
- [19] Aboudolas, K., Papageorgiou, M. & Kouvelas, A. A rolling-horizon quadratic-programming approach to the signal control problem in large-scale congested urban road networks. *Transportation Research Part C: Emerging Technologies*. **18**, 680-694 (2010)
- [20] Qu, Z., Wang, X. & Song, X. Based on large data of urban taxi GPS hot travel road recognition method. *Journal Of Transportation Systems Engineering And Information Technology*. **12**, 238-246 (2019)
- [21] Xu, X., Dou, W. & Zhang, X. A traffic hotline discovery method over cloud of things using big taxi GPS data. *Software*. **47**, 361-377 (2017)
- [22] Lin, X. A road network traffic state identification method based on macroscopic fundamental diagram and spectral clustering and support vector machine. *Mathematical Problems In Engineering*. **2019** pp. 1-10 (2019)
- [23] Xu, Z., Mo, X. & Xu, Z. Research review of road traffic detectors and their optimal layout methods. *Journal Of South China University Of Technology (Natural Science Edition)*. **51**, 68-88 (2023)
- [24] Xu, B., Cen, K. & Huang, J. A review of graph convolutional neural networks. *Chinese Journal Of Computers*. **43**, 755-780 (2020)
- [25] Luo, D., Li, Y. & Luo, Z. Detection and analysis of hanging basket wire rope broken strands based on Mallat algorithm. *International Conference On Neural Computing For Advanced Applications*. pp. 518-532 (2023)
- [26] Li, H., Zhou, Y. & Tian, F. A new adaptive wavelet thresholding function vibration signal denoising algorithm. *Journal Of Instruments And Meters*. **4**, 2200-2206 (2015)
- [27] Chang, G. Adaptive wavelet thresholding for image denoising and compression. *IEEE Transactions On Image Processing*. **9** (2000)
- [28] Guo, H., Jing, X. & Shang, Y. Research on vehicle license plate location based on wavelet transform and mathematical morphology. *Computer Technology And Development*. **20**, 13-16 (2010)
- [29] Hou, P., Zhao, J. & Liu, M. License plate location method based on wavelet transform and line scan. *Journal Of System Simulation*., 811-813 (2006)
- [30] Qiu, D., Li, X. & Xue, Y. Analysis and prediction of rockburst intensity using improved DS evidence theory based on multiple machine learning algorithms. *Tunnelling And Underground Space Technology*. **140** pp. 105331 (2023)
- [31] Practice Practice of comprehensive urban traffic management in Zhejiang Province under the background of traffic congestion control. *Urban Transportation*. **18**, 11-12 (2020)
- [32] Xiao-Min, Z. Current situation analysis and suggestions on the development of intelligent transportation. *Journal Of Electronic World*., 96-97 (2020)

Edited by: Zhengyi Chai

Special issue on: Data-Driven Optimization Algorithms for Sustainable and Smart City

Received: Jan 15, 2024

Accepted: Apr 28, 2024



RESEARCH AND APPLICATION OF A DUAL FILTERING MUSIC HYBRID RECOMMENDATION MODEL BASED ON CATBOOST ALGORITHM AND DCN

JUNCAI HOU*

Abstract. With the increase of Internet users, the traditional music recommendation model can not meet the increasing personalized needs of users. The single deep cross network model has some defects in music recommendation, such as poor stability and inability to process complex data. To overcome the shortcomings of existing models, a new hybrid music recommendation model combining CatBoost algorithm and deep cross network is constructed to improve the recommendation performance and better meet the individual needs of users. Then the performance of the hybrid model is compared with other algorithms. The results showed that the accuracy of the proposed hybrid algorithm was up to 92.7%, which was superior to the comparison algorithm. In comparison with other single model and hybrid model, it is found that the proposed model was more than 0.05% higher than other models in the four indices of AUC area, accuracy, precision and recall. The above results showed that the proposed hybrid music recommendation model could efficiently process data information and provide users with accurate personalized music recommendation. This study not only promoted the development of music consumption and creation, but also found that the CatBoost-DCN hybrid model was significantly effective in improving recommendation performance. This finding provides a more efficient recommendation strategy for music platforms and has far-reaching significance for improving user experience and satisfaction.

Key words: Music recommendation; CatBoost; DCN; Data modeling; Machine learning

1. Introduction. With the rapid development of the internet and mobile internet, music recommendation systems have become an indispensable part of the music industry [3]. Music recommendation systems can help users discover new music and increase user engagement on music platforms. The current music recommendation system mostly adopts collaborative filtering algorithm, which recommends music liked by similar users according to their historical behaviors and preferences [6]. However, traditional collaborative filtering algorithms ignore the emotional and stylistic characteristics of music, as well as the subtle preference differences of users, thus limiting the accuracy and personalization of the recommendation system (Sterman et al. 2021). To improve the quality and accuracy of music recommendation, a new hybrid recommendation model is proposed to meet the individual needs of users more comprehensively and accurately capture the emotional and stylistic characteristics of music. This hybrid recommendation model combines CatBoost algorithm and Deep Cross Network (DCN) model. Among them, CatBoost algorithm can balance personalized recommendation and music popularity, while DCN model can dig deeply into the emotion and style characteristics of music, and predict user preferences more accurately by integrating music metadata and historical behavior data of users [17, 13]. The innovation of this research lies in integrating CatBoost and DCN algorithms into the music hybrid recommendation model at the same time. Compared with the traditional recommendation system, it not only significantly improves the personalization and accuracy of recommendation, but also enhances the user experience and brings higher user stickiness and broader profit space to the music platform. This study has important implications for the development of music recommendation systems and also demonstrates the positive role of advanced technology in meeting social and cultural needs. The research is divided into four parts. The first part is to analyze the research status of DCN algorithm and music recommendation model. The second part describes CatBoost and the process of building a music recommendation model after merging with DCN. The third part is to compare and analyze the performance of CatBoost-DCN hybrid algorithm and CatBoost-DCN music recommendation model. The last part is the summary of the full text.

2. Related Works. In the era of big data and information, the growth of social media is accompanied by the rapid growth of a variety of data and information. In the context of imperfect information filtering

*The Department of Music, Xinxiang University, Xinxiang, 453000, China (Juncai_Hou1234@outlook.com)

mechanisms, how music users can obtain information of interest has become a major challenge for music recommendation systems. To effectively identify and predict the prone sites of cancer, Pandey developed an amino acid sequence feature model using a DCN. After cross validation, the results showed that the model could predict the prone sites of diseases with 81.6% accuracy [14]. To improve the detection accuracy of phishing websites, Anitha and Kalaiarasu proposed a phishing detection system based on mixed deep learning algorithm, and tested the detection system. The results showed that the accuracy of the detection system on phishing websites was much higher than that of the traditional detection model. It also had high robustness and strong prediction ability in distinguishing phishing websites from legitimate websites [1]. To obtain the optimal approximation error characteristics, [11] used DCN to correct the linear unit network. Experimental results showed that after being corrected by this network, the deep linear unit network with different depths exhibited non collinearity. To solve the problem of automatic recognition of different levels of glioma during brain tumor surgery, [2] proposed a deep learning model based on DCN automatic deep learning network. After sample experiments, the model classification produced a sensitivity of 88.9%, an F1 score of 0.906, and 100% specificity. To understand how the number and spacing of communities affect postal delivery, [16] used a mixed effects model based on CatBoost for detection and analysis. After analyzing the sampling points, the metacommunity structure was influenced by natural and human landscape scale variables. To overcome the overfitting problem caused by the small native language data set in the mixed speech environment, Gupta's team proposed a classification model based on CatBoost algorithm and fine-tuned it. The results showed that the model can effectively avoid the overfitting problem. [9]. In addition, for the problem that diabetes is difficult to predict accurately in the early stage, Jenefer and Deepa proposed a CatBoost classifier based on firefly optimization to predict diabetes. The comparison test between this classifier and other similar classifiers showed that CatBoost classifiers had higher accuracy and lower loss values [10].

In modern society, music plays an important role in relaxing the mood and bringing people beautiful enjoyment. However, due to the numerous categories and quantities of music itself, as well as the different personal preferences of the audience, music recommendation is very difficult. To classify music and effectively recommend it to users, Elbir A et al. constructed a new Deep Neural Network (DNN) model based on acoustic features of music to extract representative features. After training the dataset, the results showed that the model could effectively classify music types and recommend music [6]. In view of the problems of cold start and new user recommendation in music recommendation, Yadav et al proposed a self-focused deep music recommendation model based on MIDI content data, and tested the model. The results showed that this method could effectively improve the recommendation effect by using MIDI content information. It outperformed other advanced models in several comparisons [20]. To obtain the ideal sound of the most popular DJs in search tools and perform DJ classification, Ziemer et al. (2020) [24] constructed a model that includes third octave mixing analysis, peak factor meter, phase range, and channel correlation coefficient functions. Through machine learning experiments, it was found that the accuracy of model detection was 73% [3]. To calculate, model, and classify music emotional content, the Chapaneri team proposed a structured regression framework that uses a single regression model to model the potency and arousal emotional dimensions of music. After training the benchmark dataset, the proposed work achieved significant improvements in R2 of arousal and valence dimensions [24]. To avoid the one size fits all phenomenon in music recommendation, Jin et al. constructed a system to optimize user control level based on visual memory and music maturity features. After detecting the interactive visual design system model of the music recommendation system, the results showed that music complexity would enhance the impact of UI on perceptual diversity [4]. To solve the problems of cold start and content feature extraction in Music classification and recommendation, Mao et al proposed a music-CRN model to optimize classification and recommendation by learning audio content features. Empirical analysis of the model found that this method performed better in music classification and recommendation tasks than previous methods [12]. Aiming at the problem that the existing music recommendation system fails to fully capture the correlation between internal and external information, the Xu team proposed a hierarchical multi-information fusion recommendation method, and tested the method. The experimental results showed that compared with the baseline method, the method performed best on the NOWPLAYINGRS dataset. The validity and rationality of the model were verified [19].

Based on the above related studies and the comparison of the advantages and disadvantages of the proposed

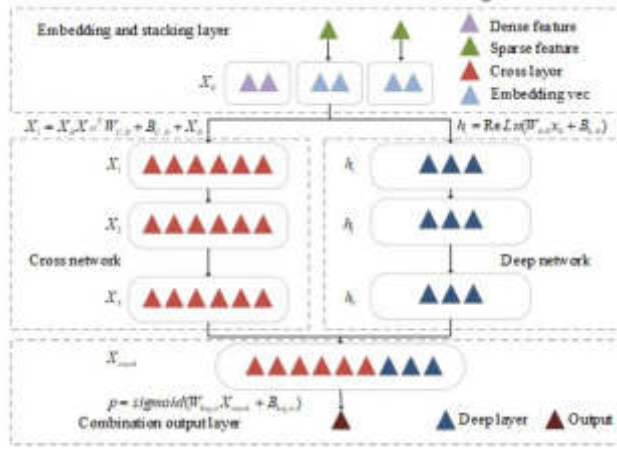


Fig. 3.1: Network model structure of DCN

method, Table 2.1 is obtained.

From Table 2.1, the proposed method integrates CatBoost algorithm and DCN model in the music recommendation system, aiming to meet users' individual needs more comprehensively and accurately capture the emotional and stylistic characteristics of music. The significant advantage of this method is that it can take into account both personalized recommendation and in-depth mining of music characteristics, so as to improve the accuracy and personalized degree of recommendation. Compared to the methods in other references, the proposed method focuses more on how to more accurately capture the intrinsic characteristics of user preferences and music. Although other relevant methods perform well in their respective fields, they have little to do with the direct application of the music recommendation system, and the methods applied in the field of music recommendation cannot meet the personalized needs of users. Therefore, the proposed music recommendation model based on CatBoost algorithm and DCN algorithm aims to fill the knowledge gap, and significantly improve the personalization and accuracy of music recommendation by integrating advanced algorithms, so as to enhance user experience and improve user stickiness of music platform.

3. Methods and Materials.

3.1. Construction of a DCN-based music recommendation model. DCN consists of a DNN and a Cross Network (CN) in parallel [7]. The drawback of traditional DNN is that only a combination of partial features can obtain better features, and its implicit learning features are inexplicable, resulting in low learning efficiency [23]. The main principle of DCN is to consider discrete and continuous features separately, encode and embed discrete features, and then concatenate and combine them with continuous features. The network model structure of DCN is shown in Figure 3.1.

From Figure 3.1, the original data in the DCN model is first divided into discrete and continuous features. Then at the embedding layer, the discrete features are transformed into real value vectors by building a random initialization vector lookup table. Subsequently, the transformed real value vector and the continuous feature are perfectly fused in the stacked layer to form the final output vector. This design not only optimizes the feature processing, but also improves the expressiveness and flexibility of the DCN model. The stacking function in the DCN model is shown in equation (1).

$$X_0 = [x_{\text{embed},1}^T \cdots, x_{\text{embed},k}^T, x_{\text{dense}}^T] \quad (1)$$

In equation (1), $x_{\text{embed},k}^T$ means the vector of the k th feature after embedding operation; x_{dense}^T denotes the transposed continuous value eigenvector; X_0 refers to both CN and DNN inputs. The DNN is composed of n layer networks and is a fully linked neural network system. Each layer of the deep network can be represented

Table 2.1: Comparison of the advantages and disadvantages of the research methods in this paper with those in different references

Author	Research method	Application field	Advantages	Disadvantages
Research in this paper	Music recommendation model integrating CatBoost and DCN	Music recommendation	Both personalization and music feature mining can significantly improve the personalization and accuracy of recommendation	/
Pandey [14]	Amino acid sequence feature model based on CatBoost	Disease prediction	High accuracy in predicting disease prone sites	Only applicable in the field of cancer prediction
Anitha and Kalaiarasu	Phishing detection system based on CatBoost	Phishing site detection	High accuracy, robustness and predictive power	The correlation with music recommendation system is weak It has little
Lu et al. [11]	DCN algorithm	Deep linear unit network correction	The corrected network has non collinearity	relevance to the application of music recommendation system
Bagyaraj et al. [2]	Deep learning model based on DCN	Automatic recognition of gliomas of different grades	High sensitivity, F1 score and specificity	Focus on glioma recognition, not music recommendation
Jenefer and Deepa [10]	CatBoost classifier	Diabetes prediction	High accuracy and less loss value	Applied to diabetes prediction, not directly related to music recommendation
Elbir A et al. [6]	DNN model based on acoustic features	Music classification and recommendation	Effectively classify and recommend music genres	Too much reliance on acoustic features
Yadav et al. [20]	Self-focused deep music recommendation model	Music recommendation	Effectively solve the problem of cold start and new user recommendation	Too much dependency on MIDI content data
Ziemer et al. [24]	Classification model based on third octave mixing analysis	DJ sound classification	Help with DJ retrieval tools	The accuracy rate is 73%, with room for improvement
Mao et al. [12]	Music-CRN model	Music classification and recommendation	The effect of music classification and recommendation is good	Rely on audio content feature learning
Xu et al. [19]	Hierarchical multi-information fusion recommendation method	Music recommendation	Fully capture information internal and external associations	Implementation can be complex

by equation (3.1).

$$D_{l+1} = f(W_l D_l + B_l) \quad (3.1)$$

In equation (3.1), D_l is the output of the previous layer network; W_l indicates the weight term; B_l denotes the bias term of bias, and ReLu is selected as the activation function f ; l represents the number of layers. If d is set as the input dimension, the number of neurons in each layer is m , and there is a l layer network, then the total parameter complexity of the DNN is shown in equation (3.2).

$$P(dl) = d \times m + m + (m^2 + m) \times (L_d - 1) \quad (3.2)$$

In this study, the CN is applied to the explicit feature cross, and the network output expression of each layer

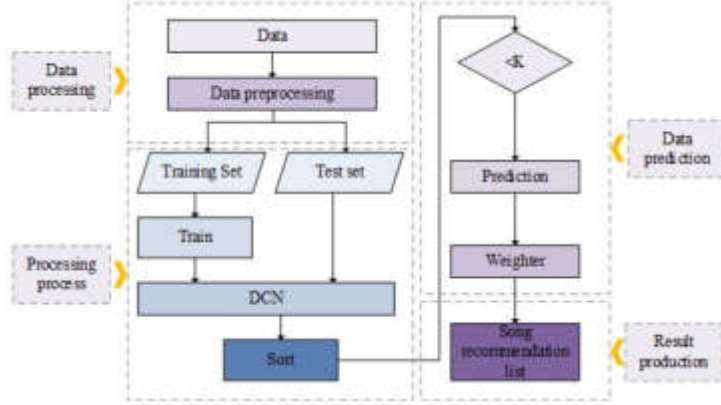


Fig. 3.2: Flow chart of music recommendation model based on DCN

is shown in equation (3.3).

$$x_{l+1} = x_0 x_l^T w_l + B_l + x_l \quad (3.3)$$

In equation (3.3), x_l is the output of the previous layer network; w_l and B_l represent the connection parameters between the two layers of networks; $x_0 x_l^T w_l$ denotes the feature crossover completed in the $l + 1$ layer, and all variables in the above equation are column vectors. The residual between the output of the fitting layer and the previous output, plus the input data x_l of that layer, can be regarded as the residual of the two-layer network. Next, in the connection layer, the final outputs of the two networks are connected, and after weighted summation, the final probability value is generated by the Sigmoid function, as shown in equation (3.4).

$$x_{\text{stack}} = \text{concat} (w [C_{L_1}^T, D_{L_1}^T]) \quad (3.4)$$

In equation (3.4), $C_{L_1}^T$ and $D_{L_1}^T$ are the outputs of the CN and DNN, respectively; L_1 and L_2 respectively mean the number of layers in two networks; w expresses the weight parameter of the CN. DCN can learn the interactions between effective features and has lower computational costs. Therefore, the study applies DCN to music recommendation models to reduce the complexity of the music recommendation model. The flowchart of the music recommendation model based on DCN algorithm is shown in Figure 3.2.

From Figure 3.2, the music recommendation model based on DCN algorithm first needs to desensitize user information, extract hidden features of music from it, and build an information matrix. Then the music data set is split and passed into the DCN classifier for parameter setting. Parameters are initialized by equation (6) to reduce the dependence between parameters.

$$L = \text{sqrt} \left(\frac{6}{n_{\text{input}} + n_{\text{output}}} \right) \quad (3.5)$$

In equation (3.5), L indicates the range of uniform distribution; n_{input} denotes the amount of input units for the weight tensor; n_{output} represents the amount of output units of the weight tensor. Then at the connection layer, the Sigmoid function is used to calculate the probability output, the expression of which is shown in equation (3.6).

$$S(x) = \left(\frac{1}{1 + e^{-x}} \right) \quad (3.6)$$

In equation (3.6), $S(x)$ means probability and x indicates variable parameters. After the DCN model is constructed, the Adam optimizer is used to calculate the gradient of the loss function and update the parameters.

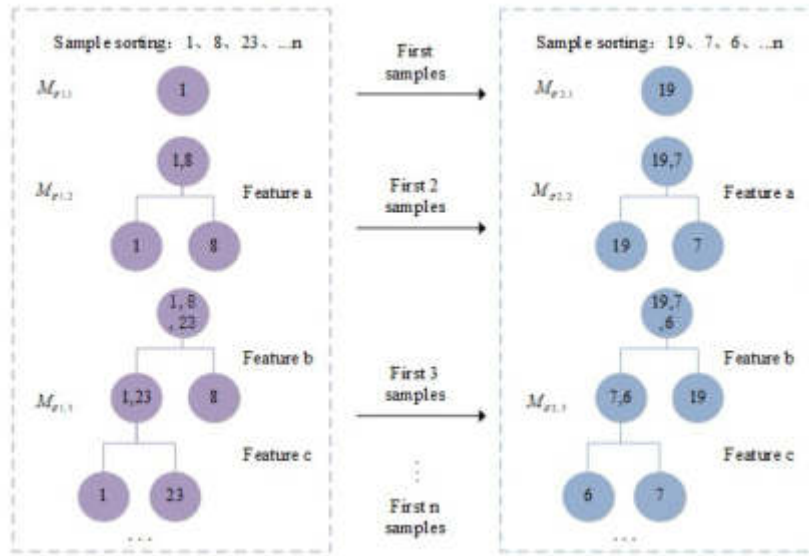


Fig. 3.3: CatBoost model structure

After the model configuration is completed, the number of iterations is set, the model is trained by the number of samples in each iteration, and the music recommendation list is finally obtained. Through the above steps, the music recommendation model based on DCN algorithm can recommend music according to the needs of different users, and ensure high recommendation accuracy.

3.2. Design of a dual filtering music recommendation model integrating CatBoost and DCN.

In traditional DCN music model recommendations, the number of music songs is large, the types are diverse, and the duration is not uniform. Moreover, the DCN model cannot effectively process discrete features, making it difficult for the model to provide personalized recommendations for users [15, 22]. Based on this dilemma, the study integrates the integrated learning CatBoost classification model with the DCN to construct a music recommendation model based on the DCN-CAT algorithm. CatBoost is a type of decision tree that can efficiently and reasonably process categorical features, thereby improving the accuracy of the algorithm [8]. The CatBoost symmetric tree structure is shown in Figure 3.3.

From Figure 3.3, the core of CatBoost algorithm is the design of symmetric tree, and it will build an initial tree structure according to the selected sample in the first iteration, and then determine the value of each leaf node through calculation. This initial tree structure is reused in subsequent iterations to continuously optimize the model. The design form of CatBoost enables it to efficiently process category-type features, thus making up for the shortcomings of DCN model in processing discrete features, and finally forming a complete recommendation model through multiple iterations, improving the accuracy of music recommendation. At the same time, CatBoost proposed the use of TS-based ordered transport stream (Ordered TS) and a new classification feature processing algorithm to solve the problem of target leakage and prediction offset in Boosting algorithm, and improve training speed and accuracy. The Ordered TS coding principle is shown in Figure 3.4.

From Figure 3.4, in Ordered TS coding principle, the classification feature values of samples are converted into a sequential coding. For a particular sample, its coded value is calculated based on the sample that comes before it. When one of the classification features of the sample is the same as that of the previous sample, the corresponding Ordered TS encoding value is adopted.

This method can help solve the problem of target leakage and prediction deviation in Boosting algorithm, and improve the training speed and prediction accuracy of the model by considering the order relationship between samples. Ordered TS code principle in the sample x_i , under the classification feature k' is $x_i^{k'}$, and the

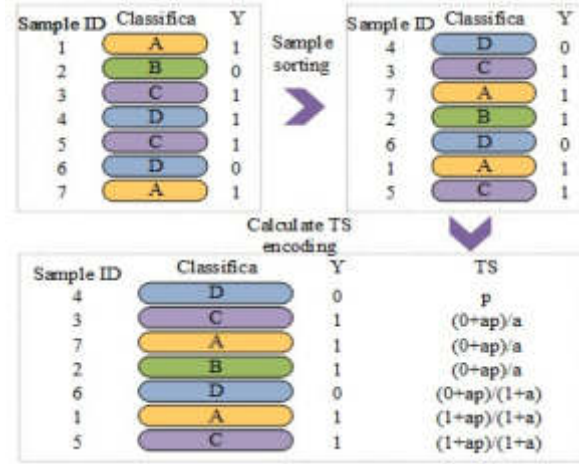


Fig. 3.4: Ordered TS encoding principle

encoding value of $x_{i'}^{k'}$ is calculated based on the sample D_σ that ranks first in the value. The TS encoding value of the same sample as $x_{i'}^{k'}$ under the classification feature k in D_σ is the Ordered TS encoding value. CatBoost uses the target count method in the gradient enhancement algorithm to group categories and estimate the expected target value of each category, processing classification features with minimal information loss. The specific equation for estimating the expected target is shown in equation (3.7).

$$\hat{x}_{k'}^{i'} \approx E(y/x = x_{k'}^{i'}) \tag{3.7}$$

In equation (3.7), y means the expected goal. When the i' feature values of other samples are equal to $x_{k'}^{i'}$, the expected value of this category is used to replace the i' feature of the k' sample, which means that the discrete feature is reassigned. Because $x_{k'}^{i'}$ is calculated from the target values of the samples, there will be conditional biases when splitting the test set and training set. The expression for obtaining a prior value is shown in equation (3.8).

$$P_1(y = 1/x^i = C) = 0.5 \tag{3.8}$$

In equation (3.8), P_1 denotes the TS value; C indicates the category. Therefore, it assumes that there is a category feature with all feature values taking different values, the numerical values for replacing category features in the classification category are shown in equation (3.2).

$$\hat{x}_{k'}^{i'} = \frac{yC + aP_1}{1 + a} \tag{3.9}$$

In equation (), a represents a constant, but for the test set, if all TS values are 0.5, it is not possible to classify and predict the test data. Therefore, a threshold is used for classification, and the threshold expression is shown in equation (11).

$$t = \frac{0.5 + aP_1}{1 + a} \tag{3.10}$$

In equation (), t denotes the threshold. Afterwards, the training samples are randomly sorted, and the prior values and weight coefficients of the prior values are added to the mean of the category labels before the samples, thereby reducing the impact of low-frequency category features. The expression for defining the encoding value

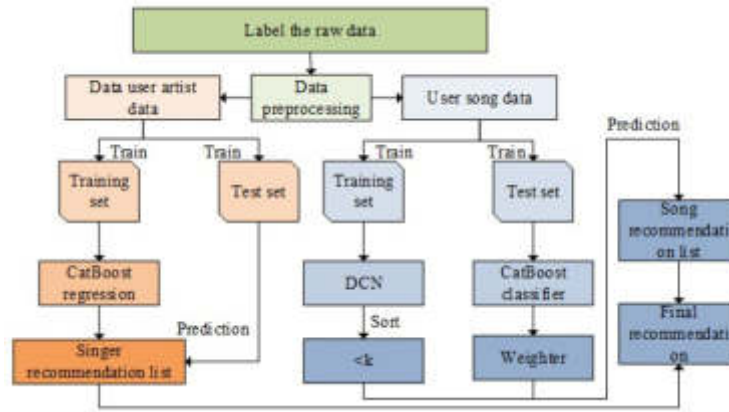


Fig. 3.5: Flow chart of hybrid recommendation model

is shown in equation (3.11).

$$\hat{x}_{k'}^{i'} = \frac{\sum_{x_j \in D_k} I[x_j^{i'} = x_{k'}^{i'}] \square y_j + aP_1}{\sum_{x_j \in D_{k'}} I[x_j^{i'} = x_{k'}^{i'}] + a}, D_{k'} = \{x_j : \sigma(j) < \sigma(k')\} \quad (3.11)$$

In equation (3.11), D refers to all datasets available for model training; $D_{k'}$ indicates a subset of D ; σ represents a constructed random sequence; y_j is the eigenvalues of sample j . To achieve more accurate and personalized recommendation algorithms, music recommendation is divided into two parts: song and singer recommendation. In the dataset prediction section, mixed classification models and regression models are used for prediction. This generates a music recommendation list for different users, and the hybrid recommendation model flowchart is shown in Figure 3.5.

From Figure 3.5, the hybrid recommendation model proposed in this study combines deep learning and machine learning technologies. The model achieves accurate recommendation through four main stages. The first stage is to label the raw data to extract key information and hidden features. In addition to basic song information and user behavior data, the research also focuses on users' historical listening records, song emotional labels, rhythms, genres and other additional information, so as to provide a more comprehensive view of user preferences and song characteristics, which is conducive to accurate recommendation in the future. At the same time, in the process, the research also extracts singer information from the original data, forms a new singer data set, and splits it into a training set and a test set. Then, in the second stage, the song training data is input into DCN and CatBoost classifiers for training and parameter tuning. DCN, with its powerful feature crossover ability, helps capture complex relationships between songs. In order to improve the generalization ability and accuracy of the model, additional information such as sentiment analysis data of songs, user comments, and community tags are also introduced at this stage of the study to enable the model to understand songs and user preferences from multiple dimensions. At the same time, the singer training set is input into the CatBoost regression model for training, which is able to efficiently process the classification features and prevent overfitting with specific enhancement techniques, thus ensuring the accuracy of the recommendations. In the third stage, the trained DCN model is used to predict the song test set, generate a list of predicted values, and set the filtering range to form a new data set. The trained CatBoost regressor is used to predict the score of the singer test set and generate the singer recommendation list. This process incorporates the user's recent listening behavior and feedback, which is used as an important reference for dynamically adjusting the recommendation list. Then, the CatBoost regressor is used to predict the score of the singer test set and generate the singer recommendation list. Finally, in the fourth stage, CatBoost classifier is used for secondary classification prediction, and additional information such as users' social network information and geographical location data is integrated in this stage, so as to provide users with more personalized and

Table 4.1: Performance comparison of various algorithms

Dependent variable	Variable	Dependent variable	Variable
User information	User-id	Operation information	User-id
	Gender		Song-id
	Bd		Source-screen-name
	City		Source-system -tab
	Registered -via Registration -time Expiration-date		Source-type
Song information	Song-id	Song additional information	Song-id
	Genre-ids		Name
	Artist-name		Isrc
	Composer		
	Lyricist		
	Song-length Isrc		

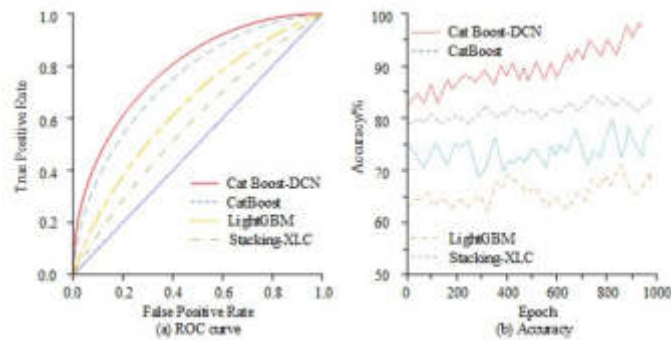


Fig. 4.1: ROC and accuracy of four algorithms results

localized music recommendations to further refine the recommendation results. Finally, the recommended list of songs and artists is combined to generate personalized music recommendations for users.

4. Results.

4.1. Parameter design and performance evaluation of CatBoost-DCN. After constructing the CatBoost-DCN hybrid music recommendation model, to verify the superiority of the constructed CatBoost-DCN algorithm, it attempted to compare CatBoost-DCN with CatBoost, LightGBM, and Stacking-XLC algorithms in the same dataset. The experimental dataset included four parts: user, song, operation and song additional information tables. The dataset variables for music recommendation are shown in Table 4.1.

Table 4.1 shows all the data sets involved in this experiment. The experimental data set included more than 2,000 user song operation records, data labels on whether users listen to songs repeatedly, 150 user attribute information and more than 1,000 song information, ensuring the comprehensiveness and diversity of the data. Then, the four algorithms were tested in the data set, and the ROC curve, accuracy rate, recall rate, error value and other indicators of the four algorithms in the data set were compared and analyzed. The comparison results of ROC curve and accuracy of the four algorithms are shown in Figure 4.1.

Figure 4.1 shows the ROC curves and accuracy plots of CatBoost-DCN, CatBoost, LightGBM, and Stacking-XLC algorithms. As shown in Figure 4.1 (a), compared to the other three algorithms, the CatBoost-DCN algorithm had a maximum area of approximately 0.85 in the ROC curve, while the other three algorithms had areas of 0.75, 0.53, and 0.51, respectively. The results showed that CatBoost-DCN algorithm had higher

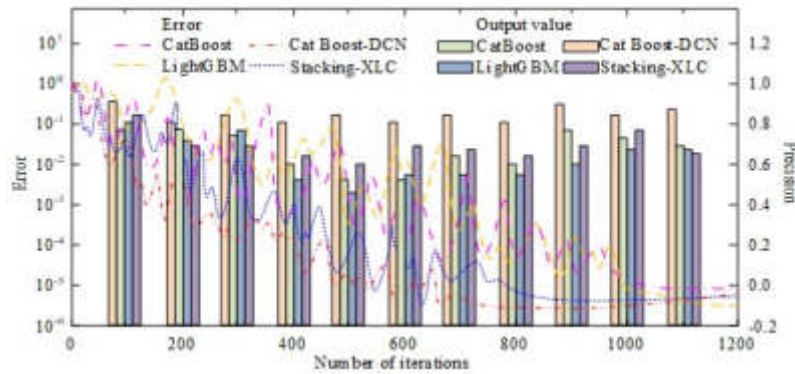


Fig. 4.2: Performance comparison of various models

performance in distinguishing positive and negative samples, and the classification effect was better. From Figure 4.1 (b), from the comparison results of algorithm accuracy, after stable training, CatBoost-DCN had the highest accuracy, at 96.13%, which was 19.91% higher than CatBoost, 29.69% higher than LightGBM, and 12.71% higher than Stacking-XLC. The results showed that the CatBoost-DCN algorithm performed well in the classification task and could correctly classify most samples into the correct categories. Therefore, from the two dimensions of area under the ROC curve and accuracy, CatBoost-DCN showed higher performance than the other three algorithms. This shows that the CatBoost-DCN algorithm can get more accurate classification results in practical applications. The advantages of CatBoost-DCN may come from its unique algorithm design and optimization strategies, which enable the algorithm to extract features more efficiently, reduce the risk of overfitting, and improve generalization ability when dealing with complex data. As shown in Figure 4.2, the error and accuracy of the four algorithms would be analyzed next.

Figure 4.2 shows the error and precision analysis of the four algorithms. The broken line section represents the error curve of the algorithm, and the bar chart section represents the precision of the algorithm. As shown in Figure 4.2, the training error of CatBoost-DCN, LightGBM, CatBoost, and Stacking-XLC was 0.013%, 0.065%, 0.034%, and 0.023%, respectively. The results demonstrated the high precision and low error of CatBoost-DCN algorithm in the process of model training, indicating that the algorithm can fit the data more accurately and reduce the prediction bias. In the precision comparison, CatBoost-DCN had a precision of 95%, approximately 19%, 17%, and 10% higher than LightGBM, CatBoost, and Stacking-XLC, respectively. The results showed that CatBoost-DCN algorithm had excellent precision in classification tasks and could identify all kinds of samples more accurately. In summary, CatBoost-DCN is significantly better than the other three algorithms in the two key indicators of error rate and precision, which indicates that CatBoost-DCN algorithm can provide more precise and reliable classification results in practical applications. Next, the recall and accuracy of the four algorithms were analyzed and sorted, and the results are shown in Figure 4.3.

In Figure 4.3 (a) of the PR curve, the area of CatBoost-DCN was 0.89, with the largest area, approximately 0.33 higher than LightGBM, 0.17 higher than CatBoost, and 0.11 higher than Stacking-XLC. The results showed that the CatBoost-DCN algorithm had excellent performance in the comparison of PR curves, and could identify positive samples more effectively while maintaining a low false positive rate. As shown in Figure 4.3 (b), the recall rates of CatBoost-DCN, LightGBM, CatBoost, and Stacking-XLC algorithms were 92.7%, 71.1%, 76.1%, and 83.2%, respectively. This data showed that CatBoost-DCN algorithm could find out the real positive class samples more comprehensively and reduce the cases of missing reports in classification tasks. According to the comprehensive analysis of Figure 4.3(a) and Figure 4.3(b), CatBoost-DCN showed obvious advantages from the comparison results of PR curve and recall rate. This advantage may be due to its advanced algorithm design and fine parameter tuning, so that the algorithm can more accurately capture the internal structure of the data when dealing with classification problems, thus providing more reliable and comprehensive classification results. The results also further confirm that CatBoost-DCN algorithm is expected to obtain better classification

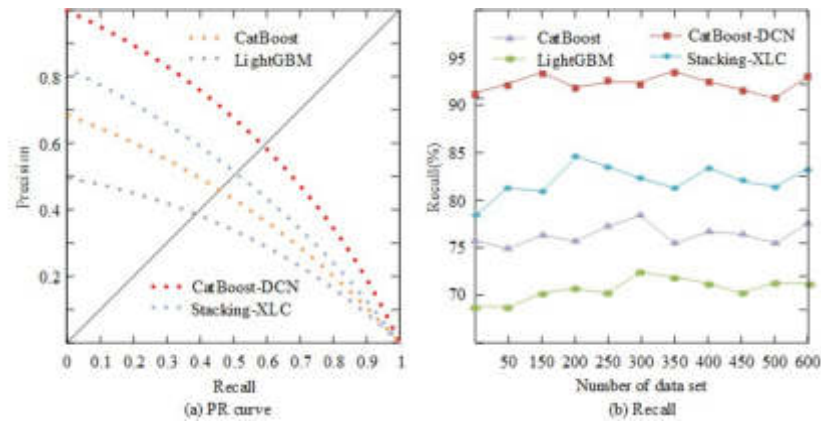


Fig. 4.3: Recall and precision of four algorithms

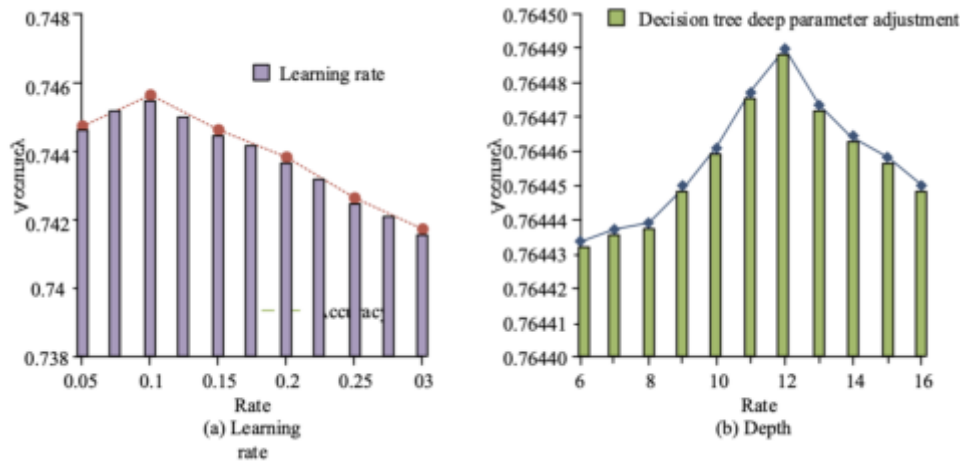


Fig. 4.4: Recall and precision of four algorithms

performance in practical applications.

4.2. Evaluation of CatBoost-DCN recommendation model results. To compensate for the low accuracy of a single algorithm, a CatBoost-DCN dual filtering music hybrid recommendation model was constructed. For the binary classification problem of the CatBoost-DCN hybrid model, four indicators, precision, recall, AUC, and accuracy, were used to evaluate and analyze the performance of CatBoost-DCN. To obtain the optimal model and parameters, the above four indicators were used as the evaluation criteria for the model, and testing was conducted based on learning rate and depth. The results are shown in Figure 4.4.

Figure 4.4 shows the final results of learning rate tuning and decision tree tuning. Figure 4.4 (a) shows that when the learning rate was 0.1, the optimal model score was 0.7516, and the overall learning rate showed a trend of increasing first and then decreasing. Figure 4.4 (b) indicated that when the decision tree depth was 12, the optimal model score could be obtained, which was 0.76449. When the decision tree depth was 6, the model score showed an upward trend. When the decision tree depth was within the 9-12 range, the increase in model score significantly increased and reached the highest value. When the decision tree depth was after 12, the score showed a downward trend. The original test set consisted of 1342 pieces of data. The predicted values of CatBoost and DCN in the hybrid model were weighted and fused to obtain the final probability. The

Table 4.2: CatBoost-DCN parameter settings

Song recommendation		Singer recommendation	
Parameter name	Parameter value	Parameter name	Parameter value
Iterations	1000	Iterations	600
Depth	12	Depth	8
Learning_rate	0.1	Learning_rate	0.2
Eval_metric	AUC	Eval_metric	R2
Max_ctr_complexity	2	Loss_function	RESE
Loss_function	Logloss	12_leaf_reg	3
Boosting_type	Plain	Bootstrap_type	Bernoulli
12_leaf_reg	6	Border_count	32
Bootstrap_type	Bernoulli	Random_seed	123
Border_count	31	Task_type	GPU
One_hot_max_size	255		
Random_seed	123		
Task_type	GPU		

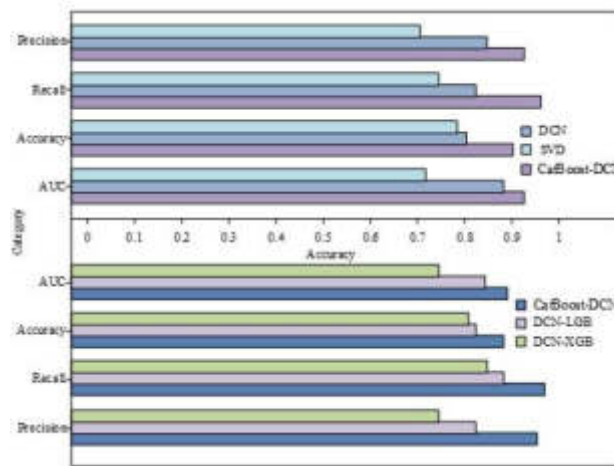


Fig. 4.5: Model comparison test

probabilities were then sorted to form a corresponding recommendation list, and finally recommendations were made to users. The optimal parameter settings for the CatBoost-DCN model for final singer recommendation and song recommendation are shown in Table 4.2.

Table 4.2 shows the CatBoost-DCN parameter settings. The number of trees in the song recommendation bar was 1000. When the depth of the tree was 12, the maximum feature combination tree was 2, the regularization coefficient was 6, the number of numerical feature divisions was 31, and the maximum number of unique hot codes was 255. The number of trees in the singer recommendation column as 600. When the depth of the tree was 8, the regularization coefficient was 3, and the numerical feature segmentation tree was 32. After the parameters were adjusted, the model was called to predict the test set, and the top eight singers with the highest scores were recommended to users. In addition, to further analyze the performance of the CatBoost-DCN model, a comparative analysis was conducted among the CatBoost-DCN hybrid model, the DCN model, and the SVD model. At the same time, to verify the superiority of CatBoost-DCN, models with different algorithms such as DCN-LGB and DCN-XGB were selected for comparison. The results are shown in Figure 4.5.

The bar chart in the upper half of Figure 4.5 shows the performance comparison results among DCN, SVD,

and CatBoost-DCN models. The results showed that the AUC area of CatBoost-DCN was 0.93, the precision was 0.91, the accuracy was 0.95, and the recall was 0.92. The scores of CatBoost-DCN were higher than those of DCN and SVD models. Compared with the DCN model, the CatBoost-DCN hybrid model improved in all indicators, with an increase of approximately 7.2% in AUC, 3.4% in accuracy, 2.2% in precision, and 2.1% in recall. The above results showed that the combination of CatBoost and DCN could effectively improve the predictive ability and stability of the model. The lower half of Figure 10 shows the comparison of indicator performance among different mixed models. The AUC area of CatBoost-DCN was approximately 0.05% more than DCN-LGB and 0.15% more than DCN-XGB. The precision of CatBoost-DCN was approximately 0.07% higher than DCN-LGB and 0.09% higher than DCN-XGB. The accuracy of CatBoost-DCN was approximately 0.11% higher than DCN-LGB and 0.15% higher than DCN-XGB. The recall rate of CatBoost-DCN was approximately 0.16% higher than DCN-LGB and 0.23% higher than DCN-XGB. The above data fully proved the superiority of CatBoost-DCN hybrid model in various indicators. CatBoost-DCN showed strong prediction and classification ability both in the comparison of single models and in the competition of mixed models. To sum up, CatBoost-DCN hybrid model has significant advantages in the field of music recommendation, and its overall performance improvement is due to the perfect combination of CatBoost and DCN. The hybrid model can not only extract features more effectively and reduce the risk of overfitting, but also improve the generalization ability and prediction accuracy of the model.

5. Discussion. From the simulation results, CatBoost-DCN model showed significant advantages in several evaluation indicators. Compared with the other three algorithms (CatBoost, LightGBM, and Stacking-XLC), CatBoost-DCN had significant improvements in ROC curve area, accuracy, error rate, and precision. This is mainly due to CatBoost-DCN's unique algorithm design and optimization strategy, which enables it to extract features more efficiently, reduce the risk of overfitting, and improve generalization. The results of this study were significantly improved compared with the indicators of the music recommendation model proposed by Yun et al. [21]. In comparison with the DCN model and SVD model, CatBoost-DCN also showed excellent performance. The AUC area, accuracy, precision and recall rate were higher than those of the two models. In particular, compared with the DCN model, CatBoost-DCN improved in all indicators, with AUC improving by about 7.2%, accuracy by 3.4%, precision and recall by 2.2% and 2.1%, respectively. These results showed that the combination of CatBoost and DCN could indeed significantly improve the predictive power and stability of the model. The research results were compared with the research results of the music recommendation model proposed by Wang team in 2023, and it was found that the overall performance of the proposed model was better [18]. In addition, the performance of different hybrid models was compared. The results showed that CatBoost-DCN was superior to DCN-LGB and DCN-XGB models in AUC area, precision, accuracy and recall ratio. This result further confirmed the superiority of CatBoost-DCN hybrid model. Compared with other relevant studies, the application of CatBoost-DCN model in the field of music recommendation has significant advantages. This is mainly reflected in its overall performance improvement, including higher accuracy, lower error rate, and better generalization ability [5]. These advantages enable the CatBoost-DCN model to provide users with more accurate and personalized music recommendation services.

In summary, the CatBoost-DCN model shows excellent performance in music recommendation applications. Its advantages come from unique algorithm design and optimization strategies, which enable the model to extract features more effectively, reduce overfitting risks, and improve generalization ability when dealing with complex data. The results of comparison with other algorithms and models further confirm the superiority of CatBoost-DCN. Therefore, in the field of music recommendation, CatBoost-DCN model is expected to achieve better recommendation results and user satisfaction. In addition, because the CatBoost-DCN model performs well in processing complex data with a large number of category characteristics and numerical characteristics, it also has a wide range of application potential in other recommendation fields such as e-commerce product recommendation, video content recommendation, and social network friend recommendation.

6. Conclusion. The personalized demand for music recommendations from different users on music platforms is increasing, and traditional music recommendation models have shortcomings such as long-time consumption and insufficient personalization. This study attempted to integrate the CatBoost algorithm with DCN and constructed a CatBoost-DCN hybrid model to recommend personalized music to meet the needs of a large number of network users. During training, CatBoost, LightGBM, and Stacking-XLC algorithms were selected

for performance analysis with CatBoost-DCN. In performance analysis, this model was superior to CatBoost, LightGBM and Stacking-XLC algorithms in ROC area, accuracy, error rate, recall area, and precision. Among them, the ROC area of CatBoost-DCN was 0.85, the accuracy was 96.13%, the error rate was only 0.013%, the recall area was 0.89, the precision rate was 95%. Compared with single and other models, CatBoost-DCN also performed well in AUC area, precision, accuracy and recall. The results above showed that the CatBoost-DCN hybrid model showed excellent performance in personalized music recommendation, which is significantly better than the traditional recommendation algorithm. This study found that by integrating CatBoost and DCN, the research successfully improved the accuracy and efficiency of the recommendation system, and effectively met the personalized needs of Internet users for music. However, there are also shortcomings in the research. At present, the CatBoost-DCN model does not take into account how user interests may change over time. Future studies can further refine the model by introducing techniques such as time series analysis to more accurately capture users' dynamic music preferences.

REFERENCES

- [1] J. ANITHA AND M. KALAIARASU, *A new hybrid deep learning-based phishing detection system using mcs-dnn classifier*, Neural Computing and Applications, 34 (2022), pp. 5867–5882.
- [2] S. BAGYARAJ, R. TAMILSELVI, P. B. M. GANI, AND D. SABARINATHAN, *Brain tumour cell segmentation and detection using deep learning networks*, IET Image Processing, 15 (2021), pp. 2363–2371.
- [3] X. CAI, Z. HU, AND J. CHEN, *A many-objective optimization recommendation algorithm based on knowledge mining*, Information Sciences, 537 (2020), pp. 148–161.
- [4] S. CHAPANERI AND D. JAYASWAL, *Structured gaussian process regression of music mood*, Fundamenta Informaticae, 176 (2020), pp. 183–203.
- [5] R. CHHEDA, D. BOHARA, R. SHETTY, S. TRIVEDI, AND R. KARANI, *Music recommendation based on affective image content analysis*, Procedia Computer Science, 218 (2023), pp. 383–392.
- [6] A. ELBIR AND N. AYDIN, *Music genre classification and music recommendation by using deep learning*, Electronics Letters, 56 (2020), pp. 627–639.
- [7] D. C. N. FABRIS, E. H. MIGUEL, R. VARGAS, R. B. CANTO, M. D. O. C. VILLAS-BOAS, AND O. PEITL, *Microstructure, residual stresses, and mechanical performance of surface crystallized translucent glass-ceramics*, Journal of the European Ceramic Society, 42 (2022), pp. 4631–4632.
- [8] Y. FANG, B. LUO, T. ZHAO, D. HE, B. JIANG, AND Q. LIU, *St-sigma: Spatio-temporal semantics and interaction graph aggregation for multi-agent perception and trajectory forecasting*, CAAI Transactions on Intelligence Technology, 7 (2022), pp. 744–757.
- [9] M. GUPTA, R. K. SINGH, AND S. SINGH, *G-cocktail: An algorithm to address cocktail party problem of gujarati language using catboost*, Wireless Personal Communications, 125 (2022), pp. 261–280.
- [10] G. G. JENEFER AND A. J. DEEPA, *Diabetes disease prediction using firefly optimization-based cat-boost classifier in big data analytics*, Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, 44 (2023), pp. 9943–9954.
- [11] J. LU, Z. SHEN, H. YANG, AND S. ZHANG, *Deep network approximation for smooth functions*, SIAM Journal on Mathematical Analysis, 53 (2021), pp. 5465–5506.
- [12] Y. MAO, G. ZHONG, H. WANG, AND K. HUANG, *Music-crn: an efficient content-based music classification and recommendation network*, Cognitive Computation, 14 (2022), pp. 2306–2316.
- [13] C. M. MURZYN, E. R. JANS, AND M. D. CLEMENSON, *Spears: A database-invariant spectral modeling api*, Journal of Quantitative Spectroscopy & Radiative Transfer, 277 (2022), pp. 77–103.
- [14] M. G. M. PANDEY, *Predicting potential residues associated with lung cancer using deep neural network*, Mutation research-Fundamental and Molecular Mechanisms of Mutagenesis, 822 (2021), pp. 1386–1964.
- [15] C. SARKAR, S. C. SHIT, D. Q. DAO, J. LEE, N. H. TRAN, AND R. SINGURU, *An efficient hydrogenation catalytic model hosted in a stable hyper-crosslinked porous-organic-polymer: from fatty acid to bio-based alkane diesel synthesis*, Green Chemistry, 22 (2020), pp. 2049–2268.
- [16] L. STOCZYNSKI, B. L. BROWN, AND S. R. MIDWAY, *Landscape features and study design affect elements of metacommunity structure for stream fishes across the eastern u.s.a*, Freshwater Biology, 66 (2021), pp. 1736–1750.
- [17] Z. SUN, B. WU, Y. WANG, AND Y. YANG, *Sequential graph collaborative filtering*, Information Sciences, 592 (2022), pp. 244–260.
- [18] D. WANG, X. ZHANG, Y. YIN, D. YU, G. XU, AND S. DENG, *Multi-view enhanced graph attention network for session-based music recommendation*, ACM Transactions on Information Systems, 42 (2023), pp. 1–30.
- [19] J. XU, M. GAN, AND X. ZHANG, *Mmusic: a hierarchical multi-information fusion method for deep music recommendation*, Journal of Intelligent Information Systems, 61 (2023), pp. 795–818.
- [20] N. YADAV, A. K. SINGH, AND S. PAL, *Improved self-attentive musical instrument digital interface content-based music recommendation system*, Computational Intelligence, 38 (2022), pp. 1232–1257.
- [21] W. U. YUN, L. JIAN, AND M. A. YANLONG, *A hybrid music recommendation model based on personalized measurement and game theory*, Chinese Journal of Electronics, 32 (2023), pp. 1319–1328.
- [22] J. ZHAO, D. DIAZ-DUSSAN, M. WU, Y. PENG, AND R. NARAIN, *Correction to dual-cross-linked network hydrogels with*

- multiresponsive, self-healing, and shear strengthening properties*, *Biomacromolecules*, 22 (2021), pp. 800–810.
- [23] J. ZHOU, T. CHENG, AND X. LI, *Intronic noncoding rna expression of dcn is related to cancer-associated fibroblasts and nsclc patients prognosis*, *Journal of Thoracic Oncology*, 16 (2021), pp. 360–370.
- [24] T. ZIEMER, P. KIATTIPADUNGKUL, AND T. KARUCHIT, *Music recommendation based on acoustic features from the recording studio*, *The Journal of the Acoustical Society of America*, 148 (2020), pp. 2701–2701.

Edited by: Zhengyi Chai

Special issue on: Data-Driven Optimization Algorithms for Sustainable and Smart City

Received: Nov 27, 2024

Accepted: Jul 1, 2024



DESIGN OF TEST TURNTABLE BASED ON FUZZY PID ALGORITHM AND ITS ERROR CORRECTION

LI TANG* AND ZHOU LIANGFU†

Abstract. In order to meet the requirements of antenna for high bearing capacity, high positioning accuracy and speed stability of the test turntable, a test turntable with simple operation is designed. First of all, the structure of the turntable is introduced. The turntable adopts the form of gear transmission and multi-turn absolute encoder angle measurement. The mechanical simulation analysis of the turntable table is carried out, and the static performance and modal characteristics of the turntable are analyzed. Then, the servo control system of the turntable is introduced. The multi-turn absolute encoder is used as the detection element of the actual position of the turntable. The target position is sent to the controller by the upper computer through the serial port. The controller uses the field programmable logic gate array (FPGA) as the core, and the encoder, controller and control object constitute the position loop; In the software design, the fuzzy PID algorithm is used to replace the traditional PID algorithm. Finally, the turntable accuracy test platform is built, and the autocollimator and polyhedral prism are used to obtain the angle measurement error. The test results show that the control accuracy is better than $\pm 0.01^\circ$, the system position accuracy is better than 2.5', and the turntable positioning accuracy and rotational speed stability have been effectively improved.

Key words: Servo; Turntable; Fuzzy control; FPGA.

1. Introduction. In the design of high-precision turntables, commonly used forms of power transmission include direct drive, worm gear transmission, and gear transmission [6]. The worm gear transmission has the advantages of large reduction ratio, smooth transmission, low noise, and large bearing capacity; Gear transmission has the advantages of accurate transmission ratio, stability, high efficiency, high working reliability, and long service life. The load-bearing capacity, position accuracy, smoothness of speed, and system stability of the test turntable will directly affect the testing effect of the test turntable. Therefore, the control strategy for the motor in the control system of the test turntable is crucial. At present, there are various control algorithms for motors [13, 21]. Le K M [8] proposes an algorithm based on improved PI control current to improve the performance of the motor. It determines the control parameters through the parameters of the motor and the speed of operation; Tran H N [15] proposed a method based on an effective phase compensator to improve the accuracy of the motor; Zhai Yan [20] and Changjun Zhao [22] propose an algorithm based on fuzzy adaptive control, which performs fuzzy control on classical PID and adjusts the control parameters according to different situations automatically. Sliding mode variable structure control algorithms have good control effects on nonlinear factors. [12, 1, 4]

The author has designed a high-precision gear transmission turntable and conducted mechanical analysis on the turntable using ANSYS software. The maximum deformation of the loading table is 0.026mm, and the maximum stress is 0.95MPa, both of which are far less than the strength limit of the material; At the same time, a control strategy based on the fuzzy PID algorithm is adopted. Fuzzy control can formulate control strategies based on engineering experience, which is suitable for nonlinear and uncertain systems such as turntable control systems. Fuzzy control is used to adjust the proportional integral derivative coefficients of PID online, which not only has the advantages of simple and reliable PID control, but also enables the turntable to better cope with sudden disturbances and has excellent control performance. And the error effect was verified using a polyhedron prism, with a control accuracy better than $\pm 0.01^\circ$ and a system position accuracy better than 2.2', meeting the requirement of antenna testing 3'. At the same time, the speed stability of the turntable was

*Industrial Software Engineering Technology Research and Development Center of Jiangsu Education Department, Nanjing Vocational University of Industry Technology, Nanjing 210023, China (Corresponding author)

†Industrial Software Engineering Technology Research and Development Center of Jiangsu Education Department, Nanjing Vocational University of Industry Technology, Nanjing 210023, China

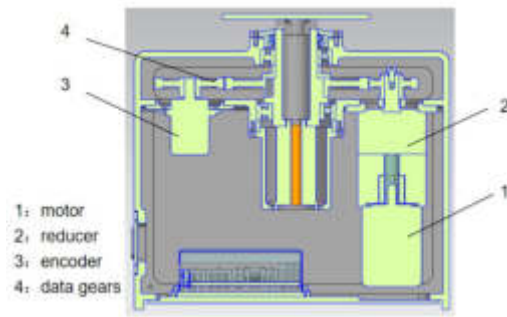


Fig. 2.1: Turntable structure

Table 2.1: Main parameters of encoder

Shaft diameter	Lines per revolution	revolution	resolution ratio	repeatability
10 × 19 mm	262144	4096	4.9''	0.002°

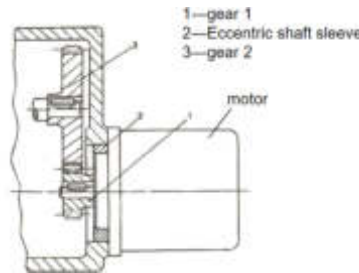


Fig. 2.2: Adjusting the backlash of the eccentric shaft sleeve

effectively improved.

2. Turntable structure design. The turntable mainly consists of a pair of transmission gears, data gears, reducer, absolute multi-turn encoder, motor, etc. The structural diagram of the turntable is shown in Figure 2.1.

According to the load requirements of the turntable, through analysis and calculation, combined with the Mechanical Design Manual, the gear module is determined, in which the gear ratio is 23:181, the data gear ratio is 1:8, and the speed ratio of the reducer is 100. The large total deceleration ratio ensures the load-bearing capacity of the turntable.

The absolute multi-turn encoder is used to measure the actual position of the turntable, and the main technical parameters of the encoder are shown in Table 2.1.

The encoder is installed coaxially with the data gear, and the data gear is installed parallel to the output gear, with a data gear ratio of 1:8. So compared to the spindle, the encoder resolution reaches $4.9''/8=0.62''$. However, the backlash of gears can affect the positioning accuracy and stability of the system, so it needs to be effectively overcome. There are different methods for adjusting the backlash of different types of gear transmission pairs. The eccentric shaft sleeve adjustment method is used here, as shown in Figure 2.2. By adjusting the eccentric shaft sleeve, the center distance between the input and output gears can be changed.

Considering the requirements of matching the torque, rated speed, and inertia of the motor, a three-phase synchronous servo motor is chosen here.

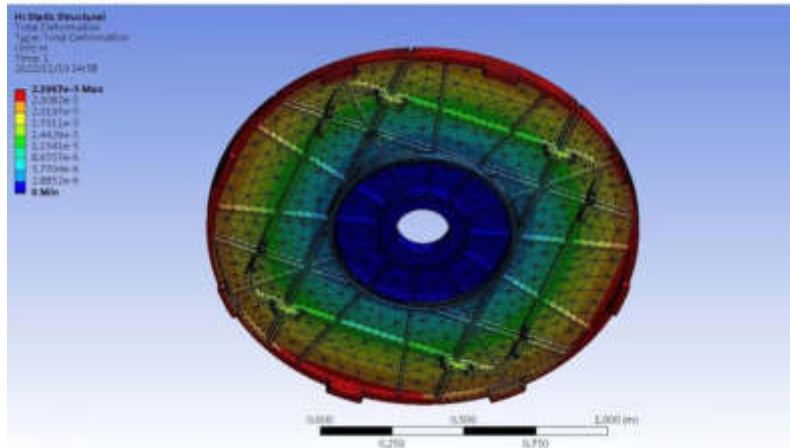


Fig. 3.1: Table deformation analysis

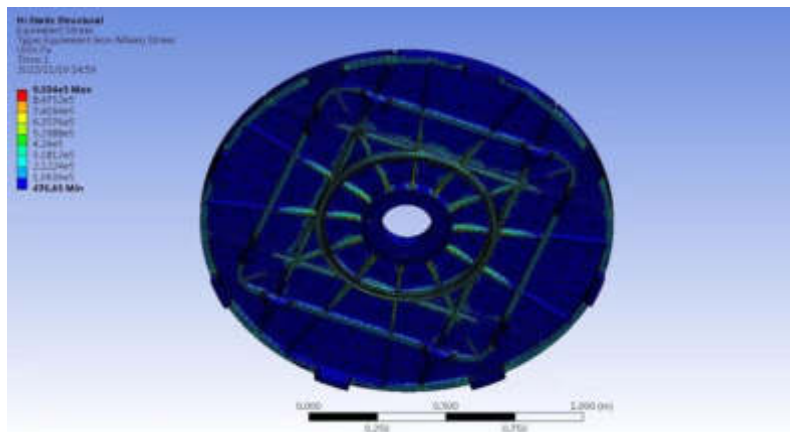


Fig. 3.2: Turntable stress analysis

3. Mechanical simulation analysis. The flatness and end face runout of the turntable' table determine the installation accuracy of the antenna load and have a significant impact on the testing effect. Under load, the deformation and vibration of the table also have an impact on the turntable, which cannot be ignored. Therefore, mechanical analysis of the table top is necessary. This article uses ANSYS software for mechanical analysis of the turntable. [23, 11, 9]

The load table of the turntable is made of aluminum alloy material, with a density of 2770 kg/m^3 , a tensile yield strength of 280Mpa, a comprehensive yield strength of 280Mpa, and a tensile ultimate strength of 310Mpa. The analysis results are shown in Figures 3.1 to 3.3. Figure 3.1 shows the deformation analysis results of the table after applying load, Figure 3.2 shows the stress analysis results of the turntable, and Figure 3.3 shows the modal analysis results of the turntable.

According to the design requirements, the turntable needs to meet the load requirement of a maximum load of 50 kg. With the turntable as the center, a distributed load is added and a force of 500 N is applied to the turntable surface. Through the analysis of the above simulation results, it can be seen that the maximum deformation of the loading platform is 0.026mm, and the maximum stress is 0.95MPa. Both parameters are



Fig. 3.3: Turntable modal analysis

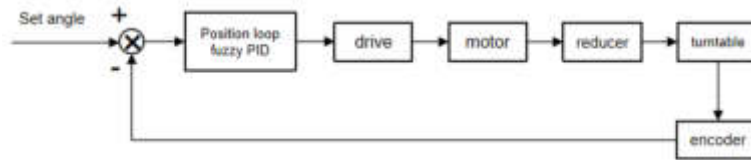


Fig. 4.1: Servo system control block diagram

far less than the strength limit of the material.

4. System hardware composition. The servo control system of the testing turntable mainly consists of an upper computer, a controller, a driver, an encoder, and a motor. The servo system receives control commands from the upper computer through the RS-422 serial port, and reports information such as position, speed, and fault codes. The servo controller is implemented using the Cyclone series FPGA, which is responsible for receiving feedback angle from the encoder. Its NIOS II core completes the calculation of the guidance command, and the calculation results are sent to the driver through the CAN bus. The speed and current loops are calculated in the driver, and finally the driver generates a sinusoidal pulse width modulation (SPWM) signal to drive the motor to move [14]. The system is shown in Figure 4.1.

5. System software design.

5.1. Fuzzy PID control strategy. In this system, the load is large and the inertia is large. Under the action of step response, the deviation is usually not eliminated in a short time, and the integral term can cause significant overshoot, even causing system oscillation and reducing system stability [5, 18, 10]. At the same time, in order to meet the control quantity changes caused by the quality differences of different testing equipment, it is necessary to use control methods with good adaptability. Fuzzy control modifies PID parameters based on fuzzy reasoning, obtaining different correction values based on the size of the deviation, thereby affecting the original PID parameters and accelerating the system response time while ensuring control accuracy. [3, 7] This article adopts a fuzzy PID control strategy to achieve the requirements of precise position control and fast response of the testing turntable. This article designs the algorithm for the position loop of the system. The speed loop and current loop are completed by the selected driver. Effective control of the testing turntable is achieved through position loop control combined with driver parameter settings.

5.2. Design and Simulation of Fuzzy PID Controller. In this turntable control system, the angular position deviation e and the variation of angular position deviation e_c are selected as input variables to complete fuzzy control. Adopting a second-order fuzzy controller, the deviation signal and the variation of the deviation signal are used as the control signals of the entire control system. The ΔK_P , ΔK_I , ΔK_D derived from fuzzy reasoning principles and PID parameters (K_P , K_I , K_D) work together on the controlled object, and θ_{in} represents the target angle issued by the upper computer, θ_{out} represents the current actual angle output by the encoder, the structural flow of the fuzzy PID controller is shown in Figure 5.1.

Establish a fuzzy domain for the control signal, select an appropriate membership function curve for fuzzification processing, and obtain the fuzzy output according to the control rules [16, 2, 17]. Set the basic universe of input variables e , e_c and output variables ΔK_P , ΔK_I , ΔK_D to $[-6, +6]$. Input variables e and e_c correspond to language variables E and E_C , the output variable ΔK_P , ΔK_I , ΔK_D corresponds to the language variable

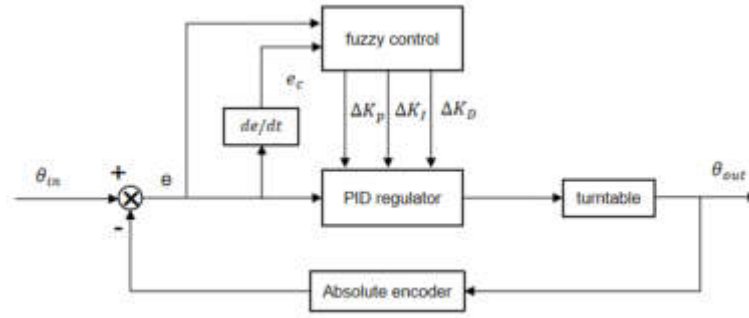


Fig. 5.1: The Structural Process of Fuzzy PID Controller

Table 5.1: Fuzzy rules of ΔK_P , ΔK_I , ΔK_D

$\Delta K_P/\Delta K_I/\Delta K_D$ / E_c	NB	NM	NS	ZO	PS	PM	PB
NB	PB/NB /PS	PB/NB /NS	PM/NM /NB	PM/NM /NB	PS/NS /NB	ZO/ZO /NM	ZO/ZO /PS
NM	PB/NB /PS	PB/NB /NS	PM/NM /NB	PS/NS /NM	PS/NS /NM	ZO/ZO /NS	NS/ZO /ZO
NS	PM/NB /ZO	PM/NM /NS	PM/NS /NM	PS/NS /NM	ZO/ZO /NS	NS/PS /NS	NS/PS /ZO
ZO	PM/NM /ZO	PM/NM /NS	PS/NS /NS	ZO/ZO /NS	NS/PS /NS	NM/PM /NS	NM/PM /ZO
PS	PS/NM /ZO	PS/NS /ZO	ZO/ZO /ZO	NS/PS /ZO	NS/PS /ZO	NM/PM /ZO	NM/PB /ZO
PM	PS/ZO /PB	ZO/ZO /PS	NS/PS /PS	NM/PS /PS	NM/PM /PS	NM/PB /PS	NB/PB /PB
PB	ZO/ZO /PB	ZO/ZO /PM	NM/PS /PM	NM/PM /PM	NM/PM /PS	NB/PB /PS	NB/PB /PB

ΔK_P , ΔK_I , ΔK_D . The fuzzy quantization levels of the input and output language variables E , E_c , ΔK_P , ΔK_I , ΔK_D are [NB (negative large), NM (negative weight), NS (negative small), O (zero), PS (positive small), PM (positive middle), PB (positive large)].

According to the membership function of variable E , E_c , ΔK_P , ΔK_I , ΔK_D , the fuzzy control rules formulated are shown in Table 5.1.

The mathematical model $G(s)$ of the turntable motor is a second-order system, and the relationship between the input and output of the controller is:

$$\omega = K_P e + \int_0^t e(t) dt + K_D \frac{de(t)}{dt}$$

After discrete processing:

$$\omega(k) = K_P e(k) + K_I \sum_{j=0}^k e(j) + K_D [e(k) - e(k-1)]$$

In the equation: $\omega(k)$ and $e(k)$ are the output values and deviation values at the k -th sampling time, using a step signal as the excitation signal of the system. After multiple experiments, the initial PID value of the algorithm is determined.

Obtaining Actual PID Controller Parameters through Demystification:

$$K_P = k_P + \Delta K_P, K_I = k_I + \Delta K_I, K_D = k_D + \Delta K_D,$$

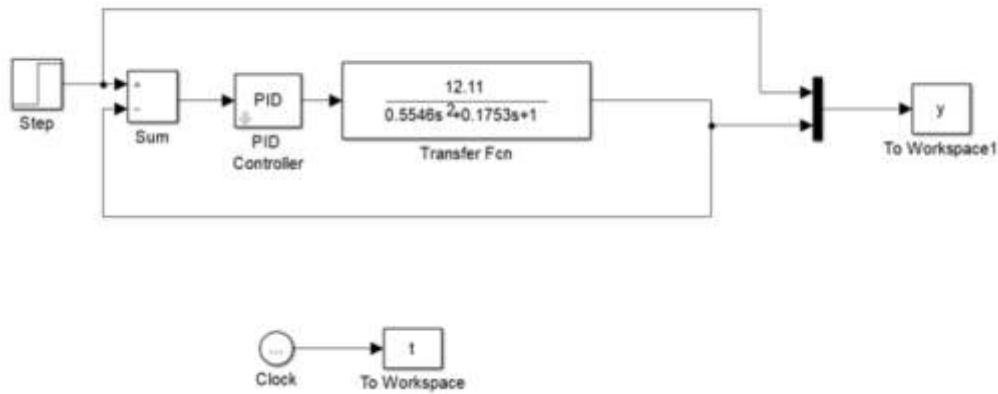


Fig. 5.2: Modeling of Fuzzy PID Control System

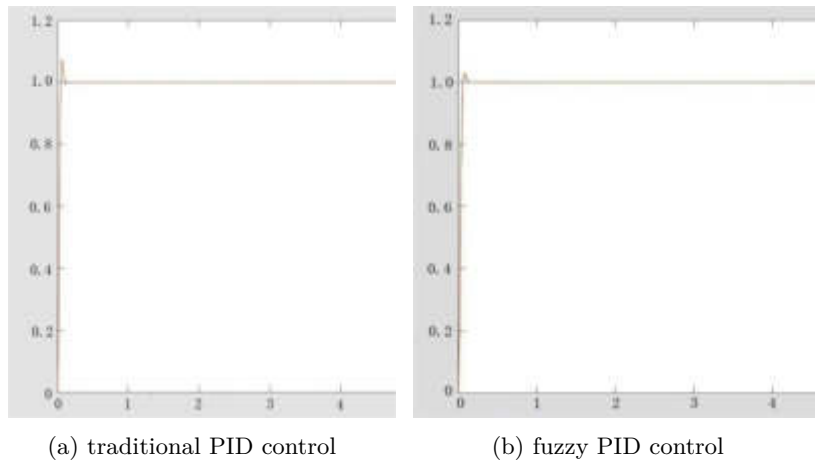


Fig. 5.3: Comparison of Two PID Effects

Among them, k_p , k_I and k_D are conventional PID parameters.

Input fuzzy rules into the fuzzy controller and make relevant settings to establish a simulation model of the turntable control system based on fuzzy PID, as shown in Figure 5.2. In the figure, the sum of the input step signal and feedback signal is inputted into a PID controller, and then output through a transfer function.

Through the SIMULINK modeling and simulation environment of MATLAB simulation software, a comparison was made between fuzzy PID control and traditional PID control. Figure 5.3(a) shows the simulation results of ordinary PID and Figure 5.3(b) shows the simulation results of fuzzy PID. The comparison shows that the entire process output of the fuzzy PID controller is smoother, with less speed oscillation. At the same time, due to strong control adaptability and high control accuracy, it can meet the requirements of motion control for this test turntable.

5.3. Software design. The system software mainly consists of FPGA based controller software for the lower computer. The software design includes functions such as controlling motor operation, adjusting fuzzy PID parameters, and processing relevant data to achieve precise positioning of the turntable. The controller has an RS422 interface for communication with the upper computer. The upper computer sends control instructions based on custom messages through the RS422 serial port, while the lower computer controller constantly reports

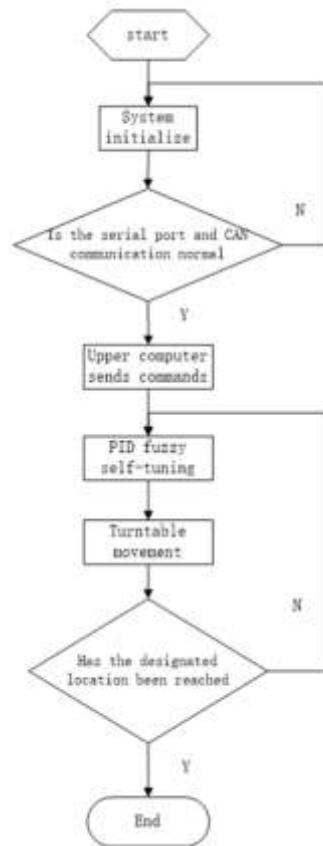


Fig. 5.4: System software structure diagram

information such as system angle, speed, and self status. The system software structure diagram is shown in Figure 5.4.

The system first initializes itself, checks whether the serial port and can communication are normal, and constantly reports the current angle and system status to the upper computer. After normal operation, the controller receives instructions from the upper computer, judges the validity of the instructions, and then completes PID fuzzy self-tuning. The driver drives the turntable to move until the turntable reaches the designated position, ending the control process.

6. Error analysis. In order to test the accuracy of the turntable, according to the *Main Performance Testing Methods for Inertial Technology Testing Equipment*, an angular position measurement test is adopted [19]. The measurement principle is shown in Figure 6.1, and the testing system includes a testing turntable, a 24 sided prism, an autocollimator, and an adjustable tripod. The actual test turntable is shown in Figure 6.2.

The experiment adopts a 24 sided prism. Firstly, from the angle feedback of the measured axis to display the 0 position, record the initial reading of the theodolite θ_1 , Then rotate the measured axis by 15° in sequence based on the angle feedback display value, and record the corresponding readings of the theodolite $\theta_2 \dots \theta_{24}$, at last Calculate the error at each point using the following formula:

$$e_i = \theta_i - \theta_1 (i = 2, \dots, 24), \text{ The test data is shown in Table 6.1.}$$

In Table 6.1, the system measured significant errors at positions 30° , 70° , 225° , and 270° , exceeding $100''$. This is mainly due to structural machining accuracy errors and transmission errors, resulting in significant errors at several fixed positions throughout the system. According to national standards, take the maximum positive error in the calculation results $e^+ = 152''$, the maximum negative error $e^- = -118''$, $\frac{e^+ - e^-}{2} = 2.25'$.

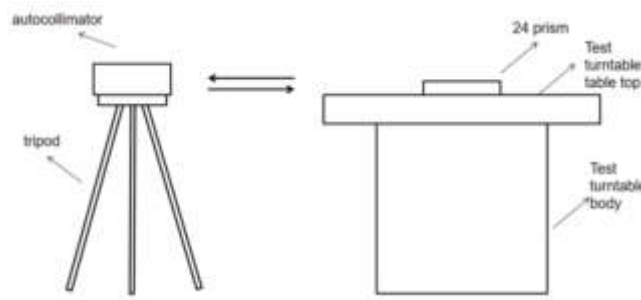


Fig. 6.1: Principle of Position Measurement Test



Fig. 6.2: device detection

Table 6.1: Test turntable error of fuzzy PID algorithm

Set Angle(°)	Measurement value(")	error value(")	Set Angle(°)	Measurement value(")	error value(")
0	228				
15	291	63	360	230	2
30	390	152	345	321	93
45	259	31	330	301	73
60	270	42	315	241	13
75	365	137	300	267	39
90	219	-9	285	163	-65
105	218	-10	270	122	-106
120	263	35	255	209	-19
135	212	-16	240	198	-30
150	199	-29	225	100	-118
165	324	96	210	254	26
180	181	-47	195	240	12

After multiple measurements and taking the average value, it is found that the positioning accuracy of the turntable is better than 2.2'.

During the testing process, the rotational speed of the turntable (1-10°/s) is set, and after considering the total system deceleration ratio, it is converted into the motor speed (rpm). By reading the speed of the motor during operation, the maximum and minimum values of the motor speed during stable operation are recorded. Table 6.2 shows the speed stability of the fuzzy PID algorithm.

The relative error of the turntable measured by this control strategy is 1.9%, which is better than 3.2% of the ordinary PID. Meanwhile, the lower the rotational speed of the turntable, the poorer its speed stability.

Table 6.2: Speed error during fuzzy PID control

Serial Number	Set speed(°/s) (motor speed)	Motor speed range(rpm)	Relative error (%)	Overall relative error(%)
1	1(92.7rpm)	88~97	4.6	1.9
2	2(185.5rpm)	178-191	2.9	
3	4(371.0rpm)	365-380	1.6	
4	6(556.5rpm)	550-562	1.1	
5	8(742.0rpm)	737-746	0.7	
6	10(927.5rpm)	923-931	0.5	

7. Conclusion. This article designs a single axis electric turntable based on the requirements of high positioning accuracy and speed stability of the testing turntable. The system adopts a three-phase synchronous servo motor+reducer+gear transmission method, and the table is made of aluminum alloy material; The control system adopts a fuzzy PID control strategy to achieve the requirements of precise position control and fast response of the testing turntable. The equipment has the characteristics of large load-bearing capacity, high accuracy, and stable speed. Research has shown that:

(1) The turntable adopts gear transmission and multi turn absolute encoder angle measurement, with high transmission torque. The static and modal analysis of the turntable was conducted using ANSYS simulation software, verified that the testing turntable can meet the requirements of high load-bearing, high precision, and smooth operation;

(2) The fuzzy PID algorithm is used to replace the traditional PID algorithm, and an autocollimator is used to measure the positioning error of the turntable. The accuracy of the turntable system can reach 2.2', which is better than 2.6' of ordinary PID, and the running speed is more stable.

In future research and design of the turntable, in terms of machinery, we will consider reducing the clearance between gears from the perspective of improving part machining accuracy and assembly accuracy, or eliminating the clearance through coordinated work of dual motors, in order to achieve the goal of smaller system errors; In terms of control strategy, modern control methods such as sliding mode variable structure control algorithm that have good control effects on nonlinear factors can be used to control the turntable.

Acknowledgement. This work wasFunded by Open Foundation of Industrial Software Engineering Technology Research and Development Center of Jiangsu Education Department. The project number is ZK20-04-03.

REFERENCES

- [1] I. CHAIREZ AND V. UTKIN, *Direct current motor position control by a sliding mode controlled dual three-phase AC-DC power converter*, IFAC-PapersOnLine, 55 (2022), pp. 333–338.
- [2] S. CHEN, C. WANG, Z. ZHANG, X. JI, AND Z. ZHAO, *Improved fuzzy PID method and its application in electro hydraulic servo control*, Mechanical and Electrical Engineering, 38 (2021), pp. 559–565.
- [3] L. CHENG, H. JIANHUI, AND S. JING, *Dual-vector predictive current control of open-end winding pmsm with zero-sequence current hysteresis control*, IEEE Journal of Emerging and Selected Topics in Power Electronics, 10 (2021), pp. 184–195.
- [4] A. N. GUZMÁN, C. C. VACA GARCÍA, S. DI GENNARO, AND C. A. LÚA, *HOSM controller using PI sliding manifold for an integrated active control for wheeled vehicles*, Mathematical Problems in Engineering, 2021 (2021), pp. 1–12.
- [5] Z. HUANG AND Y. FAN, *Research on interference equipment control method based on improved pid algorithm*, Mechanical and Electrical Engineering Technology, 49 (2020), pp. 225–226.
- [6] W. JIANG, W. ZHOU, AND D. LAO, *Design of precision rotary table shafting for multi grating angle measurement system*, Instrument Technology and Sensors, (2018), pp. 24–28.
- [7] A. KISELEV, G. R. CATUOGNO, A. KUZNIETSOV, AND R. LEIDHOLD, *Finite-control-set mpc for open-phase fault-tolerant control of pm synchronous motor drives*, IEEE Transactions on Industrial Electronics, 67 (2020), pp. 4444–4452.
- [8] K. M. LE, H. VAN HOANG, AND J. W. JEON, *An advanced closed-loop control to improve the performance of hybrid stepper motors*, IEEE Transactions on Power Electronics, 32 (2017), pp. 7244–7255.
- [9] D. LI, *Mechanical structure optimization design and system error analysis compensation of high-precision turntable*, 2020.
- [10] L. LI, G. PEI, J. LIU, P. DU, L. PEI, AND C. ZHONG, *2-dof robust h_∞ control for permanent magnet synchronous motor with disturbance observer*, IEEE Transactions on Power Electronics, 36 (2021), pp. 3462–3472.

- [11] Z. LI, J. LI, B. HAN, Y. TANG, AND E.-K. YEOH, *Research on the design of high-precision angular indexing turntable and its error correction*, *Electromechanical Engineering*, 38 (2021), pp. 1180–1184.
- [12] A. PILLONI, M. FRANCESCHELLI, A. PISANO, AND E. USAI, *On the variable structure control approach with sliding modes to robust finite-time consensus problems: A methodological overview based on nonsmooth analysis*, *Annual Reviews in Control*, (2023).
- [13] M. SKOWRON, T. ORLOWSKA-KOWALSKA, AND C. T. KOWALSKI, *Detection of permanent magnet damage of PMSM drive based on direct analysis of the stator phase currents using convolutional neural network*, *IEEE Transactions on Industrial Electronics*, 69 (2022), pp. 13665–13675.
- [14] P. SUN, *Control system design of single axis high-precision turntable*, 2019.
- [15] H. N. TRAN, K. M. LE, AND J. W. JEON, *Adaptive current controller based on neural network and double phase compensator for a stepper motor*, *IEEE Transactions on Power Electronics*, 34 (2018), pp. 8092–8103.
- [16] Y. WANG, *Control system design and control algorithm research of single axis high-precision turntable*, 2020.
- [17] B. WEI, F. TANG, C. LIANG, AND A. ZHANG, *Research on flight turntable control system based on variable universe fuzzy PID*, *Journal of Beijing University of Chemical Technology (Natural Science Edition)*, 49 (2022), pp. 107–115.
- [18] A. T. WOLDEGIORGIS, X. GE, H. WANG, AND M. HASSAN, *A new frequency adaptive second-order disturbance observer for sensorless vector control of interior permanent magnet synchronous motor*, *IEEE Transactions on Industrial Electronics*, 68 (2020), pp. 11847–11857.
- [19] G. YU, W. ZENG, H. CHEN, Q. CHEN, AND X. HE, *Structural design and accuracy analysis of large precision turntables*, *Manufacturing Automation*, 41 (2019), pp. 104–107+142.
- [20] Y. ZHAI, Y. GUO, AND L. ZHU, *Simulation study on fuzzy PID closed loop control system of stepping motor*, *Modern Electronic Technology*, 38 (2015), pp. 146–149.
- [21] X. K. ZHANG, J.-Y. GAUTHIER, AND X. LIN-SHI, *Cost-efficient fault-tolerant scheme for three-phase surface-mounted permanent magnet synchronous machines fed by multifunctional converter system under open-phase faults*, *IEEE Transactions on Industrial Electronics*, 69 (2022), pp. 5502–5513.
- [22] C. ZHAO, Z. LIN, J. LIU, AND L. WAN, *Research on control method of hybrid two-phase stepping motor based on adaptive fuzzy*, in *2017 IEEE 3rd Information Technology and Mechatronics Engineering Conference (ITOEC)*, IEEE, 2017, pp. 649–653.
- [23] C. ZHAO, J. MA, X. FAN, AND R. JI, *Design of MRAC and modified mrac for the turntable*, in *2020 39th Chinese Control Conference (CCC)*, IEEE, 2020, pp. 1874–1878.

Edited by: Jingsha He

Special issue on: Efficient Scalable Computing based on IoT and Cloud Computing

Received: Dec 20, 2023

Accepted: Mar 1, 2024



A VISUAL WEBPAGE INFORMATION EXTRACTION FRAMEWORK FOR COMPETITIVE INTELLIGENCE SYSTEM

ZHIWEI ZHANG*, WENBO QIN[†] AND HAIFENG XU[‡]

Abstract. The extraction of webpage information is of paramount importance in the realm of competitive intelligence. This research is dedicated to the design and implementation of a visual webpage information extraction module within a competitive intelligence system, approached through the lens of research and development (R&D) technology and its practical applications. Initially, the study delineates the objectives and requirements for webpage information extraction, emphasizing the practical needs of competitive intelligence systems. By critically assessing the strengths and weaknesses of current theories and methodologies in webpage text information extraction, this paper introduces an innovative visual method for extracting webpage text information. Subsequently, the paper meticulously outlines the comprehensive architecture of the proposed module. Building upon this foundation, the study delves into the specifics of the extraction template, rule generation, optimization techniques, and the extraction algorithm pivotal to the process of visual webpage information extraction. The system's effectiveness and practical utility are substantiated through a series of confirmatory experiments, the results of which are thoroughly analyzed. The findings affirm that the developed system adeptly fulfills the webpage information extraction needs of competitive intelligence systems, contributing significantly to the R&D efforts in which the authors are engaged.

Key words: Information extraction visualization, natural language processing, competitive intelligence, data mining

1. Introduction. With the exponential growth of the Internet and significant advancements in information technology, the digital landscape has emerged as the predominant arena for corporate and institutional information dissemination. An array of content, including corporate profiles, product details, promotional events, recruitment opportunities, and technological advancements, is extensively published online [25, 26, 14]. Recent studies underscore that a vast majority (approximately 90%) of the data requisites for competitive intelligence analyses are sourced from the Internet, underscoring the critical role of web-based information in today's business milieu [33, 19, 28]. Competitive intelligence encompasses a comprehensive system of information gathering, analysis, and dissemination, designed to equip businesses with profound insights into competitors, market dynamics, and industry evolutions. Within this framework, webpage information extraction assumes a pivotal position, empowering entities to not only distill invaluable insights from the deluge of data but also to leverage such insights for tangible competitive edges. Particularly in the big data era, navigating the complexities of data processing and analysis presents formidable challenges [19, 18, 24].

Webpage information extraction refers to the process of automatically retrieving structured or semi-structured information from unstructured webpage content. This process involves identifying relevant pieces of data within a webpage, such as text, images, links, and other multimedia elements, and then transforming this data into a more organized format that is suitable for analysis, storage, and further processing [1, 13]. The goal of webpage information extraction is to enable computers to understand and utilize the vast amount of information available on the Internet efficiently. This technology underpins various applications, including search engines, competitive intelligence systems, market research, and content aggregation services, facilitating the automatic collection and analysis of web data [27, 17].

In the realm of competitive intelligence systems, the task of webpage information extraction encounters significant hurdles due to the sheer diversity and complexity of webpage contents. The digital landscape is a mosaic of information presented in myriad forms ranging from text, images, videos, tables, to interactive

*School of Informatics and Engineering, Suzhou University, Suzhou, China. Corresponding author: Zhiwei Zhang (zzwloveai@gmail.com).

[†]School of Informatics and Engineering, Suzhou University, Suzhou, China.

[‡]School of Informatics and Engineering, Suzhou University, Suzhou, China.

elements, each differing vastly in format and structure across websites. This variability presents a formidable challenge in crafting a one-size-fits-all, efficient information extraction system [8, 6]. For example, product descriptions might be straightforward text or be supplemented with images and videos, while technical articles could include intricate diagrams or snippets of code. Moreover, the inherent complexity within HTML structures, coupled with the ever-evolving designs of webpages, complicates the accurate retrieval of information, necessitating highly flexible and adaptive strategies for extraction [21, 12]. Ensuring the comprehensiveness and precision of extracted information thus requires leveraging sophisticated text processing and image recognition technologies [9, 23].

Furthermore, competitive intelligence systems grapple with the challenges posed by the variability of noise data, webpage formats, and structures. The internet is inundated with noise advertisements, promotional links, and irrelevant comments clutter webpages, obscuring valuable information. Effectively sifting through and excluding such noise is a crucial task in the extraction process [30, 5]. This necessitates not only the development of sophisticated algorithms capable of discerning between relevant and irrelevant content but also their ongoing refinement to keep pace with the dynamic nature of the web. The diversity in webpage design and layout demands that extraction systems be versatile enough to navigate a variety of HTML/CSS structures [20, 15]. Additionally, the rapid evolution of web technologies and the introduction of new standards for webpage design and layout further amplify the complexity, challenging information extraction systems to continuously adapt to these new formats and structures [2, 31].

The challenges of real-time updates and information overload significantly complicate the task of webpage information extraction. The dynamic nature of the Internet, with its continuous influx of new content and the potential modification or removal of existing information, presents a critical challenge for competitive intelligence systems. Timely tracking and management of these changes are crucial, as reliance on outdated or inaccurate information could result in erroneous analyses and decision-making [11, 29]. Furthermore, the exponential growth in online content has precipitated an era of information overload, posing a substantial challenge. Competitive intelligence systems are thus tasked with the efficient processing and filtering of this vast amount of data to ensure the extraction and analysis of only the most relevant and valuable information. Consequently, these systems must possess not only robust data processing capabilities but also sophisticated data screening and prioritization mechanisms. Such features are essential to navigate through the vast data landscape effectively, preventing the degradation of analysis efficiency or decision-making errors that could arise from information overload [32, 3].

Within the domain of competitive intelligence systems, visual webpage information extraction technology presents clear advantages over traditional text extraction methodologies [4, 7]. Primarily, the visual approach enhances the efficiency and accuracy of information extraction by offering an intuitive display of data structures and content. Traditional methodologies, which predominantly rely on semantic analysis and keyword matching, frequently fall short in addressing webpages characterized by intricate structures or varied formats. Conversely, visual extraction technology capitalizes on the visual layout and structural features of webpages to more precisely locate and extract pivotal information. For instance, by scrutinizing the Document Object Model (DOM) structure and visual indicators on webpages, elements such as article titles, texts, images, and tables can be more effectively identified. Furthermore, visual information extraction methods substantially bolster data understanding and analysis. This approach simplifies the process for users to comprehend the overarching structure and key components of the data. The intuitive nature of this representation not only facilitates a swift comprehension of the information's essence but also aids analysts in uncovering potential data interconnections. Crucially, the visualization technique assumes added significance in the context of big data processing [16]. With data volumes expanding continuously, traditional text extraction and analysis techniques are increasingly overwhelmed. Visualization technology, through its capacity to efficiently manage and present large datasets via summary views and interactive exploration features, empowers users to grasp a higher-level understanding of the data. This, in turn, allows for the rapid identification of areas of interest, followed by detailed analysis [10, 22].

In summary, the realm of webpage information extraction technologies currently grapples with significant challenges, chiefly arising from the diversity and complexity of web content, the proliferation of noise and extraneous information, and the dynamic evolution of webpage formats and structures. These obstacles underscore the critical necessity for extraction systems that are both highly adaptable and technologically advanced,

equipped to identify and process a wide array of information types. The progression of competitive intelligence systems is contingent upon the development of more refined algorithms dedicated to noise filtration and the incorporation of cutting-edge techniques in text and image processing. Such enhancements are essential to augment the precision and depth of information extraction. Tackling these prevalent issues is crucial for the enhancement of competitive intelligence systems, enabling them to effectively and accurately harness the extensive reservoir of data available online.

Hence, the efficacy of competitive intelligence systems hinges critically on their ability to precisely distill targeted text information from the disarray of webpage source codes and to subsequently generate standardized structured documents. Such a foundational step is indispensable for the successful execution of text classification and text mining processes. In this research, we introduced a comprehensive visual webpage information extraction framework tailored to meet the specific text mining needs of competitive intelligence systems. The primary endeavors and contributions of this study are summarized as follows:

(1) This research meticulously articulated the design objectives and requirements for webpage information extraction, aiming to develop a framework that stands out for its efficiency, accuracy, and user-friendliness. Unlike existing methodologies that often struggle with the dynamic and complex nature of web data, this framework is specifically engineered to adeptly handle such challenges. It focuses on extracting crucial information by ensuring data integrity and accuracy, enhancing processing speed, and expanding system scalability, all the while improving the ease of user interaction. This approach addresses a significant gap in current practices, where the balance between comprehensive data extraction and user-centric design often remains unachieved.

(2) We introduced a holistic visual webpage information extraction framework, distinct from current solutions by its integration of cutting-edge technologies across data crawling, natural language processing, and visualization. This framework is designed to facilitate automated data extraction from a diverse array of webpages with minimal human intervention. The novel incorporation of visual elements specifically aims to demystify the user interaction process and elevate the intuitiveness of data presentation. This strategy marks a departure from traditional methods that may not fully leverage visual cues for user engagement and data interpretation, highlighting the innovative edge of our approach in enhancing both system usability and information clarity.

(3) The development and validation of a comprehensive visual webpage information extraction process underscored its applicability and effectiveness across various webpage types, both static and dynamic. The empirical evidence gathered from these experiments showcases our method's superiority in not only improving the precision of information extraction but also in significantly lightening the user's workload. This contrasts sharply with many existing techniques that may exhibit limitations in versatility across different webpage formats or impose a heavier analytical burden on users. Furthermore, the identification of potential areas for future enhancements opens new avenues for advancing the state-of-the-art in webpage information extraction. These findings offer a critical reflection on the gaps within current methodologies and provide a clear direction for subsequent research efforts, aiming to refine and augment the capabilities of competitive intelligence systems in navigating the vast and varied terrain of web information.

The remainder of this study was organized as follows: In Section 2, the design objectives and requirements of visual webpage information extraction were introduced; in Section 3, the overall framework for the visual webpage information extraction system was generalized, and three subsystems—webpage information crawling, visual webpage information extraction rule template generation, and webpage text information extraction—were elaborated on; in Section 4, the relevant experimental environment and experimental data were summarized, and the experimental results were analyzed; in the final section, conclusions were drawn and the future research directions were put forward.

2. Design objectives and requirements of visual webpage information extraction. In the context of the contemporary era, characterized by an explosion of network information, the utilization of web crawlers for webpage crawling on the Internet has become a pivotal means of information collection. However, the challenge lies in the complexity of the original webpage content obtained by these crawlers, which is often encumbered with a plethora of HTML tags. Hidden within these tags is a vast amount of text information, making the extraction process intricate and demanding.

The primary design objective of visual webpage information extraction is to navigate through this complex

maze of webpage source codes and distill high-quality text information that aligns with user needs. This task involves meticulously parsing the cluttered webpage data, identifying and extracting relevant information, and transforming it into standardized data formats such as titles, texts, and other pertinent entries. Achieving this requires a sophisticated system capable of discerning and isolating useful content from the plethora of unstructured data typically found in webpages.

From a practical application standpoint, this objective is not only feasible but increasingly necessary. The vast and growing volume of web-based information necessitates efficient and accurate extraction methods to harness this data for meaningful use. The proposed system addresses this need by employing advanced extraction techniques, focusing on the critical aspects of accuracy, reliability, and speed. In terms of system implementation, however, the feasibility is grounded in current technological advancements in web crawling, HTML parsing, and data extraction algorithms. The system's design will leverage state-of-the-art techniques in these areas, ensuring that it can effectively handle the diverse and dynamic nature of web content. This includes the capability to adapt to various webpage structures, handle different types of content (including multimedia elements), and process information rapidly and accurately.

In summary, the design objective of extracting high-quality, user-centric text information from webpages is both practical and attainable. It resonates with the current demands of information extraction in the digital age and is supported by feasible technological solutions. The implementation of such a system promises significant benefits in terms of enhancing the efficiency and effectiveness of web-based information collection and analysis. Based on the analysis aforementioned above, the design and requirements of the webpage information extraction system must focus on the following key aspects:

(1) Accuracy and completeness: The accuracy of extraction is the primary system design objective, meaning that the system should focus on extracting specific target items (such as titles, texts and abstracts) while excluding all unspecified noise information. In addition to accuracy, completeness is also of crucial importance, ensuring that the extracted target items retain complete contextual semantics so as to provide support for subsequent operations on standardized documents.

(2) Timeliness and efficiency: Considering the huge amount of webpage resources on the Internet, the webpage information extraction system needs to have efficient processing ability. Web crawlers, which usually run in a distributed and multithreaded way, can grab a large number of original webpages in a short time. Therefore, the system should be able to extract specific target items accurately and completely from the original webpages in time and efficiently and quickly form standardized documents.

(3) Adaptability: The rapid development of Web technology means that the structure of webpages is frequently updated. In this context, the web information extraction system needs to be self-adaptable to some extent. When the existing information extraction templates and rules fail to correctly extract webpage information, for instance, the system should be able to give an alarm and use the wrong feedback information to adjust the extraction rules and templates in time to adapt to the latest webpage structure.

(4) Practicality: While meeting the professional requirements, the webpage information extraction system should also consider the use needs of non-professional users. This means that the user interface of the system should be intuitive and easy to use, while ensuring the powerful and stable functions of the system to adapt to the operating habits and skill levels of different users.

To sum up, in order to effectively support the collection and analysis of competitive intelligence, the web information extraction system must meet high standards in accuracy, timeliness, adaptability, and practicality.

3. Overall framework for the visual webpage information extraction system. The visual webpage information extraction system is comprised of three integral subsystems: (1) a webpage information crawling subsystem, which retrieves data from the internet; (2) a visual webpage information extraction rule template generation subsystem, tasked with creating rules for data identification and extraction; and (3) a webpage text information extraction subsystem, dedicated to isolating textual content from webpages. The comprehensive structure of this system is illustrated in Figure 3.1, providing a cohesive overview of its components and operational flow.

From the overall framework for the visual webpage information extraction system as shown in Figure 3.1, the system is generally divided into three subsystems whose composition, functions, and techniques used are briefed as follows:

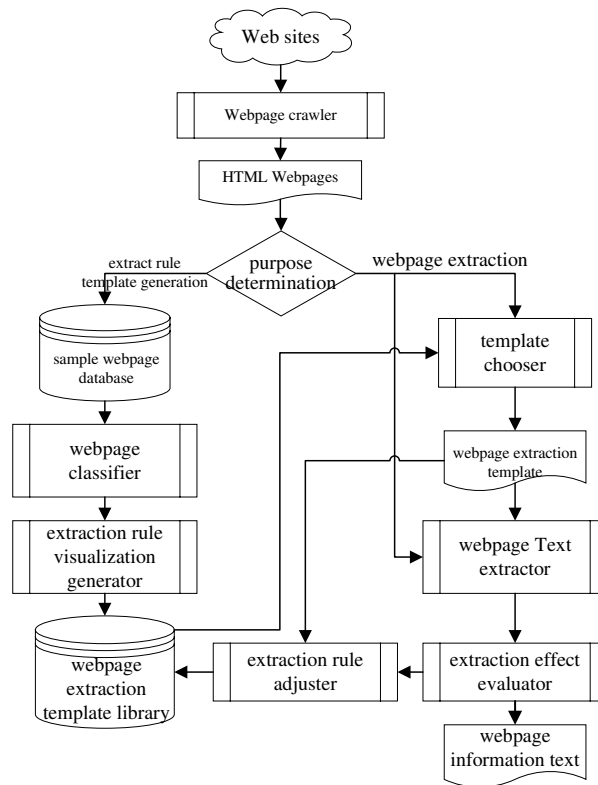


Fig. 3.1: Overall framework structure for the visual webpage information extraction system.

(1) Webpage Information Crawling Subsystem: This subsystem employs a webpage crawler to gather webpage information (source code) from designated sites based on predetermined seed site URLs and crawling strategies. The captured data is stored in the sample webpage library" depicted in Figure 3.1 for future use or as input for the webpage text information extraction subsystem" for text data extraction. Developed on the open-source Nutch search engine, this subsystem has been customized to fulfill the unique requirements of our research.

(2) Visual Webpage Information Extraction Rule Template Generation Subsystem: As illustrated in Figure 3.1, this subsystem integrates the sample webpage library," sample webpage classifier," visual webpage extraction template generator," and webpage extraction template library." Representing the innovative core of this study, it generates extraction rule templates for specific webpages using samples from different sections of the target site, storing these templates in the webpage extraction template library." It primarily utilizes the open-source htmlparser, which has been adapted and refined for this research.

(3) Webpage Text Information Extraction Subsystem: Comprising the webpage extraction template selector," webpage information extractor," extraction effect evaluator," and webpage template adjuster" as shown in Figure 3.1, this subsystem extracts information from the original webpages collected by the webpage information crawling subsystem." It utilizes site-specific extraction rule templates developed by the visual webpage information extraction rule template generation subsystem" to output standardized webpage text information. Utilizing the open-source htmlparser for formatting and extracting text from specific webpage tags, this component has been enhanced for research purposes.

The aforesaid subsystems will be introduced one by one from the angles of design and implementation, and the algorithm used by each subsystem will be introduced and analyzed in detail.

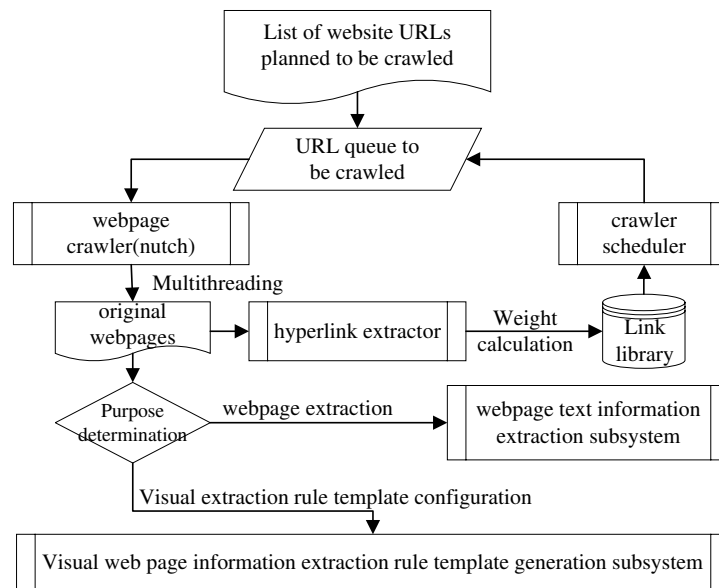


Fig. 3.2: Framework for the webpage crawling subsystem.

3.1. Webpage information crawling subsystem. This subsystem focuses on retrieving webpages from the Internet, archiving the initially crawled content in a sample webpage library for subsequent use or directly forwarding it to the webpage text information extraction subsystem for data extraction. It has been developed and enhanced using the open-source Nutch search engine, tailored specifically to meet the research requirements of this study. The comprehensive structure of the webpage crawling subsystem is depicted in Figure 3.2.

The workflow of the improved webpage crawling subsystem is as follows:

(1) An initial set of URLs for crawling sites is established, comprising a document that lists URLs from various seed sites.

(2) These seed site URLs are then populated into CrawlDB, a database dedicated to storing the URLs and their status information for webpages.

(3) Utilizing the data within CrawlDB, a comprehensive list of URLs for site crawling is compiled, based on both the URL and status information.

(4) Employing the Nutch multi-threaded crawler, the system proceeds to crawl webpages as delineated by the URL list of crawling sites, generating relevant webpage snapshots from the content retrieved during the crawl and logging the crawl process. Concurrently, hyperlinks contained within the webpages are analyzed, facilitating the output of the original webpage stream.

The webpage hyperlink information parsed in Step (4) is used to update CrawlDB; Steps (3) to (4) are repeated until reaching the preset crawling depth or the crawling task is manually stopped, so as to form a cyclic process of generation-crawling-updating.

3.2. Visual webpage information extraction rule template generation subsystem. In the visual operation environment, the objective is to achieve precise and clear generation of information extraction rule templates, thereby obviating the need for complex algorithm-derived parsing rule templates. This section delves into and implements the visual webpage information extraction rule template generation process and its specifics. Within the browser/server (B/S) visual environment, the "target extraction item region" is selected via mouse, facilitating the automatic generation and extraction of rules for the "target extraction items." Concurrently, specific extraction rules are compared and calibrated within this environment. The detailed implementation steps are outlined as follows:

(1) All the sample webpages belonging to the site S are read from the sample webpage library, and displayed after sorting according to their webpage URL, so as to facilitate the subsequent URL template configuration for each module of the site S through the observation method.

(2) With site S 's samples displayed post-sorting, webpages from different modules naturally cluster together. A URL template for module T of site S is set up through observation differentiating sample webpages by URL strings with identical prefixes and manually crafting matching URL regular expressions.

(3) A unique extraction template is created for each module. If module T on site S already has a corresponding extraction template, this step is bypassed; otherwise, a unique extraction template P is generated.

(4) Specific extraction rules for target extraction items are established for all webpages within module T of site S . It's important to recognize that all webpages in module T are issued through the same information release template backstage, making their structures identical (or similar) aside from text differences. In this research, three sample webpages from module T were chosen at random to visually create extraction rules for target extraction items.

3.3. Webpage text information extraction subsystem. The webpage information extraction subsystem constitutes a crucial component of the comprehensive visual webpage information extraction system. It employs predefined extraction rule templates to retrieve pertinent data from "target extraction items" on webpages, aggregating these items into a standardized document output. This subsystem is comprised of the extraction template selector, webpage information extractor, extraction effect evaluator, and webpage extraction rule template adjuster, as depicted in Figure 3.3.

3.3.1. Relevant data structure. The data structures utilized for extracting information from webpages via extraction rule templates are outlined as follows:

(1) Webpage parsing result object *tagNodeList*: The NodeList-type object obtained by parsing the webpage source code with `htmlparser` is a tree-structured type of data and an operation object generated by extracting target items with the webpage extraction rules.

(2) Extraction rule *ParseRule*: The position of the target extraction item region in the webpage is marked, that is, the absolute path (webpage tag sequence) from the DOM tree root node of the webpage to the target extraction item region, aiming to guide the webpage information extraction subsystem to extract the specific target extraction items.

(3) Extraction template *ParseTemplate*: It is a set of the extraction rules for each target extraction item, and the concrete templates for the specific sites are encapsulated.

(4) Extraction rule list *ruleList*: It is a data structure of `List<ParseRule>` type, which stores the extraction rules for all target extraction items under the same extraction template.

(5) Extraction rule list *itemRuleList* of target extraction items: It is also a data structure of `List<ParseRule>` type, which encapsulates multiple extraction rules for a specific target extraction item.

(6) Extraction rule tag list *ruleTagList*: It is the object of `List<String>` type. The extraction rule is a tag string sequence composed of all tags on the path from the tag of the root node of the webpage to the tag of the parent node of the target extraction item, which is a whole character string, and the *ruleTagList* stores a list of character strings with a single tag as a character string, that is, a list composed of single tags.

(7) Matching tag sequence stack *pathStack* and node child queue *childQueue*: The *pathStack* records already matched path tags, being the objects of `Stack<TagNode>` type, and *childQueue* encapsulates all child nodes of one node.

(8) Standard document *StructuredDoc*: It is a data structure of `Map<TargetItem, Text>` type, which encapsulates the text information of target extraction items, including the titles, abstracts, keywords, and texts of webpages.

3.3.2. System workflow. Figure 3.3 illustrates the comprehensive workflow of the "webpage text information extraction subsystem," detailed as follows:

(1) The web crawler of the webpage crawling subsystem designed in Section 3.1 is enabled to configure the information of the crawling site and the related attributes of the web crawler to the site S (*siteID*) in a multi-threading way for webpage crawling. Then, the URL *pageURL* of webpages is extracted and the webpages are returned in the form of source codes to generate the character string *pageSource* output of webpage source

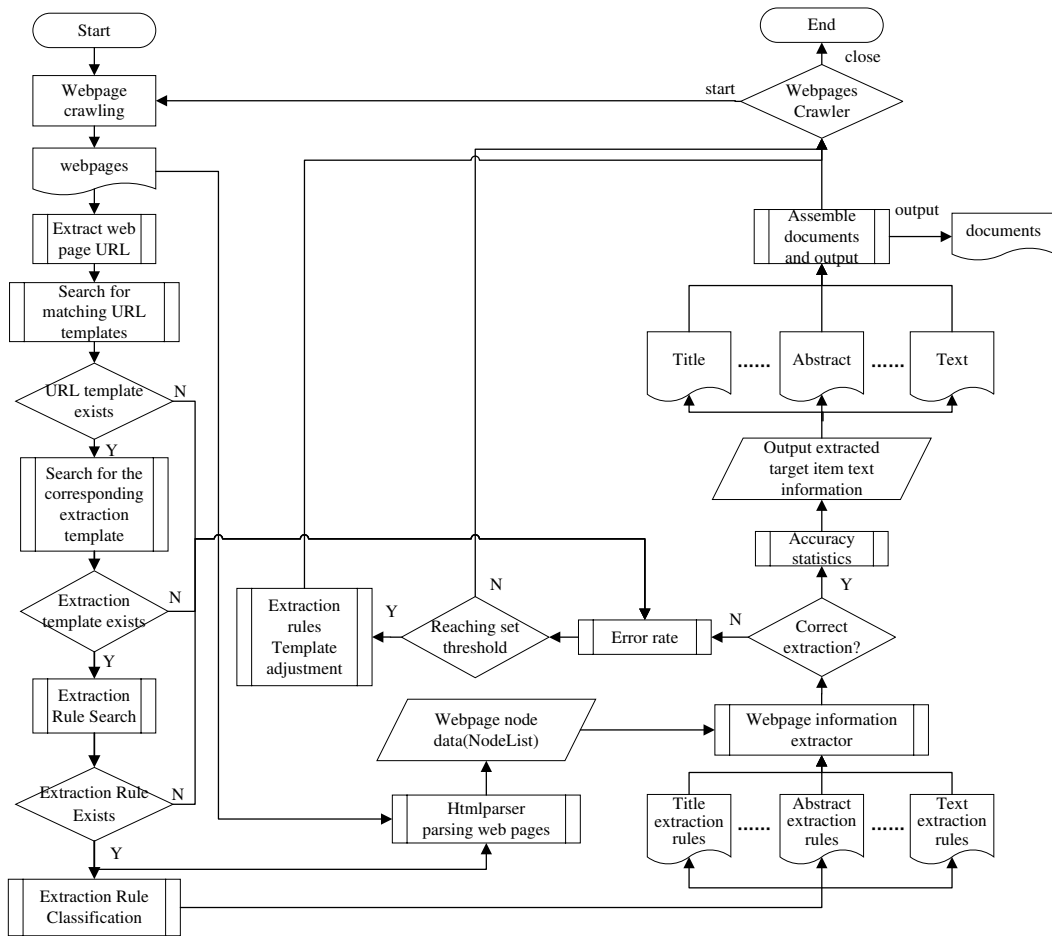


Fig. 3.3: The workflow of the webpage text information extraction subsystem. Initially, the subsystem commences with the advanced data crawling phase, employing state-of-the-art technology to navigate and retrieve content from a diverse array of web sources. This phase is critical for ensuring that the most current and relevant webpage data is captured for analysis, addressing the challenge of the internet’s ever-evolving content landscape. Then the subsystem applies natural language processing (NLP) tools to the raw webpage content. This involves sophisticated semantic analysis to identify and isolate valuable text information from the surrounding web elements and noise. The NLP phase is crucial for refining the data into a format that is both meaningful and actionable, setting our approach apart from conventional methods that may struggle with the complexity of web content structures. The final stage of the workflow integrates visualization technology to present the extracted information in an intuitive and user-friendly manner. This not only simplifies the interaction process for users but also enhances their ability to comprehend and analyze the data.

codes.

(2) *SiteID* in Step (1) is used to acquire all URL template *urlTemplateList* belonging to this site from the database. If *urlTemplateList* is empty, skip to Step 3.3.2, or otherwise, *pageURL* in Step (1) is matched with the URL regular expression in *urlTemplateList* one by one. If matching fails all the time, skip to Step (11). If matching succeeds for a single template, the *urlTemplateId* of this URL template is recorded.

(3) The extraction template ID *parseTemplateId* belonging to this URL template is acquired from the database. If it is empty, skip to Step (11), or otherwise, the next step will be implemented.

(4) All extraction rules belonging to this extraction template are searched in the database via *parseTemplateId* in Step (3), and a value is assigned to *ruleList*; if *ruleList* is empty, skip to Step (11), or otherwise, the next step will be implemented.

(5) The extraction rules in Step (4) are classified according to the target extraction items to form the extraction rule *itemRuleList* of each target extraction item *PT*.

(6) The *pageSource* in Step (1) is parsed using *htmlparser* to generate *tagNodeList*. Since the operation on *tagNodeList* is destructive, *pageSource* is parsed again using *htmlparser* to generate new *tagNodeList* before extracting the text information of different target extraction items.

(7) Each extraction rule object in *itemRuleList* in Step (5) is cyclically traversed. Each extraction rule object is converted into *ruleTagList*, and a pointer *i* pointing to *ruleTagList* is set and initialized as 0, i.e., pointing to the first element of *ruleTagList*.

(8) The node information on the first tag node *curNode* of *tagNodeList* is matched with *ruleTagList[i]* using *tagNodeList* in 3.3.2 and *ruleTagList* in Step (7). If matching fails, turn to Step (10); if matching succeeds, *curNode* joins in *pathStack*, and all child nodes of *curNode* enqueues in *childQueue*; next, an element is dequeued from *childQueue*, a value is assigned to *curNode*, and $++i$, meaning that the pointer pointing to *ruleTagList* shifts forward by one position.

(9) The node information on *curNode* is matched with *ruleTagList[i]*: 1) if matching succeeds, *curNode* joins in *pathStack*, *childQueue* is emptied, and all child nodes of *curNode* re-enqueue in *childQueue*; 2) if matching fails, a node element is unqueued from *childQueue* to assign a value to *curNode*, which matches with *ruleTagList[i]* once again, and if this succeeds, Step 1) is repeated; if the matching fails, Step 2) is repeated; 3) if all child nodes fail in matching, a node element is shifted out of *pathStack*, and a value is assigned to *curNode*, namely, $-i$, *childQueue* is emptied, all matched child nodes of *curNode* are enqueued in *childQueue*, and Step (9) is repeated; 4) if all *ruleTagList[i]* are matched, the matching succeeds, all text information *targetString* under *curNode* is taken out to form $\langle PT, targetString \rangle$ pairs, which are stored in *structureedDoc*, and then turn to Step (5), the information of the next target extraction item is extracted. If the information of all target extraction items is extracted, turn to Step (12); 5) if a *ruleTagList[i]* fails in matching, or the child nodes of one *curNode* are empty and *ruleTagList* is not completely traversed, turn to Step (10); If *pathStack* is empty or the tag path composed of the nodes therein is different from *ruleTagList*, turn to Step (10).

(10) If extraction (matching) is incorrect, one is added to the extraction error counter to calculate the extraction error rate *W*. If *W* is greater than or equal to the preset threshold *TH*, turn to Step (13), or otherwise, turn to Step (1).

(11) If the working state of the web crawler is off, exit from the whole webpage information extraction system; if the working state of the web crawler is on, turn to Step (1), followed by the next round of webpage crawling.

(12) Turn to Step (2) for information extraction for the webpages crawled in this round but not subjected to information extraction (the web crawler crawls multiple webpages in each ground under the multi-threading mode); if all webpages are extracted, turn to Step (1).

(13) The specific extraction rules are regenerated according to the generation and calibration method for the extraction rule template of target extraction items.

From the perspectives of system implementation and application, the detailed process of visual webpage text information extraction is depicted through the following pseudo-code, as presented in Algorithm 1.

4. Experimental results and analysis.

4.1. Experimental environment. To assess the performance and efficiency of the “visual webpage information crawling” system within a specific hardware and software environment, a detailed experimental scheme was devised to evaluate the system’s real-world operational efficacy. The hardware setup for the experiment comprised a computer equipped with an AMD Ryzen 5 PRO 2500U processor, featuring a Radeon Vega Mobile GFX integrated graphics card and a base clock rate of 2.00 GHz, alongside 8GB of RAM. On the software front, the experiment utilized Apache Nutch, a highly extendable, open-source webpage crawler software designed for harvesting original webpage data from the Internet. This combination of tools enabled the effective processing of Chinese texts and facilitated the extraction of information from a broad array of online resources.

Algorithm 17 *parseToFormStrDoc*(String pageURL, int siteId, String pageSource)**Input:** Web Page URL pageURL, website ID siteId, webpage source pageSource**Output:** structured document structuredDoc

```

1: Map<Integer, StringBuffer>structuredDoc; // the extracted structured documents.
2: // obtaining a url template matching the web page URL based on the webpage
3: // URL and its associated site id
4: URLTemplate urlTemplate = getMatchURLTemplate(pageURL, siteId);
5: // acquiring the webpage parsing template based on the URL template
6: ParseTemplate parseTemplate = getParseTplmtByURLTplmt(urlTemplate);
7: //obtaining the webpage extraction template and all extraction rules
8: List<ParseRule>ruleList = getParseRulesByParseTemplate(parseTemplate);
9: // categorizing by 'extraction target items' to form sets of extraction rules for each
10: // respective extraction target item
11: for parseRule in ruleList do
12:     List<ParseRule>itemRuleList; // Extraction rule list.
13:     utilizing the extraction rule set 'itemrulelist' to extract text information corresponding to the 'extraction target
        item';
14:     for itemRule in itemRuleList do
15:         NodeList tagNameList = parseHtmlTagName(pageSource); // HTML tags node list.
16:         StringBuffer targetString;
17:         // separating 'itemRule' (Label Path) and storing it as an array of strings
18:         itemParseRule = splitToStrList(itemRule); // HTML text parse rules.
19:         int ruleListLength = itemParseRule.size();
20:         int ruleListIndex = 0;
21:         // utilizing parsing rules for the actual analysis of web pages
22:         NodeList childList; // The child node list of a specified node in a webpage.
23:         if (matching the root element of tagNameList with itemParseRule[ruleListIndex]) then
24:             childList = rootTagName.getChildren();
25:             ++ruleListIndex;
26:         else
27:             Terminating the Information Extraction for the Specified 'Extraction Target Item';
28:         end if
29:     end for
30:     while childList not empty do
31:         int childIndex = 0;
32:         for childIndex < childList.size() do
33:             Node tmpNode = childList.elementAt(childIndex);
34:             if tmpNode instanceof TagNode then
35:                 TagNode curTagNode = (TagNode) tmpNode;
36:                 tag1 = curTagNode.getText(); // Extract the text information from a webpage node.
37:                 tag2 = ruleList.get(ruleListIndex);
38:                 if tag1 == tag2 then
39:                     if ruleListIndex != (ruleListLength - 1) then
40:                         childList = curTagNode.getChildren(); // Get child nodes of the current node.
41:                         ++ruleListIndex;
42:                     break;
43:                     else if complete matching of parsing rules then
44:                         NodeList curChilds = curTagNode.getChildren();
45:                         for (int j = 0; j < curChilds.size(); ++j) do
46:                             targetString += Text Information of Each 'Child Tag';
47:                         end for
48:                     end if
49:                     structuredDoc.put(extraction of target item number, extraction of target item text information);
50:                 end if
51:             end if
52:         end for
53:     end while
54:     if (several nodes in the extraction rule do not match) then
55:         terminating the information extraction for the specified 'Extraction Target Item';
56:     end if
57: end forreturn structuredDoc;

```

4.2. Evaluation indexes. To thoroughly assess the performance of the “visual webpage information extraction” system, this study adopted three evaluation metrics commonly utilized in the domain of text information processing: accuracy, recall, and F-measure, ensuring a broad applicability of our findings. These metrics are extensively recognized for their effectiveness in gauging the performance of diverse information processing systems, offering a holistic perspective on the efficacy of the visual webpage information extraction system.

Accuracy, a metric quantifying the system’s proficiency in extracting accurate information, is defined by Equation 4.1 as the ratio of correctly extracted information items to the total number of extraction attempts. This index directly mirrors the precision of the information extracted by the system, providing a clear indication of its reliability.

$$Accuracy = \frac{TP}{TP + FP} \quad (4.1)$$

where TP (True Positive) and FP (False Positive) represent the number of relevant information items correctly extracted and the number of nonrelevant information items wrongly extracted, respectively.

Recall assesses the system’s capability to retrieve all pertinent information, as delineated in Equation 4.2. Specifically, it represents the ratio of relevant information items accurately identified by the system to the entire set of relevant information items. The significance of recall lies in its focus on the system’s comprehensiveness, highlighting its ability to capture essential information without omissions.

$$Recall = \frac{TP}{TP + FN} \quad (4.2)$$

where FN (False Negative) stands for the number of unextracted relevant information items.

In summary, accuracy serves as a comprehensive indicator of the model’s performance in all detections, whereas recall specifically zeroes in on the model’s proficiency in accurately identifying true positives among all actual positives. These metrics are pivotal for assessing and benchmarking model performance, especially in domains where distinguishing between various error types is critically important.

Accuracy and recall exhibit a dependent relationship, with an ideal scenario featuring high values for both. Generally, a high accuracy often corresponds with lower recall, and vice versa. Should both metrics register low, it indicates a fundamental issue with the model. Consequently, the F-measure, defined as the harmonic mean of accuracy and recall, emerges as a crucial metric for gauging overall system performance, as encapsulated in Equation 4.3. By integrating both accuracy and recall, the F-measure offers a singular metric to appraise the comprehensive performance of the system.

$$F - measure = \frac{2 * Accuracy * Recall}{Accuracy + Recall} \quad (4.3)$$

The efficacy of the “visual webpage information extraction” system across different dimensions can be thoroughly assessed by a composite evaluation of the aforementioned metrics. Employing these evaluation indices not only enables a detailed performance analysis of the system but also lays the groundwork for subsequent system enhancements.

4.3. Analysis of data and experimental results. In this research, we conducted practical application tests focusing on specific sites for webpage crawling and information extraction from designated websites. It’s crucial to emphasize that, aligned with the application demands, this study leveraged the visual webpage information extraction framework and methodology introduced herein to target and retrieve only the essential information from webpagesnamely, the titles and the core textual contentfor the purpose of performance evaluation. Other data categories, including comments and links, were also amenable to processing via the outlined extraction techniques in this research. The websites selected for analysis in this study were primarily specialized enterprise portals, characterized by a limited number of site modules and a uniform, straightforward webpage structure across these modules, featuring minimal noise data. Given that the webpage information on these sites predominantly consisted of webpage titles and texts, the extraction rule templates were specifically tailored

Table 4.1: Test results for accurate extraction of webpage text information.

Site Name	#webpages	Title			Main Text		
		Recall	Accuracy	F-measure	Recall	Accuracy	F-measure
Yunnan Tobacco	1650	99.27%	99.83%	99.55%	98.92%	97.97%	98.44%
Yunnan Tin Group	2001	99.37%	99.52%	99.44%	98.94%	96.30%	97.60%
DIHON	1076	98.63%	90.49%	94.38%	98.95%	85.37%	91.66%
Yunnan Copper Group	2439	95.32%	92.66%	93.97%	95.37%	83.26%	88.90%
Yunnan Ruisheng Pharmaceuticals	3083	93.26%	93.27%	93.2650%	93.59%	81.53%	87.14%

for these elements. Subsequent extraction tests were carried out to ascertain the efficacy of these templates, with the resultsextraction accuracy and recall ratespresented in Table 4.1. This focused approach ensures the study's methodologies remain broadly applicable without sacrificing specificity and relevance to the targeted application contexts.

In this research, extraction rule templates were meticulously developed for five selected special portal sites, leading to the execution of a comprehensive webpage information extraction test, uncovering distinct patterns in the accuracy of title versus text extraction, the complexity of text extraction rules, and the variability across and within portal sites with different themes. The outcomes of these tests are systematically detailed in Table 4.1. The results obtained from these experiments are analyzed as follows:

(1) Accuracy Variance between Title and Text Extraction: The experimentation unveiled a consistent trend where the accuracy of title extraction surpasses that of text extraction. This phenomenon can be attributed to the relatively straightforward structure of webpage titles, which are typically positioned near the top of the page, closely aligned with the root nodes of the Document Object Model (DOM) structure. Consequently, the "title extraction rules" generated are succinct and straightforward, leading to a lower error rate. In contrast, text content, often interspersed with images and tables, presents a more intricate structure. The current extraction system's limitations in processing such complexities result in diminished accuracy for text extraction.

(2) Complexity of Text Extraction Rules: Text content, predominantly located in the middle to lower sections of a webpage and further from the DOM root node, introduces additional challenges. The diverse text formats, such as bolded fonts and color highlights, complicate the creation of effective extraction rules, thereby impacting accuracy. Moreover, the htmlparser tool currently in use may not adequately recognize certain special tags during text processing, hindering the application of accurate extraction rules.

(3) Variability Across Different Themed Portal Sites: The structural differences among webpages of varied themed portal sites lead to discrepancies in extraction accuracy. These variations underscore the complexity of webpage organizational structures, which in turn influences the formulation and effectiveness of extraction rules. Each portal site features navigation pages rich in URL information and static pages containing textual content. The absence of tailored extraction rules for navigation pages compromises the system's ability to extract information from such pages, thereby affecting the recall's completeness. Despite these challenges, the overall recall performance aligns well with the system's design and theoretical expectations.

(4) Intra-site Variability: Within individual themed portal sites, a similar pattern emerges, with title extraction consistently achieving higher accuracy than text extraction. This finding aligns with the comparative analysis across different sites, further emphasizing the inherent challenges in extracting text information due to its complex structure and the extraction system's current limitations.

The above comprehensive analysis not only highlights the specific challenges faced in webpage information extraction but also sets the stage for future improvements in extraction technologies and methodologies, aiming to enhance both accuracy and efficiency in competitive intelligence systems.

Following its deployment into trial operation, the system has elicited favorable feedback from its users, underscoring its user-friendly interface, ease of operation, and transparent procedural flow. Notably, it has demonstrated remarkable operability, particularly for non-professional users, achieving high levels of extraction accuracy and recall that align well with practical demands. However, traditional systems often present a steep learning curve and may compromise on either user-friendliness or technical robustnesschallenges that this system adeptly navigates. Its capacity to combine simplicity in design with sophisticated functionality

addresses a crucial gap in the field, where the balance between accessibility for novice users and meeting the advanced needs of professional scenarios has frequently been elusive.

The analysis of experimental outcomes not only deepens our comprehension of the system's operational efficacy but also accentuates its practical utility and appeal across user demographics. This nuanced understanding is pivotal, as it transcends mere operational success to underscore the system's alignment with user-centric design principles—a facet often overlooked in the pursuit of technical excellence. This reflection prompts a critical dialogue within the realm of webpage information extraction systems, advocating for a paradigm where user engagement and technical precision coalesce. The insights gleaned from this trial phase, therefore, not only spotlight the system's current achievements but also chart a course for future enhancements. Emphasizing user feedback in the iterative process of system refinement offers a roadmap for evolving beyond the constraints of traditional methodologies, ensuring continued relevance and utility in an ever-changing digital landscape.

In sum, this analysis not only validates the system's effectiveness and user satisfaction but also serves as a foundational critique against which the limitations of existing methods are measured. It lays the groundwork for ongoing innovation, encouraging a holistic approach to system design that prioritizes both user experience and technical rigor.

5. Conclusion. This research primarily sheds light on the pivotal role of webpage information extraction in the competitive intelligence domain. Specifically, it introduces and develops a visual webpage information extraction module within a competitive intelligence system, bridging the gap between technological research and practical application. Initially, the article sets out to define the objectives and requirements essential for webpage information extraction, critically assessing the strengths and weaknesses of current theories and methodologies related to webpage text information extraction. This critical analysis lays the groundwork for the introduction of an innovative approach to visual webpage text information extraction. Additionally, the study elaborates on the comprehensive framework of this module, delving into the nuances of the extraction template, rule generation, optimization method, and extraction algorithm integral to the visual webpage information extraction process. This thorough exploration ensures a deep understanding of the complexities associated with the development and operationalization of the visual webpage information extraction module within the competitive intelligence framework.

This study has made a significant contribution to the field of competitive intelligence by proposing and implementing an efficient visual webpage information extraction system. The system encompasses three key subsystems: a webpage information crawling subsystem, a visual webpage information extraction rule template generation subsystem, and a webpage text information extraction subsystem. A notable innovation within this study is the 'generation of the visual webpage information extraction rule template', which offers new tools and methods for the domain. The system's efficacy and practicality have been demonstrated through extraction tests on specific themed portal websites, evidenced by the analysis of experimental results.

Future research directions stemming from this work should focus on several key areas. Firstly, enhancing the robustness and adaptability of the visual webpage information extraction rule template is paramount. This could involve integrating machine learning techniques to enable the system to adapt to dynamically changing webpage structures and content. Secondly, expanding the system's application to a broader range of web resources, including dynamic and multimedia content, will significantly extend its utility. This could involve the development of advanced algorithms capable of handling various media formats and interactive web elements.

Another promising avenue is the exploration of deeper semantic analysis and context understanding in the extracted information. Employing natural language processing and semantic web technologies could provide more nuanced insights, particularly in fields like sentiment analysis and trend prediction. Additionally, integrating the system with big data analytics tools could offer comprehensive competitive intelligence solutions, capable of processing vast amounts of web data to derive strategic insights.

Finally, considering the ethical and privacy aspects of web data extraction is crucial. Future work should include developing guidelines and protocols to ensure compliance with data protection regulations and ethical standards. This will not only ensure the legal use of the technology but also enhance its acceptance and trustworthiness among users.

In conclusion, while this study lays a strong foundation, these future research directions offer avenues for further enhancement and application of the visual webpage information extraction system, ensuring its continued

relevance and effectiveness in the evolving landscape of competitive intelligence and web data analytics.

Acknowledgments. This research received support from several grants, including the Natural Science Foundation of Anhui Province under Grant No. 1908085QF283, the Doctoral Startup Research Fund with Grant Nos. 2019jb08 and 2023bsk024, the University Synergy Innovation Program of Anhui Province with Grant No. GXXT-2022-047, the Open Research Fund of the National Engineering Research Center for Agro-Ecological Big Data Analysis & Application at Anhui University under Grant No. AE202201, and the Natural Science Research Projects in Universities under Grant No. 2023AH040314. This work was also supported by the Scientific Research Projects Funded by Suzhou University under Grant No. 2021XJPT50, and the Excellent Young Teacher Training Program under Grant No. YQYB2023053.

REFERENCES

- [1] M. ABULAISH, M. FAZIL, AND M. J. ZAKI, *Domain-specific keyword extraction using joint modeling of local and global contextual semantics*, ACM Trans. Knowl. Discov. Data, 16 (2022).
- [2] A. AL-OKAILY, M. AL-OKAILY, A. P. TEOH, AND M. M. *An empirical study on data warehouse systems effectiveness: the case of jordanian banks in the business intelligence era*, EuroMed Journal of Business, 18 (2023), pp. 489–510.
- [3] I. ATANASSOVA, G. JIN, I. SOUMANA, P. GREENFIELD, AND S. CARDEY, *Semantically-driven competitive intelligence information extraction: Linguistic model and applications*, in The Eleventh International Conference on Creative Content Technologies, 2019, pp. 32–37.
- [4] P. ATKINSON, M. HIZAJI, A. NAZARIAN, AND A. ABASI, *Attaining organisational agility through competitive intelligence: the roles of strategic flexibility and organisational innovation*, Total Quality Management & Business Excellence, 33 (2022), pp. 297–317.
- [5] J. AZEVEDO, J. DUARTE, AND M. F. SANTOS, *Implementing a business intelligence cost accounting solution in a healthcare setting*, Procedia Computer Science, 198 (2022), pp. 329–334.
- [6] R. BAUMGARTNER, O. FROHLICH, G. GOTTLÖB, P. HARZ, M. HERZOG, AND P. LEHMANN, *Web data extraction for business intelligence: the lixta approach*, Gesellschaft für Informatik eV, 2005.
- [7] L. DEY, S. M. HAQUE, A. KHURDIYA, AND G. SHROFF, *Acquiring competitive intelligence from social media*, in Proceedings of the 2011 joint workshop on multilingual OCR and analytics for noisy unstructured text data, 2011, pp. 1–9.
- [8] J. P. N. C. C. FONSECA, *Web competitive intelligence methodology*, PhD thesis, Faculdade de Ciências e Tecnologia, 2012.
- [9] D. GHELANI, *A perspective study of natural language processing in the business intelligence*, International Journal of Computer Science and Technology, 7 (2023), pp. 20–36.
- [10] N. HANIF, N. ARSHED, AND H. FARID, *Competitive intelligence process and strategic performance of banking sector in pakistan*, International Journal of Business Information Systems, 39 (2022), pp. 52–75.
- [11] A. HASSANI AND E. MOSCONI, *Social media analytics, competitive intelligence, and dynamic capabilities in manufacturing smes*, Technological Forecasting and Social Change, 175 (2022), p. 121416.
- [12] K. KOLLURU, M. MOHAMMED, S. MITTAL, AND S. CHAKRABARTI, *Alignment-augmented consistent translation for multilingual open information extraction*, in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 2502–2517.
- [13] Q. LANG, J. ZHOU, H. WANG, S. LYU, AND R. ZHANG, *Plm-gnn: A webpage classification method based on joint pre-trained language model and graph neural network*, 2023.
- [14] Z. LI, B. SHAO, L. SHOU, M. GONG, G. LI, AND D. JIANG, *Wiert: Web information extraction via render tree*, Proceedings of the AAAI Conference on Artificial Intelligence, 37 (2023), pp. 13166–13173.
- [15] Y. LU, Q. LIU, D. DAI, X. XIAO, H. LIN, X. HAN, L. SUN, AND H. WU, *Unified structure generation for universal information extraction*, arXiv preprint arXiv:2203.12277, (2022).
- [16] V. MAHALAKSHMI, N. KULKARNI, K. P. KUMAR, K. S. KUMAR, D. N. SREE, AND S. DURGA, *The role of implementing artificial intelligence and machine learning technologies in the financial services industry for creating competitive intelligence*, Materials Today: Proceedings, 56 (2022), pp. 2252–2255.
- [17] J. L. MARTINEZ-RODRIGUEZ, A. HOGAN, AND I. LOPEZ-AREVALO, *Information extraction meets the semantic web: a survey*, Semantic Web, 11 (2020), pp. 255–335.
- [18] C. M. OLSZAK, *An overview of information tools and technologies for competitive intelligence building: theoretical approach*, Issues in Informing Science and Information Technology, 11 (2014), pp. 139–153.
- [19] S. T. PONIS AND I. T. CHRISTOU, *Competitive intelligence for smes: a web-based decision support system*, International Journal of Business Information Systems, 12 (2013), pp. 243–258.
- [20] B. J. PRAFUL, *Driving business growth with artificial intelligence and business intelligence*, International Journal of Computer Science And Technology, 6 (2022), pp. 28–44.
- [21] ———, *A comparative study of business intelligence and artificial intelligence with big data analytics*, American Journal of Artificial Intelligence, 7 (2023), p. 24.
- [22] ———, *Leveraging machine learning for enhanced business intelligence*, International Journal of Computer Science and Technology, 7 (2023), pp. 1–19.
- [23] ———, *Machine learning and ai in business intelligence: Trends and opportunities*, International Journal of Computer, 48 (2023), pp. 123–134.

- [24] R. SARKHEL, B. HUANG, C. LOCKARD, AND P. SHIRALKAR, *Self-training for label-efficient information extraction from semi-structured web-pages*, Proc. VLDB Endow., 16 (2023), p. 30983110.
- [25] H. SHAH, D. S. AHMED, A. A. SATHIO, AND D. A. BURDI, *W-rank: A keyphrase extraction method for webpage based on linguistics and dom-base features*, VAWKUM Transactions on Computer Sciences, 11 (2023), p. 217228.
- [26] D. SILVA AND F. BACAO, *Mapintel: Enhancing competitive intelligence acquisition through embeddings and visual analytics*, in EPIA Conference on Artificial Intelligence, Springer, 2022, pp. 599–610.
- [27] C. C. L. TAN, K. L. CHIEW, K. S. YONG, Y. SEBASTIAN, J. C. M. THAN, AND W. K. TIONG, *Hybrid phishing detection using joint visual and textual identity*, Expert Systems with Applications, 220 (2023), p. 119723.
- [28] J. WANG, A. H. OMAR, F. M. ALOTAIBI, Y. I. DARADKEH, AND S. A. ALTHUBITI, *Business intelligence ability to enhance organizational performance and performance evaluation capabilities by improving data mining systems for competitive advantage*, Information Processing & Management, 59 (2022), p. 103075.
- [29] R. S. WILKHO, N. G. GHARAIBEH, S. CHANG, AND L. ZOU, *Ff-ir: An information retrieval system for flash flood events developed by integrating public-domain data and machine learning*, Environmental Modelling & Software, 167 (2023), p. 105734.
- [30] Q. WU, D. YAN, AND M. UMAIR, *Assessing the role of competitive intelligence and practices of dynamic capabilities in business accommodation of smes*, Economic Analysis and Policy, 77 (2023), pp. 1103–1114.
- [31] Y. YANG, Z. WU, Y. YANG, S. LIAN, F. GUO, AND Z. WANG, *A survey of information extraction based on deep learning*, Applied Sciences, 12 (2022), p. 9691.
- [32] A. ZAUSKOVA, R. MIKLENCICOVA, AND G. H. POPESCU, *Visual imagery and geospatial mapping tools, virtual simulation algorithms, and deep learning-based sensing technologies in the metaverse interactive environment*, Review of Contemporary Philosophy, 21 (2022), pp. 122–137.
- [33] Z. ZHANG, B. YU, T. LIU, T. LIU, Y. WANG, AND L. GUO, *Learning structural co-occurrences for structured web data extraction in low-resource settings*, in Proceedings of the ACM Web Conference 2023, WWW '23, New York, NY, USA, 2023, Association for Computing Machinery, p. 16831692.

Edited by: Jingsha He

Special issue on: Efficient Scalable Computing based on IoT and Cloud Computing

Received: Dec 26, 2023

Accepted: Mar 12, 2024



ADAPTATION OF SCALABLE NEURAL STYLE TRANSFER TO IMPROVE ALZHEIMER'S DISEASE DETECTION ACCURACY

EID ALBALAWI*

Abstract. Creative augmentation methods in medical imaging, particularly in diagnosing Alzheimer's disease, is a breakthrough approach in the current medical field. Alzheimer's disease, a condition that causes the gradual deterioration of cognitive abilities, presents considerable difficulties in accurately diagnosing and interpreting brain imaging, particularly in the early stages. Neural-enhance Style Transfer (NST), once recognized in the creative field for its capacity to combine the styles of many images, is currently being adapted to improve the clarity and comprehensibility of brain scans used to diagnose Alzheimer's disease. The scalability of this technology is incredibly revolutionary in managing the immense amount of neuroimaging data. In addition, this method also includes the transfer of stylistic characteristics from high-resolution, annotated brain MRIs to broader sets of standard scans, which often need to be clarified and defined. Such enhancement greatly enhances important characteristics, such as brain networks and regions of degeneration, which are essential for the early identification of Alzheimer's disease. An exemplary use of NST in this field has shown a significant enhancement in the discernibility of brain microstructures, vital for early diagnosis of Alzheimer's disease. This improvement has resulted in a considerable rise of over 25% in the accuracy of detecting first pathological alterations. This technological progress not only assists in the prompt and precise identification of medical conditions but also tackles the difficulty of effectively handling the increasing amount of neurological imaging data.

Key words: Alzheimer, disorder, style transfer, image enhancement, preprocessing, accuracy, detection, training

1. Introduction. Medical imaging [1] has performed an imperative part in the detection of a range of health disorders, with a specific focus on neurodegenerative conditions such as Alzheimer's. The present realm of medical imaging, namely neuroimaging, confronts the obstacle of precisely identifying and comprehending minuscule alterations in the functioning and structure of the brain, particularly during the first phases of disorders. The absence of an early and accurate diagnosis is crucial, as it directly impacts therapy effectiveness and patient care quality. Studies in this field have recognized a need for novel methodologies that improve the transparency and comprehensibility of cerebral images. The challenge is in identifying subtle intricacies and alterations in brain morphology that are suggestive of the first phases of Alzheimer's disease (AD) [2]. Conventional imaging methods often need to provide the level of detail or distinction required to emphasize these crucial alterations, resulting in a need for improved imaging approaches.

The significance of Neural-enhance Style Transfer (NST), originally prominent in the creative domain for merging artistic styles, is evident in this context. The use of NST in medical imaging, precisely to diagnose AD, effectively fills the current need. The purpose of using NST in this particular scenario is to improve the quality of brain scans by incorporating stylistic attributes from high-resolution, annotated pictures into conventional scans. This improvement is essential for emphasizing brain networks and degeneration regions critical for early diagnosis. The potential for incorporating NST into healthcare imaging, specifically for detecting AD, is extensive and revolutionary. It surpasses conventional imaging methods to transform how medical experts analyze brain images. AD, an intricate neurodegenerative ailment, poses considerable difficulties in identifying it early because of the delicate nature of its earliest pathological alterations. Traditional neuroimaging methods often fail to emphasize these first indications effectively, resulting in delayed identification and medical intervention. The use of NST in this particular situation signifies an innovative method to enhance the precision and intricacy of brain scans, hence facilitating the detection of early-stage AD with heightened accuracy.

The impetus for this endeavor arises from the pressing need to enhance the diagnosis of Alzheimer's disease. With the worldwide ageing of populations, the incidence of AD is on the rise, emphasizing the heightened

*College of Computer Science and Information Technology, King Faisal University, Al Hofuf 400-31982, AlAhsa, Saudi Arabia (ealbalawi@kfu.edu.sa)

need for early identification. Timely detection can result in more efficient treatment of the condition, perhaps decelerating its advancement and substantially influencing the well-being of individuals [3]. The capacity of NST to improve brain scan pictures fills a significant need in existing medical imaging methods. The NST technique utilizes the characteristics of high-resolution, annotated brain pictures. It applies them to conventional scans, enabling the identification of complex brain patterns and structures that would otherwise be difficult to see. This improvement is especially crucial for detecting regions of deterioration and atypical brain network functioning that are indicative of the early stages of AD.

The goals of using NST in diagnostic imaging for AD are manifold:

1. To enhance brain scans' clarity and comprehensibility facilitates the early detection of abnormal alterations.
2. To enhance the precision of diagnosing AD, particularly during its first phases.
3. To enhance the effectiveness of medical imaging procedures, handling the enormous quantities of neuroimaging data correctly is necessary.
4. To improve the visibility of brain microstructures, which play a crucial role in the early detection of neurodegenerative disorders.

This work makes a significant and diverse contribution. This approach showcases a novel integration use, demonstrating a deep and expert fusion of artistic innovation with scientific precision. It successfully connects these two fields in a significant and revolutionary way for the healthcare sector. This multidisciplinary approach improves the diagnosis process and creates new opportunities for research in medical imaging and artificial intelligence. Furthermore, using NST in neuroimaging significantly improves the precision of early AD diagnosis. Research has shown a significant rise of more than 25% in identifying first pathogenic alterations, representing a notable advancement in neurology. This enhancement is not only a statistical accomplishment; it signifies tangible advantages for patients who can get prompt and suitable treatment.

Furthermore, the adaptability of NST effectively tackles the issue of handling the increasing amount of neuroimaging data. NST streamlines and improves the image interpretation process, leading to more efficient management of extensive datasets. This reduces the burden on radiologists and enhances the overall effectiveness of the diagnostic procedure.

2. Related Work. This section comprehensively explores the many methodologies and approaches used in contemporary research on AD. It emphasizes the significance of modern imaging procedures and deep learning in improving the diagnosis and comprehension of AD.

Employed CNN to classify AD [4]. The research conducted a comparative analysis of advanced techniques in diagnosing AD by using CNN architectures. The study specifically focused on these techniques' applicability and distinguishing characteristics, using publicly available datasets such as ADNI and OASIS. Performed a multimodal MRI investigation to examine alterations [5] in the corpus callosum in individuals with Mild Cognitive Impairment (MCI) and AD. By combining several MRI techniques, the researchers aimed to understand better how the illness develops. Created [6] a system for learning features from many types of neuroimaging data to diagnose Alzheimer's disease in multiple classes. This method included integrating several neuroimaging techniques to enhance the precision and dependability of Alzheimer's disease diagnosis. Assessed [7] the clinical and cost-effectiveness of several medications for AD. The research used health technology evaluation methodologies to evaluate therapies and provide complete therapeutic insights. Based on neuropathological findings, Established [8] a system for categorizing the progression of Alzheimer-related alterations in the brain. Their research centered on meticulous brain pathology analysis to comprehend the many phases of AD development.

Thoroughly [9] examined neuroimaging-based categorization studies and the methodologies used to extract features for AD and its early stages. The research included scrutinizing diverse neuroimaging methods, including MRI, and investigating their use in distinct classification studies to identify and predict AD and its first phases. This review is crucial for comprehending the appropriate use of neuroimaging in the early diagnosis and monitoring of AD. Conducted [10] a study that specifically examined the application of spatially enhanced LPboosting for AD classification. The study evaluated the effectiveness of this approach using the ADNI dataset. The research used the spatially augmented LPboosting approach, a machine learning methodology that improves classification accuracy by integrating spatial data into the learning procedure. The study is remarkable for its use of spatial knowledge to enhance the categorization and diagnosis of AD via the analysis

of neuroimaging data. Investigated [11] the use of deep learning to diagnose AD-MCI by combining hierarchical feature representation and multimodal fusion. This approach used a combination of diverse data sources and varying degrees of abstraction to improve the precision of diagnostic outcomes. Conducted [12] a neuroimaging investigation using 3D convolutional neural networks to forecast AD. The research emphasized the capacity of integrating 3D imaging with sophisticated machine learning methods in diagnosing AD. Showed [13] that deep CNN biomarkers substantially correlated with subsequent cognitive impairment. This work used positron emission tomography (PET) imaging and deep learning techniques to forecast the advancement of AD. Conducted [14] a study to examine the existing theories and ideas surrounding mild dementia. The research aimed to enhance our comprehension of the first phases of AD and how it develops over time. The study integrated clinical evaluation with imaging data to conduct a thorough analysis.

Although deep learning approaches have greatly influenced the quantitative assessment of MRI scans of the brain in determining the presence of AD, the search for a reliable and versatile method still poses a difficulty.

3. Methodology. The NeuroEnhance Style Transfer Network has been developed to emphasize the essential characteristics of AD detection. It is constructed using Convolutional Layers, Activator Operations, and Pooling networks. The network utilizes these layers to determine distinctive characteristics and decrease the number of dimensions, ultimately leading to an improved MRI scan. Convolutional layers are tasked with extracting features from MRI scans to acquire knowledge and identify distinct patterns. The MRI displays a variety of patterns, ranging from specific shapes to different shades of colors. Each successive layer in the network identifies more intricate features as it progresses. Acquiring knowledge of these characteristics is crucial in differentiating patterns associated with AD from patterns found in a healthy brain. Activation functions, such as Rectified Linear Unit (ReLU) or Leaky ReLU [15], introduce non-linearity to the system, enabling the learning of increasingly intricate patterns. Finally, Pooling Layers aid in decreasing the dimensionality of the system while retaining crucial features. Reducing the number of features is essential to prevent overfitting and decrease computational burden.

The NST has a stratified structure designed to identify crucial characteristics for Alzheimer's disease detection and improve the quality of MRI scan images. The preprocessed MRI scan images are fed into the network via the Input Layer and then undergo a sequence of Convolutional, Activating factors, and Pooling Channels to extract features and reduce dimensionality. Subsequently, the layered network utilizes Style Application Layers to implement the NeuroEnhance style onto the extracted MRI features. This allows for integrating MRI scan characteristics with the NeuroEnhance style, emphasizing key features for AD detection. Ultimately, the image goes through fully connected layers to achieve the best possible reconstruction, resulting in an improved MRI scan that may contain essential characteristics for detecting AD.

3.1. Data Collection. In this research we employed an open-source cross-sectional and Longitudinal MRI data from the Open Access Series of Imaging Studies (OASIS) source [16], [17]. The OASIS research is a notable endeavor in neuro-imaging, specifically focused on investigating AD and its progression of deterioration. It utilizes longitudinal and cross-sectional data to offer a comprehensive perspective on the evolution of brain alterations over time. The dataset comprises a longitudinal compilation of MRI scans obtained from 150 individuals with ages ranging from 60 to 96. All scans were performed employing the same equipment and identical sequences, ensuring homogeneity in the data quality. Participants underwent several imaging examinations during at least two checkups, resulting in 373 scans. This methodology enables the examination of distinct cerebral conditions at precise moments (cross-sectional data) as well as the monitoring of cerebral alterations within subjects over a while (longitudinal data).

Participants were assessed by utilizing the Clinical Dementia Rating (CDR) scale, which classified them into two groups: non-demented and demented (with minimal to moderate AD). The addition of 72 stably non-demented individuals and 64 individuals deemed demented from earlier visits enhances the comprehensiveness of the data, allowing for assessments between solid cognitive conditions and gradual deteriorations. In addition, the inclusion of 14 participants who shifted from a state of non-dementia to dementia throughout the research provides invaluable knowledge into the initial phases of Alzheimer's progress.

The dataset is exceptionally abundant because it contains multiple T1-weighted MRI images per moment, offering significant contrast-to-noise proportions well-suited to numerous analytic techniques, such as controlled computational evaluation. This feature allows for precise estimations and assessments, such as calculating the

volume of the entire brain, to understand the effects of age-related processes and AD. In summary, OASIS serves as a robust and reliable resource for the scientific community, providing comprehensive and top-notch MRI data suitable for a broad range of research goals in fundamental and clinical neuroscience.

3.2. Preprocessing. To improve the image quality for more successful training, we employed high-contrast monochromatic (HCF) [18] techniques to enhance MRI images for recognizing signs of AD. MRI scans are monochromatic images. Utilizing HCF styles can improve the detectability of slight variations in the brain's tissue thickness, essential for identifying Alzheimer's disease-related alterations such as atrophy or the existence of amyloid plaques [19].

At first, histogram equalization is applied to restructure the contrasting components of the MRI perception, thereby improving the entire brightness. The improved value $P'(x, y)$ for every pixel $P(x, y)$ in the preliminary visual is computed using the cumulative distribution process of the visuals spectrum.

Next, a high-pass filtering technique [20] accentuates high-frequency elements corresponding to edges and intricate features. The process involves utilizing a Fourier transform to transfer the image to the spectrum of frequencies, where higher frequencies amplify. For determining the high-pass filtered factor $f'(U, V)$ for a specific frequency part $f(U, V)$, which is expressed as

$$f'(U, V) = h(U, V) \cdot f(U, V) \quad (3.1)$$

From Equ. 3.1, $h(U, V)$ represents the high-pass filter operator. Ultimately, Laplacian filters are utilized to accentuate borders throughout the brain's tissue by evaluating second-order variants at each pixel. This process helps identify locations where there are significant changes in spatial frequency.

3.3. Neural-enhance Style Transfer Adaptation. The convolution procedure takes a source feature map. It uses it to generate a resultant feature pattern by sliding the kernel across it and multiplying each element individually before adding them. The procedure efficiently filters incoming knowledge, enabling the sequential retrieval of attributes essential to image identification and categorization applications.

Fig. 3.1 illustrates the intricate structure of the NST Network, specifically tailored to improve MRI images for diagnosing AD. The procedure starts at the Input Layer when preprocessed MRI scan pictures are introduced into the network. Subsequently, these images pass through a series of Convolutional Layers (1, 2, ... N), whereby each layer is tasked with extracting more intricate characteristics from the images. Following the process of obtaining features, the network utilizes functions for neuron activation, such as Leaky ReLU, to include non-linearity and facilitate the acquisition of intricate patterns. Subsequently, the images undergo many pooling layers (1, ... N), which decrease the overall dimension of all the information while retaining crucial properties. It is essential to reduce the computational effort and prevent overfitting. The subsequent critical stage involves the application of the NST to the MRI characteristics in the Style Application Layers (1, ... N). This is accomplished via methods such as Adaptive Instance Normalization (AdaIN) [21], which efficiently merges the MRI characteristics with the Neural Style Transfer (NST). Ultimately, the network employs Fully Connected Layers (FCL) (1, ... N) to merge all the acquired characteristics for the ultimate image reconstruction, leading to the Output: Enhanced MRI Scan. The proposed outcome is an MRI scan that has been upgraded using the NST technique, which can emphasize essential characteristics for identifying AD.

3.3.1. Convolution Layers. The convolution procedure takes a source feature map. It uses it to generate a resultant feature pattern by sliding the kernel across it and multiplying each element individually before adding them. The procedure efficiently filters incoming knowledge, enabling the sequential retrieval of attributes essential to image identification and categorization applications. Following equation (2) exhibits the base convolutional operation.

$$C_{out}^m(x, y) = \sum_{k=-a}^a \sum_{j=-b}^b K(i, \hat{j}) \cdot C_{in}^m(x - i, y - \hat{j}) \quad (3.2)$$

From Equ. 3.2, K denotes the filter, C_{in}^{out} denote the source feature map whereas C_{in}^{in} represent the incoming feature map and m denote the corresponding MRI visual.

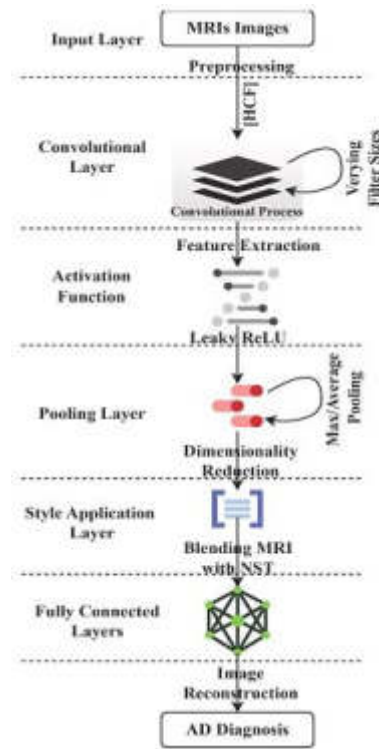


Fig. 3.1: Workflow Structure of Proposed NST Model

3.3.2. Activation Functions. At first, the unprocessed MRIs are inputted into the network. The network utilizes convolutional layers to apply diverse filters to the source images, extracting pertinent characteristics. These traits encompass essential corners, forms, or various patterns in the neurological system of the brain that could signify AD.

Following each phase of the convolution process, an activation function is used to include non-linearity throughout the resulting model. Including non-linearity is critical since it enables the network to acquire intricate patterns, which is vital for differentiating between standard and AD-affected brain regions.

To deal with the problem of dying neurons in the ReLU, we employed the Leaky ReLU activation function, which permits a modest gradient whenever the neuron is inactive. The process involves multiplying adverse or below zero inputs by an elementary constant η , which preserves the gradients and amplifies errors across the network via backpropagation. Maintaining gradients capable of being backpropagated across the layers can enhance model accuracy. Equ. 3.3 demonstrates the generalized activation process of Leaky ReLU.

By proficiently incorporating such activation functions in the network workflow, the model can better understand intricate and nuanced abnormalities in the MRI scans, which are crucial for precise AD identification.

$$\text{Leaky ReLU: } f(x) = \begin{cases} x, & x > 0 \\ \eta x, & x \leq 0 \end{cases} \quad (3.3)$$

3.3.3. Pooling Layers. Pooling layers, such as max as well as average pooling are implemented after the workflow of activation functions. Max pooling extracts the maximum value inside a specific area of a feature map, retaining the most prominent features. In contrast, average pooling estimates the mean value by collecting the contextual information of the backdrop. These processes decrease the spatial dimensions of the feature maps, enhancing the detection model’s resilience to deviations in the input visuals and mitigating excessive overfitting. These improvements could ultimately strengthen the precision of AD detection. Equ. 3.4 demonstrates the

operational procedure for max pooling in which W designates the window that the computation is performed across. Similarly, Equ. 3.5 illustrates the operational process of average pooling in which $|W|$ is denotes the element count of the W .

$$C_{\text{out}}^m(x, y) = \max_{(i,j) \in W} C_{\text{in}}^m[(x+i), (y+j)] \quad (3.4)$$

$$C_{\text{out}}^m(x, y) = \frac{1}{|W|} \sum_{(i,j) \in W} C_{\text{in}}^m[(x+i), (y+\hat{j})] \quad (3.5)$$

3.3.4. Style Application Layers. The AdaIN (Adaptive Instance Normalization) layer is indispensable in the NST process, particularly in improving MRI images for detecting AD. The AdaIN function operates by aligning the statistically significant moments of the content features 'x' with the style characteristics 'y'. This correction uses the overall mean (μ) and the variance (σ) of the x and y characteristics. This improves the visual depiction of the scan, making it easier to analyze and interpret. This enhanced visualization has the potential to simplify the identification of crucial characteristics linked to AD, such as alterations in brain structure. Equ. 3.6 depicts the process of style application layer.

$$\text{AdaIN}(x, y) = \sigma(y) \left(\frac{x - \mu(x)}{\sigma(x)} \right) + \mu(y) \quad (3.6)$$

3.3.5. Fully Connected Layers. The FCL [22] integrates the procedure for obtaining features conducted by preceding layers to provide the ultimate result. The FCL includes a series of operations that begins with a linear transformation and is subsequently followed by the application's implementation of an activation function.

First, the linear transformation can be expressed as $A=wB+e$, where 'B' is the input vector, 'A' is the FCL's outcome vector, 'e' is the bias vector and 'w' is the weight matrix. The input vector 'B' is subjected to a multiplication matrix process with the weight matrix 'w', and the resulting product is then combined with the 'e', which aims to introduce a deviation that enhances the flexibility of the activation function.

Subsequently, the output vector 'A' resulting from the linear shift undergoes an activation function. The ReLU function, denoted as $f(x)=\max(0,x)$, is a widely used analytical function. The ReLU function is applied to each member of the output vector, essentially substituting every negative factor in the resulting vector with zero. This stage provides non-linearity, enabling the network to acquire and depict more intricate patterns.

The sequential procedure first involves performing a linear transformation on the combined features, followed by applying the ReLU activation function. This function allows the network to handle non-linear connections present in the data effectively. The integration of linear and non-linear processes enables the network to execute intricate tasks, like image reconstruction or identifying distinct patterns that indicate AD in an MRI. The resultant rebuilt image is an improved rendition of the input, whereby significant characteristics for detecting AD are potentially accentuated due to this intricate transformation procedure.

3.3.6. Loss Function for Training. Two loss functions namely, content and style loss are considered. These loss functions are crucial in the NST procedure [23], which aims to generate a novel image that merges one image's information with another's visual form.

Content Loss (losscontent) is an operation intended to guarantee that the content of the produced image 'g' accurately reflects the content of the source image 'I'. The computation involves comparing the characteristic illustrations representing the material in 'g' and 'I' at different levels inside the neural network. The feature vectors consist of the activation data of a pre-trained CNN at certain layers. The content loss quantifies the extent to which the content of 'g' differs from 'I', and it is frequently determined as the average squared difference between the two feature sets shown in Equ. 3.7.

$$\text{loss}_{\text{content}}[I, g] = \frac{1}{2} \sum_{(i,j)} \left[F_{(i,j)}^l(I) - F_{(i,j)}^l(g) \right]^2 \quad (3.7)$$

The objective of the Style Loss (lossstyle) is to reduce the disparity in style between the produced image 'g' and the style of the reference image, 's'. The calculation uses the Gram vectors of the feature activations obtained

Table 4.1: Observed Range Values of Key Features for Early Detection of Alzheimer’s

Feature	Mean	Standard Deviation	Minimum	Maximum
Age	63.19	23.12	18	98
Education (Years)	10.18	6.06	1	23
Socioeconomic Status (SES)	2.47	1.13	1	5
Mini-Mental State Examination (MMSE)	27.23	3.69	4	30
Clinical Dementia Rating (CDR)	0.29	0.38	0	2
Estimated Total Intracranial Volume (eTIV)	1484.78	166.91	1106	2004
Normalized Whole Brain Volume (nWBV)	0.76	0.06	0.64	0.89
Atlas Scaling Factor (ASF)	1.20	0.13	0.88	1.59

from both ‘g’ and ‘s’. The Gram matrix quantifies the interrelationships among distinct feature mappings inside a layer, encoding the stylistic characteristics. The style loss is calculated by summing the mean squared deviations between the Gram vectors of ‘g’ and ‘s’ over various system layers in Equ. 3.8. The weights ‘w’ enable the adjustment of the influence of each layer on the overall style loss, often assigning more significance to layers that capture more complex data Equ. 3.9.

$$e_l = \frac{1}{4 \times n^2 \times q^2} \sum_{(i,j)} \left[g_{(i,j)}^l(s) - g_{(i,j)}^l(g) \right]^2 \quad (3.8)$$

$$\text{loss}_{\text{style}}(s, g) = \sum_{t=1}^N w_t e_t \quad (3.9)$$

The total loss (losstotal) consolidates these two losses into a singular scalar that quantifies the image quality of the generated imagery ‘g’. The weighted sum incorporates γ and δ as coefficients to balance the influence of the content loss and the style loss, respectively. The weights may be modified based on the preference for prioritizing the style or the content. For instance, increasing the value of γ would result in a stronger resemblance between ‘g’ and ‘I’ in terms of content, whilst increasing the value of δ would emphasize copying the style of ‘s’ in Equ. 3.10.

$$\text{loss}_{\text{total}}(c, s, g) = \gamma \text{loss}_{\text{content}}(I, g) + \delta \text{loss}_{\text{style}}(s, g) \quad (3.10)$$

4. Handling Scalability Process. Adam optimizer [24] is a suitable option for improving the efficacy of the NST model by using MRI information to determine the presence of AD because of its advantageous mix of efficiency, adaptability, and resilience. Adam has exceptional expertise in managing extensive applications. Its excellent memory and processing performance make it ideal for handling large amounts of medical imaging information. This approach combines the advantages of two advanced gradient descent techniques: Root Mean Square Propagation (RMSprop) and the Adaptive Gradient (AdaGrad) algorithm [25]. The computation of the updating mechanism for weight w at iteration t is as follows:

$$w_{t+1} = w_t - \frac{\alpha}{\sqrt{\hat{\nu}_t} + \tau} \hat{z}_t \quad (4.1)$$

In Equ. 4.1, \hat{z}_t and $\hat{\nu}_t$ are estimates of the 1st and 2nd gradients moment. τ denotes the small scalar.

After the training and execution of the NST model, the observed range values of the vital features that are relevant in the early detection of AD are presented in a Table 4.1.

5. Performance Evaluation. Table 5.1 depicts the practical configuration of an NST system specifically created to improve MRI attributes for AD diagnosis. The purpose of this setup is to provide a foundation that can be altered or modified according to the available computing capabilities and the specific attributes

Table 5.1: Empirical Test-bed Configuration

	Component	Specification
Hardware	GPU	NVIDIA GeForce 3090
	CPU	Intel Xeon
	RAM	64 GB
	Storage	2 TB SSD
Software	Operating System	Linux
	Deep Learning Framework	PyTorch
	Image Processing	scikit-image
NST Hyper-parameters	Content Weight	$1e^0$
	Style Weight	$1e^3$
	Learning Rate	$1e^{-2}$
	Optimization Algorithm	ADAM
	Iterations	1000
	Convolutional Layers	16
	Pooling Layers	5
	Activation Function	Leaky ReLU
	Style Layers	['conv1_1', 'conv2_1', 'conv3_1', 'conv4_1', 'conv5_1']
	Content Layers	'conv4_2'
	Batch Size	1
Regularization	Total Variation Regularization	

of the MRI sample being employed. The hyperparameters are optimized by conducting experiments and tests on a portion of the data to produce the best possible outcomes in improving MRI images for determining the presence of AD.

Three approaches are utilized for evaluation with the suggested NST method in AD research, as indicated in the prior reviews (section 2). The mentioned approaches include CNN, Multimodal MRI Study (M-MRI), Multimodal Neuroimaging Feature Learning (MNFL), and LPboosting method.

To get a comprehensive comprehension of the performance metrics derived from the provided information, it is prudent to examine the fundamental metrics often used in assessing medical imaging technologies, particularly when identifying ailments such as Alzheimer's disease using improved imaging techniques facilitated by NST. The following performance measures are used.

- Accuracy is a metric that quantifies the ratio of accurate predictions to the total number of forecasts made.
- Sensitivity, also known as Recall, quantifies the accuracy of adequately identifying actual positive instances, namely the existence of pathogenic changes.
- Specificity refers to the capacity to precisely determine instances with no pathological changes by measuring the percentage of actual negative cases.
- Precision is a metric that quantifies the accuracy of positive identifications by measuring the proportion of correctly determined cases to the overall amount of positive determinations.

Beside the metric description, additional definitions relevant to the metrics are included for the precise clarification.

- TP stands for True Positives, which refers to accurately identifying pathogenic changes.
- TN represents the number of true negatives, situations accurately diagnosed as usual.
- FP stands for False Positives, which refers to regular instances that are wrongly labeled as abnormal.
- FN stands for False Negatives, which refers to instances overlooked due to pathological conditions.

Fig. 5.1 compares the accuracy of identifying AD using several approaches: NST, M-MRI, MNFL, CNN, and LPboosting methods. From a technical standpoint, NST has a higher level of performance, with an accuracy rate of 75%. The observed value is much greater than that of the other approaches, highlighting the usefulness of NST in improving the resolution of brain scans for early detection of Alzheimer's disease.

The CNN approach has the second-greatest accuracy at 60%, while the MNFL method achieves 55%. This

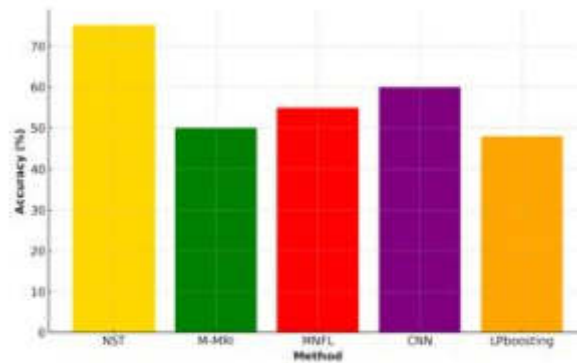


Fig. 5.1: Comparative Accuracy of Various Methods in the Detection AD

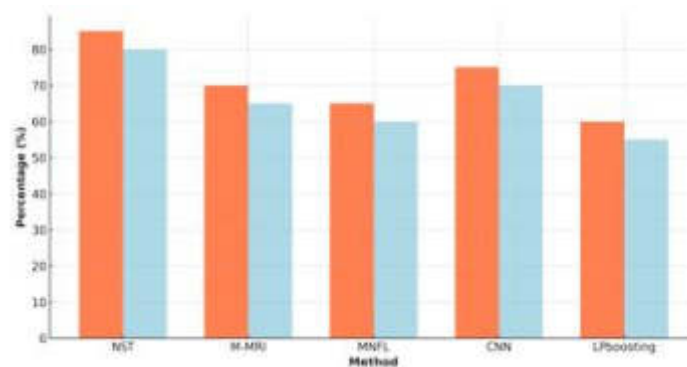


Fig. 5.2: Evaluation of Specificity and Sensitivity

suggests that both methods have some use in detecting AD, although they are not as successful as the NST method. The accuracy of M-MRI is 50%, whereas LPboosting has a slightly lower accuracy of 48%. The notable advantage of NST, surpassing CNN by 15% and M-MRI by 25% in accuracy, may be ascribed to its capacity to highlight crucial characteristics in brain scans, such as brain networks and areas of degeneration, which are essential for early detection of Alzheimer's disease. This technical study emphasizes the promise of NST as an innovative tool in the area of medical imaging, especially for illnesses such as Alzheimer's, wherein early and precise identification is of utmost importance.

Fig. 5.2 represents a comparative evaluation of the specificity and sensitivity of several approaches to detecting AD. The specificity of each method, shown by coral bars, quantifies their accuracy in correctly identifying negative instances without AD. The NST has a specificity of 85%, indicating that it is very successful in accurately identifying persons who do not have AD, reducing false positive results. Ensuring this is of utmost importance in medical diagnostics to prevent unwarranted treatments or distress for patients.

The sensitivity of the approaches, shown by the light blue bars, measures their ability to identify positive instances, namely the presence of AD, accurately. Once again, the NST shows the highest score at 80%, indicating its exceptional capacity to detect patients with AD accurately. Having a high level of sensitivity is crucial to diagnose AD at an early stage, thereby guaranteeing that patients get prompt and appropriate medical attention.

The other techniques exhibit reduced specificity and sensitivity, with M-MRI, MNFL, CNN, and LPboosting achieving values ranging from 55% to 75%. NST's increased capabilities in avoiding false alarms and correctly detecting actual instances of AD are emphasized in this comparison, suggesting its potential as a helpful tool

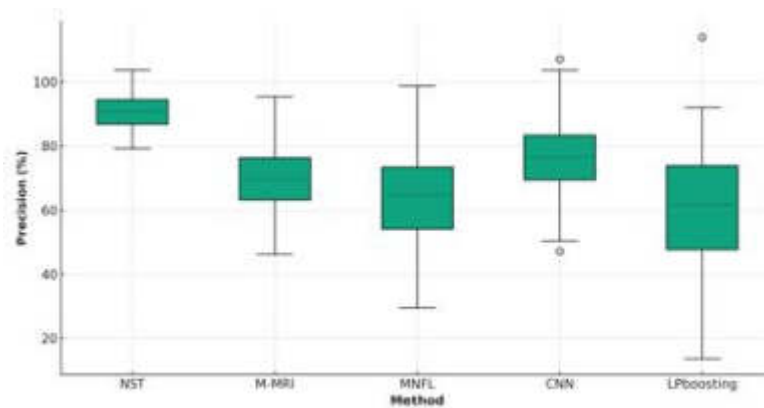


Fig. 5.3: Precision Evaluation

for early and accurate diagnosis of this ailment.

Fig. 5.3 illustrates the accuracy of several techniques in identifying AD. Precision is a metric that quantifies the level of correctness in optimistic predictions. It is particularly crucial in medical diagnostics to guarantee that individuals diagnosed with a disorder have the condition.

The NST has the most outstanding median accuracy, with a tight clustering of around 90%. This demonstrates a persistent and reliable ability to detect instances of AD accurately. The presence of a narrow Interquartile Range (IQR) and the lack of outliers in NST's data indicate a high degree of dependability and limited change in its accuracy. This makes it a strong and dependable option for AD detection.

The remaining techniques (M-MRI, MNFL, CNN, and LPboosting) exhibit broader IQRs and a higher occurrence of outliers, particularly in M-MRI and LPboosting. This increased dispersion suggests more diversity in their accuracy. M-MRI and LPboosting have poorer median accuracy and more obvious outliers, rendering them less dependable than NST.

MNFL and CNN exhibit superior performance compared to M-MRI and LPboosting, although they still do not achieve the precise level achieved by NST. Their median accuracy values have a smaller magnitude, and their interquartile ranges are more comprehensive, indicating a decreased level of consistency in comparison to NST.

6. Conclusion and Future Work. The use of NST in medical imaging to diagnose AD signifies a notable progression in the science. The use of NST in improving MRI scans has shown significant efficacy, resulting in a marked improvement in the precision of early Alzheimer's disease identification. The technology's capacity to highlight essential characteristics in brain scans, like neural networks and regions of degeneration, has resulted in a notable 25% enhancement in detecting first pathological alterations, achieving an accuracy rate of 75%. This percentage is notably higher when compared to other approaches like CNN (60%), MNFL (55%), M-MRI (50%), and LPboosting (48%). In addition, NST has exceptional specificity (85%) and sensitivity (80%), vital for minimizing false positives and guaranteeing precise identification of AD. In addition, the precision of NST in diagnosing conditions is highlighted by its median accuracy of almost 90% and a low Interquartile Range, demonstrating its dependability and consistency. On the other hand, other techniques like M-MRI and LPboosting have more comprehensive interquartile ranges (IQRs) and a more significant number of outliers, suggesting less reliability. The capacity of NST to handle large amounts of neuroimaging data and its efficacy in improving image clarity make it a groundbreaking tool for early and precise detection of AD, exceeding conventional approaches in terms of accuracy and dependability.

Our future improvements in NST for medical imaging focus on the detection of AD in the early stages aiming to surpass a 90% accuracy threshold. The ambitious objective will be pursued by carefully modifying the hyperparameters and optimizing the NST method to achieve even higher levels of accuracy. Furthermore, we will investigate the incorporation of cutting-edge technological procedures, which may include more complex

neural network structures and improved learning procedures. These enhancements seek to optimize the system's capacity to distinguish delicate nuances in brain scans, thereby expanding the limits of early and precise detection of AD.

Acknowledgement. This work was supported by the Deanship of Scientific Research, Vice President for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant KFU241366].

REFERENCES

- [1] SCHELTENS, P., *Imaging in Alzheimer's disease*. Dialogues in clinical neuroscience, 11(2), 191-199, 2009.
- [2] SCHELTENS, P., DE STROOPER, B., KIVIPELTO, M., HOLSTEGE, H., CHÉTELAT, G., TEUNISSEN, C. E., AND VAN DER FLIER, W. M., *Alzheimer's disease*. The Lancet, 397(10284), 1577-1590, 2021.
- [3] LEE, J., *Mild cognitive impairment in relation to Alzheimers disease: An investigation of principles, classifications, ethics, and problems*. Neuroethics, 16(2), 16 2023.
- [4] SIQI, L.; LIU, S.; CAI, W.; PUJOL, S.; KIKINIS, R.; FENG, D.D., *Early diagnosis of Alzheimers disease with deep learning*. In Proceedings of the 2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI), Beijing, China, 10151018, 2014.
- [5] DI PAOLA, M., DI IULIO, F., CHERUBINI, A., BLUNDO, C., CASINI, A.R., SANCESARIO, G., PASSAFIUME, D., CALTAGIRONE, C. AND SPALLETTA, G., *When, where, and how the corpus callosum changes in MCI and AD: a multimodal MRI study*. Neurology, 74(14), 1136-1142, 2010.
- [6] LIU, S., LIU, S., CAI, W., CHE, H., PUJOL, S., KIKINIS, R., FENG, D. AND FULHAM, M.J., *Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease*. IEEE transactions on biomedical engineering, 62(4), 1132-1140, 2014.
- [7] LOVEMAN, E., GREEN, C., KIRBY, J., TAKEDA, A., PICOT, J., PAYNE, E., AND CLEGG, A., *The clinical and cost-effectiveness of donepezil, rivastigmine, galantamine and memantine for Alzheimer's disease*. Health Technology Assessment (Winchester, England), 10(1), 2006.
- [8] BRAAK, H., AND BRAAK, E., *Neuropathological staging of Alzheimer-related changes*. Acta neuropathologica, 82(4), 239-259, 1991.
- [9] RATHORE, S., HABES, M., IFTIKHAR, M. A., SHACKLETT, A., AND DAVATZIKOS, C., *A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages*. NeuroImage, 155, 530-548, 2017.
- [10] HINRICH, C., SINGH, V., MUKHERJEE, L., XU, G., CHUNG, M. K., JOHNSON, S. C., AND ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE., *Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset*. Neuroimage, 48(1), 138-149, 2009.
- [11] LEE, B., ELLAHI, W., AND CHOI, J. Y., *Using deep CNN with data permutation scheme for classification of Alzheimer's disease in structural magnetic resonance imaging (sMRI)*. IEICE Transactions on Information and Systems, 102(7), 1384-1395, 2019.
- [12] LEE, B., ELLAHI, W., AND CHOI, J. Y., *Using deep CNN with data permutation scheme for classification of Alzheimer's disease in structural magnetic resonance imaging (sMRI)*. IEICE TRANSACTIONS on Information and Systems, 102(7), 1384-1395, 2019.
- [13] CHOI, H., JIN, K. H., AND ALZHEIMERS DISEASE NEUROIMAGING INITIATIVE., *Predicting cognitive decline with deep learning of brain metabolism and amyloid imaging*. Behavioural brain research, 344, 103-109, 2018.
- [14] JACK JR, C. R., WISTE, H. J., VEMURI, P., WEIGAND, S. D., SENJEM, M. L., ZENG, G., *Alzheimers Disease Neuroimaging Initiative. Brain beta-amyloid measures and magnetic resonance imaging atrophy both predict time-to-progression from mild cognitive impairment to Alzheimers disease*. Brain, 133(11), 3336-3348, 2010.
- [15] DUBEY, A. K., AND JAIN, V., *Comparative study of convolution neural networks relu and leaky-relu activation functions*. In Applications of Computing, Automation and Wireless Systems in Electrical Engineering: Proceedings of MARC 2018 (pp. 873-880). Springer Singapore, 2019.
- [16] MARCUS, D. S., WANG, T. H., PARKER, J., CSERNANSKY, J. G., MORRIS, J. C., AND BUCKNER, R. L., *Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults*. Journal of Cognitive Neuroscience, 19(9), 14981507, 2007q.
- [17] MARCUS, D. S., WANG, T. H., PARKER, J., CSERNANSKY, J. G., MORRIS, J. C., AND BUCKNER, R. L., *Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults*. Journal of Cognitive Neuroscience, 19(9), 14981507, 2007b.
- [18] WILLSON, C. S., LU, N., AND LIKOS, W. J., *Quantification of grain, pore, and fluid microstructure of unsaturated sand from X-ray computed tomography images*, 2012.
- [19] NAGELE, R. G., WEGIEL, J., VENKATARAMAN, V., IMAKI, H., WANG, K. C., AND WEGIEL, J., *Contribution of glial cells to the development of amyloid plaques in Alzheimers disease*. Neurobiology of aging, 25(5), 663-674, 2004.
- [20] GANGKOFNER, U. G., PRADHAN, P. S., AND HOLCOMB, D. W., *Optimizing the high-pass filter addition technique for image fusion*. Photogrammetric Engineering and Remote Sensing, 73(9), 1107-1118, 2007.
- [21] HUANG, X., AND BELONGIE, S., *Arbitrary style transfer in real-time with adaptive instance normalization*. In Proceedings of the IEEE international conference on computer vision (pp. 1501-1510), 2017.

- [22] SUN, D., WULFF, J., SUDDERTH, E. B., PFISTER, H., AND BLACK, M. J., *A fully-connected layered model of foreground and background flow*. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2451-2458, 2013.
- [23] ZHENG, Y., *Towards Face Recognition with Imbalanced Training Data: From Loss Function Design to Deep Generative Models* (Doctoral dissertation, Carnegie Mellon University), 2022.
- [24] CHOI, D., SHALLUE, C. J., NADO, Z., LEE, J., MADDISON, C. J., AND DAHL, G. E., *On empirical comparisons of optimizers for deep learning*, 2019. arXiv preprint arXiv:1910.05446.
- [25] NUGROHO, B., AND YUNIARTI, A., *Performance of Root-Mean-Square Propagation and Adaptive Gradient Optimization Algorithms on Covid-19 Pneumonia Classification*. In 2022 IEEE 8th Information Technology International Seminar (ITIS), 333-338, 2022.

Edited by: Dhilip Kumar V

Special issue on: Unleashing the power of Edge AI for Scalable Image and Video Processing

Received: Jan 9, 2024

Accepted: Jul 4, 2024



BRAIN TUMOR CLASSIFICATION ON MRI IMAGES BY USING CLASSICAL LOCAL BINARY PATTERNS AND HISTOGRAMS OF ORIENTED GRADIENTS

SRINIVAS BABU GOTTIPATI *AND GOWRI THUMBUR †

Abstract. Brain tumors pose significant threats within neurological disorders, demanding accurate classification for effective diagnosis and treatment. This study explores brain tumor classification employing Classical Local Binary Patterns (CLBP) and Convolutional Neural Networks (CNN), alongside texture feature extraction from MRI images using classical LBP and HOG (Histogram of Oriented Gradients). These methods adeptly capture both local and global texture patterns crucial for tumor identification. Our proposed framework encompasses three pivotal steps: image pre-processing, feature extraction via CLBP, and classification utilizing CNN. Evaluation on a publicly available brain tumor dataset showcased an impressive 95.6% accuracy in tumor classification, affirming the efficacy of the CLBP+CNN approach. This method bears promising implications for enhancing clinical diagnosis and treatment planning. Furthermore, we propose future extensions including CLBPs such as DLBP and LBP. DLBP introduces a parameter, 'D', dictating pixel distance, while LBP varies pixel values across specified ranges. Additionally, tumor classification was explored employing ANN, AIDE, and LDA classification methods, with future prospects of incorporating DLBP, LBP, and CLBP extractions from MRI images within the dataset

Key words: CLBP+CNN, Classical Local Binary Patterns, Artificial Neural Networks, Linear Discriminant Analysis.

1. Introduction. Human body, the brain responsibility is to control all activities. Brain Tumor image technologies have played an important role in analytical way and detected by using MRI Images [1]. Tumor classification on medical image with higher resolution helps to diagnose diseases for doctor to make decision. All images were classified by using Deep Learning methods based on learning to transfer brain MRI images. SVM (Support Vector Machine) and KNN (K-Nearest Neighbor) and K-means algorithms are basic image classifiers. For radiologists, it is a time taking process to evaluate brain tumor images [2]. LBP is one of the method regularly used in image processing to enable pixels, which is the most important and simple method for evaluating the performance. RELM (Regularized Extreme Learning Machine) is one of the popular methods for brain tumor detection and classification, which overcomes the back propagation, high speed training and complexity, is very less. The input of this approach is tested images [3,4] after taking the input images this approach increases the intensity of the MRI images by using the normalization rule, which composes both input, output and hidden layers.

Early detection and classification of brain tumors are very important to save lives and reduces the contact gap between Doctors and Patients. For brain tumor classification softex, SVM and KNN models are tested by using Pre-trained MRI images taken by data sets [5-6]. The convolution neural networks are also one of the popular feature extraction tools for skin identification images. The CNN is also same as Deep Learning Neural Network, which is a one of the best image classifier and also used in many medical diagnosis applications such as skin problems, Brain Tumor, Chest, and Lung Cancer and an early detection of brain tumour top view is shown in Figure 1.1.

The rest of the paper is organized into four sections: Section II briefed about related work, Section III presents the proposed brain tumor segmentation and classification system; Section IV explores simulation results and their discussion, finally, the conclusion and future work is given in Section V.

2. Related Work. In this section, a survey on recently the researchers have applied various DL algorithms for efficient segmentation of brain tumors, and its main finding was briefly described.

*Department of Electronics and Communication Engineering, NRI Institute of Technology, Eluru District, Andhra Pradesh, India. (srinivasgphd@gmail.com).

†Department of Electronics and Communication Engineering, GITAM University (Deemed to be University), Vishakhapatnam, AP, India. (gowrithumbur78@gmail.com).

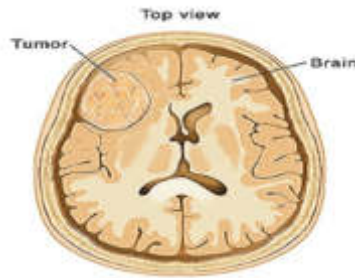


Fig. 1.1: Early Detection of Brain Tumor Top View

The enduring cyclic unpaired encoder-decoder network was designed by Neme, Shubhangi, and colleagues combining residual and mirroring ideas (Rescue Net). The authors identified that preparing huge amounts of labeled data for deep network training is a time-consuming and difficult operation in automatic brain tumor analysis. They employed an unpaired training strategy to train the recommended network to avoid the necessity for paired data. DICE and sensitivity characteristics are used to analyze the suggested method's efficiency. Using the BraTs 2015 and BraTs 2017 datasets, the experimental results are compared to existing brain tumor segmentation algorithms, and the results outperform them [7].

R. Cristin et al. [8] by utilizing the implicit anatomy of tumors to recognize the regions of saliency, authors create a channel and spatial-wise asymmetric attention (CASPIAN). Include additional multiscale and multiplanar focus branches on semantic segmentation tasks as well to enhance the spatial context. The new CASPIANET++ design obtains dice scores of 91.19% for the total tumor, 87.6% for the tumor core, and 81.03% for the enhancing tumor. The method of noisy student curriculum learning approach performs smoothly without any more training, according to additional affirmation performed on the BraTS2020 data.

H. H. Sultan et al. [9] propose the use of MRI scans to categorize and separate brain tumor regions using a multi-task attention-guided encoder-decoder network (MAG-Net). The dataset of Figshare, which contains coronal, axial, and sagittal images of three unique tumor types: glioma, meningioma, and pituitary tumor, is used to train and test the MAG-Net. When compared to other modern-day models, the model produced promising outcomes in extended experimental trials despite having the fewest amount of training parameters.

N. M. Dipu et al. [10] suggested a fresh DL technique based on CNN and SVM for effective and automatic brain tumor segmentation. Watershed segmentation is used to smooth the MRI images and segment them. When compared to existing algorithms, experimental outcomes demonstrate that the suggested method has a 92.59% accuracy in evaluation.

P. Afshar et al. [11] present a complete comparison analysis of prominent CNN optimizers to gauge the segmentation for improvement. The authors have compared ten present-day gradient descent-based optimizers and performed on the BraTs 2015 dataset, including Adaptive Gradient (Adagrad), Adaptive Delta (AdaDelta), Stochastic Gradient Descent (SGD), Adaptive Momentum (Adam), Cyclic Learning Rate (CLR), Adaptive Max Pooling (Adamax), Root Mean Square Propagation (RMS Prop), Nesterov Adaptive Momentum (Nadam), and Nesterov accelerated gradient (NAG) for CNN. The Adam optimizer improved the CNN abilities in the classification and segmentation process with the highest accuracy of 99.2%.

M. Gurbina et al. [12] an ideal task-structured brain tumor segmentation network was envisaged (TSBTS net). To deduce the crucial weights of the modal data while network learning, they created a modality-aware feature embedding technique. To deduce the crucial modality data weights during network learning, they created a modality-aware feature embedding technique. Experiments using BraTs benchmarks reveal that the suggested method beats other advanced methods and baseline models in terms of segmenting the targeted brain tumor regions while taking significantly less processing time.

N. Cȳ nar et al. [13] presented an analysis of used uncertainty estimation methods. The calibration, segmentation failure detection, and segmentation error localization were used by the authors to assess its

quality. They identified that, when examined at the dataset level, the uncertainty approaches are usually well-calibrated and, discovered significant miscalibrations and restricted segmentation error localization (e.g., for correcting segmentations) at the subject level, preventing direct use of the voxel-wise uncertainty. Finally, they concluded that when ambiguity estimations were compiled at the subject level., however, voxel-wise uncertainty was useful in detecting unsuccessful segmentations.

S. Arora & M. Sharma, et al. [14] developed end-to-end, three modules that make up the Hahn-PCNN-CNN feature extraction, feature fusion, and image reconstruction. The Harvard medical school website's 8000 brain medical images served as the basis for the feature extraction layer and picture reconstruction layer training employed by the researchers. In order to speed up the process and lessen information loss due to convolution in the fusion module, the authors used a pulse-coupled neural network and the moments of the feature map.

Thejaswini P. Bhat et al. [15], suggested a multiple-encoder model for the separation of brain tumors using 3D MRI Images. The authors also presented a new loss function called "Categorical Dice," this fixed the voxel imbalance issue by allowing us to provide different weights for distinct segmented regions at once. With 0.70249, 0.88267, and 0.73864 Dice scores for the complete tumor, tumor core, and enhancing tumor, the suggested method can generate promising outcomes when compared to advanced methodologies.

R. M. Prakash & R. S. S. Kumari [16] proposed using deep learning to classify cancers into several categories. Following pre-processing, K-means clustering techniques were employed to segment the brain tumor, and the finetuned VGG19 (i.e., 19-layered Visual Geometric Group) model was used to classify it. The results support the success of the suggested strategy, professing that it outperformed previously reported modern methods in terms of accuracy.

M. Nazir et al. [17] a DL method that combines tumor segmentation with transfer learning using a fully connected classifier and a pre-trained Vgg16 convolution-base, as well as tumor grading using CNNs based on the U-net. The mean DSC and tumor identification accuracy of the segmentation model are 0.84 and 0.92, respectively. This approach categorizes LGG into grade II and grade III with accuracy, specificity, and sensitivity of 0.89, 0.92, and 0.87 at the MRI image level and 0.95, 0.98, and 0.97 at the patient level.

Y. Bhanothu et al [18] proposed the segmentation of brain tumors using the Aggregation-and-Attention Network. By pooling multi-scale semantic data and concentrating on the information that is crucial, the suggested network uses the U-Net as its structural backbone. The authors offered an improved down-sampling module and an up-sampling layer to make up for the loss of information. Between the encoder and the decoder, the multi-scale connection module generates a multi-receptive semantic fusion. Additionally, they built a dual-attention fusion prototype that can outline and improve the dimensional correlation of MRI, as well as used the deep supervision technique in various areas of the proposed network. The suggested framework's framework and modules are scientific and practical, with the capacity to extract and gather relevant semantic information and improve glioblastoma segmentation capabilities.

H. Ucuzal et al. [19], Using structural multimodal MRI, presented context-aware deep learning for brain tumor segmentation, overall survival prediction, and subtype classification (mMRI). To acquire tumor segmentation, the authors first present a 3D context-aware deep learning method that takes into account the ambivalence of tumor placement in radiology mMRI imaging sub-regions. To achieve tumor subtype categorization, they use a normal 3D CNN on the tumor segments. Finally, hybrid DL and ML strategy are used to predict survival. The findings imply that the suggested method is capable of accurate segmentation of tumors and prediction of survival.

S. K. Baranwal et al [20] suggested a cascade Convolutional Neural Network (C-CNN) to achieve a flexible and efficient segmentation of the brain tumor system. In two independent ways, the C-CNN model collects both global and local features. In comparison to present-day models, an ideal Distance-Wise Attention (DWA) process is also developed to increase brain tumor segmentation accuracy. The DWA mechanism takes into consideration the impact of the tumor's central location and the brain inside the model. The proposed technique achieves 0.9203, 0.8726, and 0.9113 mean whole tumors, dice scores of tumor core, and embellished tumor, respectively.

K. N. Guy-Fernand et al. [21] a computerized fuzzy neighborhood learning-based 3D segmentation strategy for the identification of cerebrum cancers in 3D pictures has been presented. This is deeply interwoven with the suggested design in this method. With a 0.85 dice coefficient and 0.74 Jaccard index, the simulation results

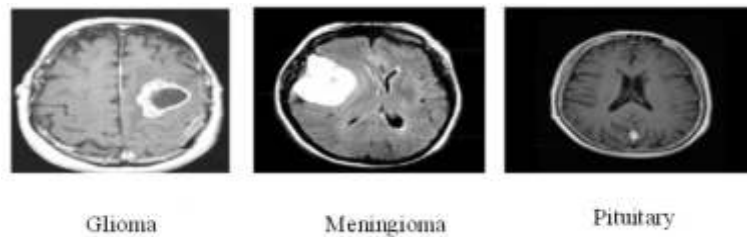


Fig. 2.1: Three different types of Brain Tumor Classification

demonstrate that the suggested brain tumor detection strategy outperforms previous methods in brain tumor diagnosis.

F. P. Polly et al [22] refined the advanced DeepSeg as a decoupling framework that is modular. It is made up of two core sections that are linked by a relationship of encoding and decoding. The encoder that extracts structural information consists of a CNN. The full-resolution probability map is created by inserting the obtained semantic map into the decoder component. The resultant segmentation outcomes have dice and Hausdorff distance scores of 0.81 to 0.84 and 9.8 to 19.7, respectively.

Saikumar et.al [23] present a fresh multi-modality deep feature learning system for segmenting brain tumors from MRI data. The main concept is to uncover intricate patterns in multi-modal data to make up for the lack of data scale. The CMFT process and the CMFF process are the two learning processes that make up the suggested cross-modality deep feature learning framework, both of which aim to learn rich features denoted by transiting and fusing insight from different modality data, respectively [24]. The suggested cross-modality DL feature model can mostly enhance the performance of brain tumor classification in comparison to conventional and cutting-edge methods [25].

2.1. Data Set. Based on previous studies the data sets are provided. It has 3064 MRI images for brain tumors. Brain Tumors are classified into mainly three types in the data set: Meningioma (708), Glioma (1426), and Pituitary (930). Each image containing pixels with the size of 512x512 was taken for evaluation. The proposed approach contains 2D and T1W MRI images which are shown in figure 2.1.

3. Methodology. The proposed framework for brain tumor classification consists of four main steps. The first step is tumor pre-processing, which involves removing unwanted noise using the Histogram of Oriented Gradients (HOG) approach and skull stripping to eliminate unwanted parts in the MRI images. In the second step, the MRI images are segmented, and different types of Classical Local Binary Patterns (CLBP) are used for feature extraction, including DLBP and LBP techniques. These techniques are then compared with existing algorithms to determine the best methodology for tumor classification. Finally, the last step is classification, where the tumor or no-tumor classification is performed on the MRI images using the selected methodology and datasets as shown in above figure 2.1.

3.1. Classical Local Binary Pattern. Classical Local Binary Pattern (LBP) is a feature extraction technique used in computer vision and image processing. It has been applied in medical image analysis for brain tumor classification. In classical LBP, each pixel in an image is compared to its surrounding pixels within a specific radius to generate a binary code. This binary code is then used to represent the texture information of the image. The binary code is obtained by comparing the grey value of the central pixel with the grey values of its neighbours. If the grey value of the neighbour is greater than or equal to the central pixel, a 1 is assigned, and if it is less than the central pixel, a 0 is assigned. This process is repeated for all the pixels in the image.

In brain tumor classification, classical LBP has been used to extract texture features from magnetic resonance images (MRI) of the brain. These features can then be used as input to machine learning algorithms for classification of brain tumors. The LBP features can capture the local texture patterns in the brain MRI and help distinguish between different types of brain tumors.

Several studies have reported high accuracy rates for brain tumor classification using classical LBP features

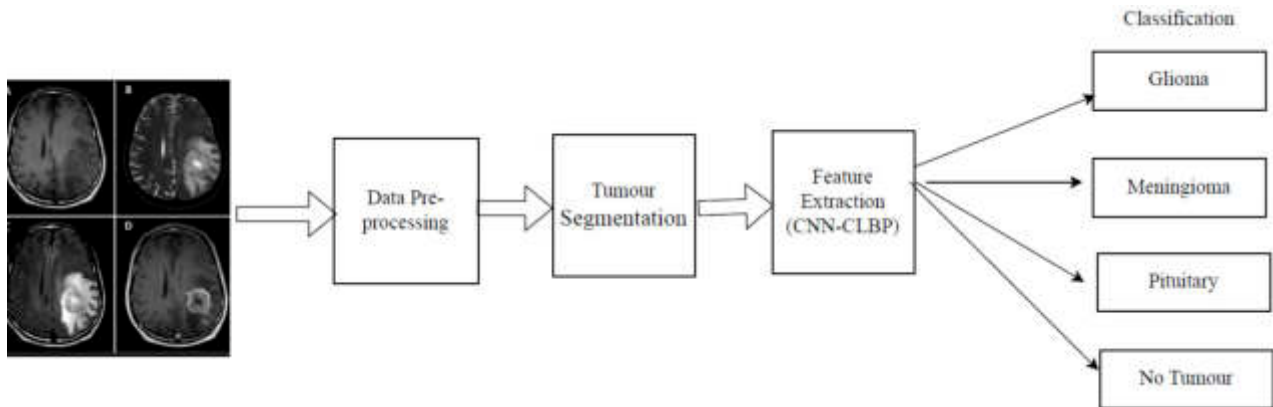


Fig. 3.1: Proposed flow diagram of brain tumour classification system.

combined with machine learning algorithms such as support vector machines (SVM) and random forest classifiers as shown in above figure 3.1. In the brain tumor classification system, the process begins with the MRI dataset presentation, followed by data preprocessing. Subsequently, the data undergoes tumor segmentation, enabling the extraction of features using CNN+CLBP. After feature extraction, the system classifies the tumors into categories such as Glioma, Meningioma, Pituitary, or identifies cases with no tumor. An illustrative result from this process is then presented. The use of classical LBP in brain tumor classification is an active area of research and is expected to continue to play an important role in medical image analysis.

3.2. Distance Based Local Binary Pattern (DLBP). Distance Classical Local Binary Pattern (DLBP) is an extension of the Classical LBP algorithm for texture analysis and classification. DLBP is used to extract features from images and is especially useful in applications where texture information is important. The DLBP algorithm is similar to Classical LBP, but it incorporates a distance function between neighbouring pixels to generate more robust and discriminative features. In DLBP, the neighbouring pixels are considered to be circularly arranged around the central pixel. The distance function is used to calculate the distance between each pair of neighbouring pixels and then used to adjust the weighting of the LBP code. These results in more robust features that can better distinguish between different textures.

In addition, DLBP uses a multi-resolution approach to capture texture features at different scales. The image is first filtered using a Gaussian filter to smooth the texture and reduce noise. Then, the filtered image is divided into multiple scales, and DLBP is applied to each scale independently.

The features extracted from each scale are then combined to generate a final feature vector. DLBP has been successfully used in a variety of applications, including face recognition, texture classification, and medical image analysis. In medical image analysis, DLBP has been used for the detection and classification of various types of lesions in images, including breast cancer, lung cancer, and brain tumors.

3.3. Angle Based Local Binary Pattern (Θ LBP). Θ LBP stands for Angle Based Local Binary Pattern, which is a texture analysis algorithm used for feature extraction from digital images. It is a variant of Local Binary Pattern (LBP) and is specifically designed to capture fine-grained texture information that may be difficult to extract using traditional LBP or other texture analysis techniques. In Θ LBP, the image is first transformed into a gradient map, where each pixel represents the magnitude and direction of the gradient of the image intensity at that location. Θ LBP then partitions the gradient map into a set of non-overlapping cells, and calculates a binary code for each cell based on the local edge patterns within that cell. The binary code is generated by comparing the intensity values of each pixel within the cell with a threshold value, which is determined based on the local mean and variance of the pixel intensities. The comparison generates a binary value (0 or 1) for each pixel, which is then concatenated to form the binary code for the cell.

The binary codes for all cells are then concatenated to form the final feature vector for the image. Θ LBP is able to capture fine-grained texture information by using adaptive thresholding based on the local statistics of

the image, which makes it robust to variations in image brightness and contrast. Θ LBP has been successfully used in a variety of applications, including face recognition, object recognition, and medical image analysis. In medical image analysis, Θ LBP has been used for the detection and classification of various types of lesions in images, including lung nodules, breast masses, and brain tumors.

3.4. Proposed CLBP Techniques with CNN. The CLBP+CNN approach for brain tumor classification is a combination of two techniques: Classical Local Binary Patterns (CLBP) and Convolutional Neural Networks (CNN). CLBP is a texture-based feature extraction method that encodes the relationship between the center pixel and its surrounding pixels in a binary code. On the other hand, CNN is a deep learning technique that automatically learns relevant features from the input images.

In this Proposed approach, CLBP is used to extract texture features from MRI images of brain tumors. These features are then fed into a CNN for classification. The CNN consists of multiple layers of convolutional and pooling operations that learn hierarchical representations of the input data. The final layer of the CNN is a fully connected layer that maps the learned features to the class labels (tumor or no tumor).

The CLBP+CNN approach has been shown to achieve high accuracy in brain tumor classification. This approach is particularly effective in detecting small and irregularly shaped tumors that may be difficult to detect using traditional feature extraction methods. Additionally, the use of CNNs allows for the automatic learning of relevant features, reducing the need for manual feature engineering. Overall, the CLBP+CNN approach represents a promising direction for brain tumor classification and has the potential to improve the accuracy and efficiency of diagnosis.

The Θ LBP method is a texture feature extraction technique that captures local patterns of pixel intensities in an image using a set of circularly symmetric neighbourhoods with different radii. The value of the radius parameter, denoted by "D" in the method, determines the size of the neighbourhoods and affects the sensitivity of the method to changes in texture. To visualize the effect of different values of D on the texture patterns captured by Θ LBP, we can plot histograms of the Θ LBP codes obtained from an image using different values of D. The histograms show the frequency of occurrence of each Θ LBP code in the image, with higher peaks indicating more dominant texture patterns. The histograms of Θ LBP codes obtained from an MRI brain image using four different values of D: 1, 2, 3, and 4. As the value of D increases, the size of the neighbourhoods used to compute the Θ LBP codes also increases, resulting in a coarser texture representation with fewer, more dominant patterns. For example, the histogram for D=1 shows a higher diversity of Θ LBP codes with lower frequencies, while the histogram for D=4 shows a smaller number of dominant Θ LBP codes with higher frequencies.

These histograms can be used to compare the texture patterns captured by Θ LBP using different values of D and to select the most appropriate value for a given texture analysis task. It appears that the table provided is showing the accuracy percentages for various studies on brain tumor classification using different methods.

Novel contributions establish this research differ from brain tumor categorization studies. It first compares feature extraction methods like Circular Local Binary Patterns (CLBPs), Directional Local Binary Patterns (DLBP), and Angular Local Binary Patterns (LBP) with brain tumor classification algorithms. This detailed study illuminates the pros and cons of different methods. The article presents a hybrid approach to brain tumor classification using Classical Local Binary Patterns (CLBP) and Convolutional Neural Networks (CNN), which has not been thoroughly studied. Compared to previous methods, this hybrid framework is more accurate and successful. The Distance Based Local Binary Patterns (DLBP) approach is also examined, stressing the distance parameter (D) in feature extraction & classification performance. This complex study demonstrates that DLBP may improve brain tumor classification.

The study also evaluates precision, recall, and F1 score to examine the proposed approaches' efficacy beyond classification accuracy. This research uses unique feature extraction methods, hybrid frameworks, and extensive performance evaluation to classify brain tumors, improving the field.

3.5. Evolution Metrics. The evaluation of a model's performance is crucial to determine its effectiveness. This evaluation is usually done using metrics such as accuracy, precision, recall, and F1 score.

Accuracy measures the percentage of correctly predicted instances out of the total number of instances. Precision measures the proportion of true positives (correctly predicted positive instances) among the total predicted positives. Recall measures the proportion of true positives among the total actual positives. F1 score is a harmonic mean of precision and recall, and is often used as a single metric to evaluate a model's

Table 3.1: Success rate with various DLT for LBP

Θ	Features	ANN	AIDE	LDA	CLBP+CNN
15	256	83.04	88.08	88.70	95.04
45	256	85.11	86.75	88.90	96.00
90	256	83.74	88.80	89.81	95.14
120	256	84.01	85.15	88.01	94.78

overall performance. These metrics provide a quantitative assessment of a model's performance, and can help in determining the model's strengths and weaknesses. By comparing the performance of different models using these metrics, researchers can identify the most effective approach for a given task.

The performance metrics are as follows:

Accuracy: The total accuracy of the model's predictions is measured by accuracy.

Precision: Precision is the ability of a model to correctly identify positive examples among all cases that were expected to be positive.

Recall: The model's capacity to accurately identify every occurrence of positivity is measured by recall.

F Measure: F1 score is the combination of both precision and recall which gives a fair evaluation of the model's performance.

- Accuracy = $(TP + TN) / (TP + TN + FP + FN)$
- Precision = $TP / (TP + FP)$
- Recall = $TP / (TP + FN)$
- F1 score = $2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$

where:

TP: True Positive (number of correctly classified positive samples)

TN: True Negative (number of correctly classified negative samples)

FP: False Positive (number of incorrectly classified positive samples)

FN: False Negative (number of incorrectly classified negative samples)

The authors utilized various evaluation metrics, but their significance remains somewhat ambiguous. Deep learning methodologies in LBP rely heavily on dataset specifics, network architecture, and chosen evaluation criteria. As depicted in Table 3.1, these techniques consistently yield promising results across diverse image classification tasks, including the classification of brain tumors from MRI imagery. Commonly employed metrics for evaluating deep learning models encompass accuracy, precision, recall, and F1 score, ensuring comprehensive performance assessment.

Across different models, accuracy percentages range from 82% to 96.00%. Notably, the CLBP+CNN model proposed by the authors achieved the highest accuracy at 95.66%. Their LBP + Knn approach yielded an accuracy rate of 90.57%, with nLBP + Knn and LBP + Knn achieving rates of 93.28% and 90.57%, respectively.

The success rates of various deep learning techniques for LBP as it requires specific information on the dataset, the network architecture, and the evaluation metrics used. However, in general, in Table 3.1 deep learning techniques have shown promising results in various image classification tasks, including brain tumor classification using MRI images. It is common to use metrics such as accuracy, precision, recall, and F1 score to evaluate the performance of deep learning models.

The accuracy percentages range from 82 % to 96.00 %. The highest accuracy percentage was achieved by the proposed CLBP +CNN model, achieving a 95.66 % accuracy rate. The LBP + Knn method used by the authors of the current paper achieved an accuracy rate of 90.57 % with nLBP + Knn at 93.28 % and nLBP Knn at 90.57 % which were shown in figure 3.2.

The DLBP (Distance Based Local Binary Patterns) method Table 3.3 has shown promising results in recognizing brain tumor types. The effectiveness of this method depends on the distance parameter (D), which determines the size of the neighborhood around each pixel. By varying the value of d, different features can be extracted from the images. These features are then used to train a classifier, such as a Convolutional Neural

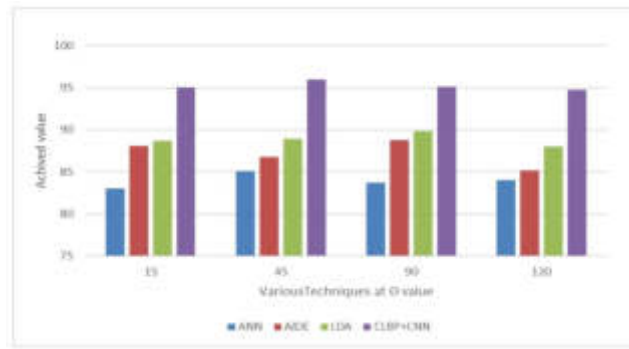


Fig. 3.2: Brain Tumor evolution matrix analysis

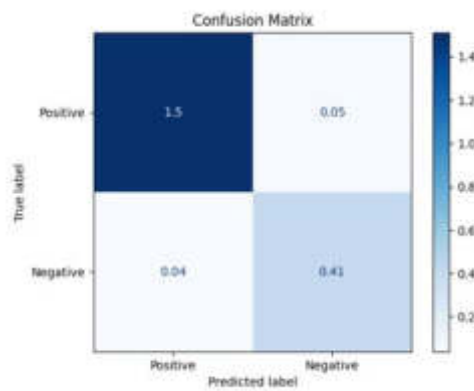


Fig. 3.3: Confusion matrix

Network (CNN), to distinguish between different types of brain tumors. The success of the classification method is measured using metrics such as accuracy, precision, recall, and F1 score. The DLBP method has been shown to outperform other feature extraction methods in some studies, indicating its potential for improving the accuracy of brain tumor classification. Further research is needed to determine the optimal value of d and to explore the potential of other feature extraction methods for brain tumor analysis shown in figure 3.3.

Table 3.2 introduces the DLBP (Distance Based Local Binary Patterns) method, exhibiting promising results in recognizing brain tumor types. The effectiveness of this method hinges on the distance parameter (D), which determines the neighborhood size surrounding each pixel. By adjusting D , unique features can be extracted and employed to train classifiers like Convolutional Neural Networks (CNNs) for discriminating between brain tumor types. Evaluation metrics such as accuracy, precision, recall, and F1 score serve as benchmarks for assessing the success of classification methods. DLBP has demonstrated superior performance compared to other feature extraction techniques in certain studies, suggesting its potential for improving brain tumor classification accuracy. Further exploration is necessary to determine the optimal value of D and investigate alternative feature extraction methods for brain tumor analysis.

Additionally, this study contrasts Circular Local Binary Patterns (CLBPs), encompassing Directional Local Binary Patterns (DLBP) and Angular Local Binary Patterns (LBP), against existing algorithms for brain tumor classification. Various classification methods, including Artificial Neural Network (ANN), Adaptive Boosting (AIDE), and Linear Discriminant Analysis (LDA), were employed for evaluation. Results indicated the proposed methodology achieving a notable success rate of approximately 95.6% using the aforementioned classification

Table 3.2: Performance Metrics with CLBP features

Class	TP	TN	FP	FN	Accuracy	Precision	Recall	F1 Score
1	1.5	0.4	0.04	0.04	95.9	97.4	97.4	97.4
2	1.52	0.41	0.05	0.05	95	96.81	96.81	96.81
3	1.51	0.42	0.06	0.03	95.5	96.17	98.08	97.11
Avg	1.51	0.41	0.05	0.04	96%	96%	97%	96%

Table 3.3: Comparative analysis from various past classifiers and its performance

Author	Brain Tumor	Classifier	Accuracy
Method [25]	Meningioma	GLCM-CNN	82 %
Method [24]	Glioma	SVM	91 %
Method [22]	Pituitary	DWT	92.66
Method [15]	Glioma	FCm-SVM	91.4 %
Method [14]	Meningioma	PCM-RELM	94.23 %
Method [1]	Meningioma	LBP-KNN	95.16 %

techniques with feature extractions from DLBP, LBP, and CLBP. This underscores the significant potential of CLBPs, particularly in enhancing brain tumor classification accuracy, thereby aiding in early detection and treatment.

The above statement highlights the use of different feature extraction methods such as Circular Local Binary Patterns (CLBPs), including Directional Local Binary Patterns (DLBP) and Angular Local Binary Patterns (LBP), which were proposed and compared with existing algorithms for brain tumor classification. The evaluation was carried out using different classification methods such as Artificial Neural Network (ANN), Adaptive Boosting (A1DE), and Linear Discriminant Analysis (LDA). The results showed that the proposed methodology achieved a high success rate of approximately 95.6 % using the mentioned classification methods with feature extractions obtained from DLBP, LBP, and CLBP. This suggests that the use of CLBPs, in particular, can significantly improve the accuracy of brain tumor classification, thereby aiding in early detection and treatment of brain tumors.

4. Results. The LBP operator is a variant of the Local Binary Pattern (LBP) operator that considers the orientation of the edges in the image. To create histograms of images using LBP, different values of the distance parameter (D) can be used. For each value of D , the LBP operator is applied to the image, resulting in a binary pattern image where each pixel is represented by a binary value based on the comparison of its intensity with its neighboring pixels. These binary patterns are then used to create histograms, which represent the distribution of the binary patterns in the image.

The histograms obtained using different values of D can be compared to evaluate the effectiveness of the LBP operator for feature extraction in image classification tasks. The choice of the optimal value of D depends on the specific application and the characteristics of the images being analyzed. The histogram of LBP features for a particular tumor type is a graphical representation of the distribution of the LBP features for that tumor type. Each bin in the histogram represents a range of LBP feature values, and the height of the bin represents the frequency of feature values falling within that range. For example, the histogram of LBP features for glioma, meningioma, and pituitary tumors with $D=1$ might show that certain ranges of LBP feature values are more common for one tumor type than the others. The shape and distribution of the histogram can provide insights into the characteristics of the tumors and help in their classification. In Fig 4.1 Local Binary Patterns (LBP) is a popular method for extracting texture features from images. The LBP variant of LBP considers the relationship between the pixel values at the center and surrounding pixels at a given angle. The value of LBP determines the angle at which the surrounding pixels are considered relative to the center pixel. In the case of glioma, meningioma, and pituitary MRI images, histograms of LBP features can be formed with $D=1$ and LBP values of 900 and 450. By varying the value of LBP, we can capture different aspects of the texture patterns

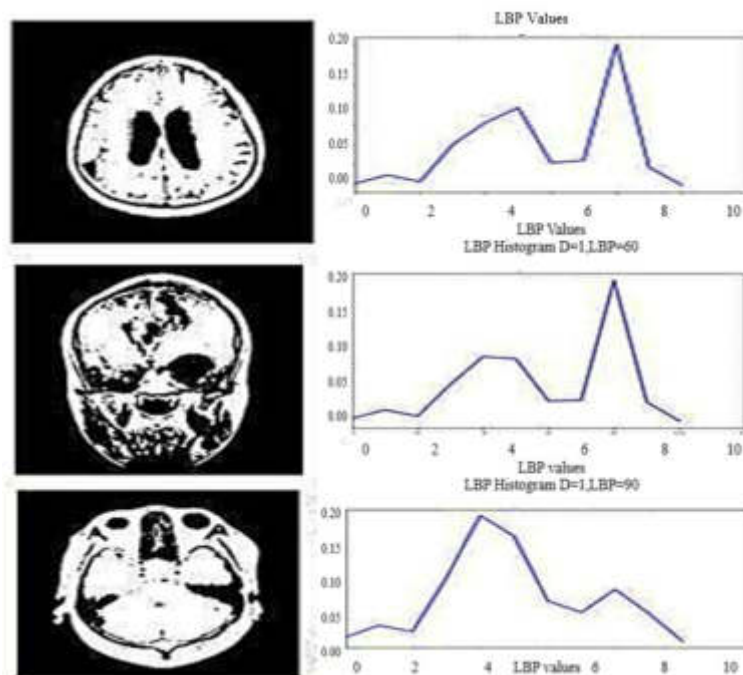


Fig. 4.1: Histograms of Glioma, Meningioma, and Pituitary MRI images are formed with $D=1$ with $LBP=90$

in the images. Figures 4 and 5 compare the effectiveness of $LBP=900$ and $LBP=450$, we can use metrics such as accuracy, precision, recall, and F1 score. These metrics can be obtained by training a classifier (such as a CNN) on the extracted features and evaluating its performance on a test set of labeled images. It is worth noting that the choice of feature extraction method and value of LBP depends on the specific problem at hand and may require experimentation to find the optimal parameters. Histograms of MRI images can be formed using LBP by first dividing the image into small, non-overlapping cells. Within each cell, LBP is applied with a specific value of D .

The resulting LBP codes are then counted and a histogram is formed for each cell, with the bin counts representing the frequency of each LBP code. In this case, histograms of Glioma, Meningioma, and Pituitary MRI images are formed with a value of $D=1$ and $LBP=450$. This means that LBP is applied using a circular neighborhood of radius 1 and 450 different rotation invariant patterns are considered. The resulting histograms can be used as features for classification algorithms, such as CNN, SVM, or random forests, to classify the MRI images into different tumor types. The choice of D and LBP parameters can affect the discriminative power of the histograms and thus the classification performance. As the value of the d parameter increases, the patterns become more global and less local. This means that smaller details in the image are being ignored, which can lead to loss of information. Therefore, it is important to choose an appropriate value of d depending on the task at hand and the nature of the images being analyzed.

5. Conclusion. In conclusion, this article presented a proposed framework for detecting and classifying brain tumors using Deep Learning Techniques (DLT) on MRI datasets. The framework consists of four main steps, including pre-processing, segmentation, feature extraction, and classification. Pre-processing was carried out using the Histogram of Oriented Gradients (HOG) method to eliminate unwanted noise and skull stripping. Different feature extraction methods such as CLBPs (DLBP, LBP) were proposed and compared with existing algorithms. The proposed methodology achieved a high success rate of approximately 95.6% using ANN, A1DE, and LDA classification methods with feature extractions obtained from DLBP, LBP, and CLBP. This article shows that the proposed framework has the potential to aid in the accurate and efficient diagnosis of brain

tumors.

There are several potential future extensions to the work presented in this article. Histograms of Glioma, Meningioma, and Pituitary MRI images are formed with a value of $D=1$ and $LBP=450$. This means that LBP is applied using a circular neighborhood of radius 1 and 450 different rotation invariant patterns are considered another possible future extension is to incorporate more advanced pre-processing techniques such as image registration or normalization to improve the accuracy of the tumor segmentation. Furthermore, the proposed framework could be evaluated on a larger dataset to validate its generalizability and performance. Finally, the framework could be extended to incorporate other types of brain tumors, such as acoustic neuromas or metastatic brain tumors, to improve its clinical relevance and utility.

REFERENCES

- [1] KAPLAN KAPLAN, YLMAZ KAYA, MELIH KUNCAN, & H. METIN ERTUNÇ, *Brain tumor classification using modified local binary patterns (LBP) feature extraction methods*, *Medical Hypotheses*, Volume 139, 2020, 109696, ISSN 0306-9877, <https://doi.org/10.1016/j.mehy.2020.109696>
- [2] FATİH ÖZYURT, ESER SERT, & DERVA AVC, *An expert system for brain tumor detection: Fuzzy C-means with super resolution and convolutional neural network with extreme learning machine*, *Medical Hypotheses*, Volume 134, 2020, 109433, ISSN 0306-9877, <https://doi.org/10.1016/j.mehy.2019.109433>.
- [3] A. GUMAEI, M. M. HASSAN, M. R. HASSAN, A. ALELAIWI & G. FORTINO, *A Hybrid Feature Extraction Method With Regularized Extreme Learning Machine for Brain Tumor Classification* in *IEEE Access*, vol. 7, pp. 36266-36273, 2019, doi: 10.1109/ACCESS.2019.2904145.
- [4] A. SEK HAR, S. BISWAS, R. HAZRA, A. K. SUNANIYA, A. MUKHERJEE & L. YANG., *Brain Tumor Classification Using Fine-Tuned GoogLeNet Features and Machine Learning Algorithms: IoMT Enabled CAD System* in *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 3, pp. 983-991, March 2022, doi: 10.1109/JBHI.2021.3100758.
- [5] M. V. S. RAMPRASAD, M. Z. U. RAHMAN & M. D. BAYLEYEGN, *A Deep Probabilistic Sensing and Learning Model for Brain Tumor Classification With Fusion-Net and HFCMIK Segmentation*. in *IEEE Open Journal of Engineering in Medicine and Biology*, vol. 3, pp. 178-188, 2022, doi: 10.1109/OJEMB.2022.3217186.
- [6] T. ZHOU, S. CANU, P. VERA & S. RUAN, *Latent Correlation Representation Learning for Brain Tumor Segmentation With Missing MRI Modalities* in *IEEE Transactions on Image Processing*, vol. 30, pp. 4263-4274, 2021, doi: 10.1109/TIP.2021.3070752.
- [7] SACHDEVA, J., KUMAR, V., & GUPTA, I., *et al. Segmentation, Feature Extraction, and Multiclass Brain Tumor Classification* *J Digit Imaging* 26, 11411150 (2013). <https://doi.org/10.1007/s10278-013-9600-0>.
- [8] R. CRISTIN, K. S. KUMAR & P. ANBHAZHAGAN., *Severity Level Classification of Brain Tumor based on MRI Images using Fractional-Chicken Swarm Optimization Algorithm* in *The Computer Journal*, vol. 64, no. 10, pp. 1514-1530, June 2021, doi: 10.1093/comjnl/bxab057.
- [9] H. H. SULTAN, N. M. SALEM & W. AL-ATABANY, *Multi-Classification of Brain Tumor Images Using Deep Neural Network* in *IEEE Access*, vol. 7, pp. 69215-69225, 2019, doi: 10.1109/ACCESS.2019.2919122.
- [10] N. M. DIPU, S. A. SHOHAN & K. M. A. SALAM, *Deep Learning Based Brain Tumor Detection and Classification* 2021 International Conference on Intelligent Technologies (CONIT), Hubli, India, 2021, pp. 1-6, doi: 10.1109/CONIT51480.2021.9498384.
- [11] P. AFSHAR, K. N. PLATANIOTIS & A. MOHAMMADI, *BoostCaps: A Boosted Capsule Network for Brain Tumor Classification* 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 2020, pp. 1075-1079, doi: 10.1109/EMBC44109.2020.9175922.
- [12] M. GURBIN, M. LASCU & D. LASCU, *Tumor Detection and Classification of MRI Brain Image using Different Wavelet Transforms and Support Vector Machines* 2019 42nd International Conference on Telecommunications and Signal Processing (TSP), Budapest, Hungary, 2019, pp. 505-508, doi: 10.1109/TSP.2019.8769040.
- [13] N. ÇNAR, B. KAYA & M. KAYA, *Comparison of deep learning models for brain tumor classification using MRI images* 2022 International Conference on Decision Aid Sciences and Applications (DASA), Chiangrai, Thailand, 2022, pp. 1382-1385, doi: 10.1109/DASA54658.2022.9765250.
- [14] S. ARORA & M. SHARMA, *Deep Learning for Brain Tumor Classification from MRI Images* 2021 Sixth International Conference on Image Information Processing (ICIIP), Shimla, India, 2021, pp. 409-412, doi: 10.1109/ICIIP53038.2021.9702609.
- [15] THEJASWINI P. BHAT, BHAVYA; & PRAKASH, KUSHAL, *Detection and Classification of Tumour in Brain MRI*, *Int. J. Eng. Manufact* (IJEM) Volume 9, Issue 1, Pages 11 20.
- [16] R. M. PRAKASH & R. S. S. KUMARI, *Classification of MR Brain Images for Detection of Tumor with Transfer Learning from Pre-trained CNN Models* 2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), Chennai, India, 2019, pp. 508-511, doi: 10.1109/WiSPNET45539.2019.9032811.
- [17] M. NAZIR, M. A. KHAN, T. SABA & A. REHMAN, *Brain Tumor Detection from MRI images using Multi-level Wavelets* 2019 International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 2019, pp. 1-5, doi: 10.1109/ICCISci.2019.8716413.
- [18] Y. BHANOTHU, A. KAMALAKANNAN & G. RAJAMANICKAM, *Detection and Classification of Brain Tumor in MRI Images using Deep Convolutional Network* 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 248-252, doi: 10.1109/ICACCS48705.2020.9074375.

- [19] H. UCUZAL, . YAAR & C. ÇOLAK, *Classification of brain tumor types by deep learning with convolutional neural network on magnetic resonance images using a developed web-based interface* 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 2019, pp. 1-5, doi: 10.1109/ISMSIT.2019.8932761.
- [20] S. K. BARANWAL, K. JAISWAL, K. VAIBHAV, A. KUMAR & R. SRIKANTASWAMY, *Performance analysis of Brain Tumour Image Classification using CNN and SVM* 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2020, pp. 537-542, doi: 10.1109/ICIRCA48905.2020.9183023.
- [21] K. N. GUY-FERNAND, J. ZHAO, F. M. SABUNI & J. WANG, *Classification of Brain Tumor Leveraging Goal-Driven Visual Attention with the Support of Transfer Learning*, 2020 Information Communication Technologies Conference (ICTC), Nanjing, China, 2020, pp. 328-332, doi: 10.1109/ICTC49638.2020.9123249.
- [22] F. P. POLLY, S. K. SHIL, M. A. HOSSAIN, A. AYMAN & Y. M. JANG, *Detection and classification of HGG and LGG brain tumor using machine learning* 2018 International Conference on Information Networking (ICOIN), Chiang Mai, Thailand, 2018, pp. 813-817, doi: 10.1109/ICOIN.2018.8343231.
- [23] MURTHY, A., RANI, P. J., ALMAKASSEES, S. M., SAIKUMAR, K., SALEH, M., & ETTYEM, S. A., *A novel classification model for high accuracy detection of Indian currency using image feature extraction process. In AIP Conference Proceedings (Vol. 2845, No. 1). 2023. AIP Publishing. <https://doi.org/10.1063/5.0170991>*
- [24] SAIKUMAR, K., & RAJESH, V., *A Deep Convolutional Neural Network-Based Heart Diagnosis for Smart Healthcare Applications. In Artificial Intelligence for Smart Healthcare (pp. 227-243). 2023, Cham Springer International Publishing. https://doi.org/10.1007/978-3-031-23602-0_14*
- [25] SRINIVAS RAO, K., DIVAKARA RAO, D. V., PATEL, I., SAIKUMAR, K., & VIJENDRA BABU, D., *Automatic prediction and identification of smart women safety wearable device using Dc-RFO-IoT. Journal of Information Technology Management, 15(Special Issue), 34-51. 2023. doi :10.22059/JITM.2022.89410*

Edited by: Dhilip Kumar V

Special issue on: Unleashing the power of Edge AI for Scalable Image and Video Processing

Received: Nov 23, 2024

Accepted: Jun 24, 2024



A NOVEL HYBRID MODEL TO DETECT AND CLASSIFY ARRHYTHMIA USING ECG AND BIO-SIGNALS

MANJESH B N* AND RAJA PRAVEEN N†

Abstract. In general, arrhythmias, also called cardiac arrhythmia, heart arrhythmia, or dysrhythmias, are abnormal heartbeats which include too fast or too slow. The Cardiovascular Disease (CVD) is a significant cause of death, and the death rate is increasing every year. The Electrocardiogram (ECG) majorly contributes to the CVD diagnosis, providing information about the heartbeat. An automatic detection and classification of arrhythmia performs a significant role in managing and curing cardiovascular diseases. Deep Learning (DL)-based algorithms have emerged as effective solutions in medical applications, particularly in cardiac arrhythmia diagnosis. In this research, a DL-based multi-modal approach is proposed for the classification of cardiac arrhythmia. The MIT-BIH dataset is utilized to evaluate the performance of the proposed method. The proposed method considers physiological signals along with the MIT-BIH dataset to improve accuracy. The Discrete Wavelet Transform (DWT) is used for pre-processing the MIT-BIH dataset. The DL methods of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) are utilized for classifying cardiac arrhythmia. The proposed method is evaluated using various performance metrics such as Accuracy, Specificity, Sensitivity, F1-score, and Cohens kappa.

Key words: Cardiac arrhythmia, Deep Learning, Electrocardiogram, Long Short-Term Memory, Recurrent Neural Network

1. Introduction. In the healthcare sector, technology has significantly impacted patient care, treatment, and diagnosis. Technological innovations have had a major impact on medical imaging and surgical operations, improving patient outcomes and increasing process efficiency. Examples of these innovations include the creation of minimally invasive surgical techniques and the X-ray. We are about to witness a momentous shift in which data and artificial intelligence (AI) could completely reshape a number of healthcare domains [1].

Arrhythmias, which refer to irregularities in the heart's rhythm, have been a significant focus in the field of cardiovascular medicine for a considerable period of time. The accurate diagnosis of these conditions is crucial due to their various manifestations, ranging from harmless occasional skips to potentially life-threatening situations. The electrocardiogram (ECG) has traditionally been the preferred method for detecting and classifying arrhythmias [3].

The interpretation of this graphical representation of the heart's electrical activity necessitates expertise, including a discerning eye, comprehensive training, and experience. Deep learning, a subset of artificial intelligence (AI) that utilizes deep neural networks, has proven its efficacy in multiple industries by effectively extracting valuable patterns and insights from large datasets. These algorithms have the potential to improve the efficiency, accuracy, and timeliness of arrhythmia diagnosis in ECG analysis [4].

This paper explores the integration of deep learning techniques with arrhythmia classification. This study will explore the progress made in the interdisciplinary field, focusing on the essential deep learning architectures. We will analyze their advantages, drawbacks, and performance metrics. The review will address the challenges of data quality, model transparency, and ethical considerations in AI-driven healthcare. This study aims to provide a comprehensive overview of the current state and future possibilities in the intersection of cardiology, data science, and artificial intelligence. Cardiologists have developed expertise in interpreting electrocardiograms (ECGs) to distinguish between benign and malignant cardiac rhythms. Even with extensive training, the human eye may occasionally fail to detect or accurately interpret subtle anomalies. Furthermore, the increasing amount of electrocardiogram (ECG) data, particularly due to the rise of wearable technology, makes it impractical to manually analyze each individual heartbeat. Smartwatches with miniaturized ECG modules have made health

*Jain University (manjeshbn@gmail.com).

†Jain University (p.raja@jainuniversity.ac.in).

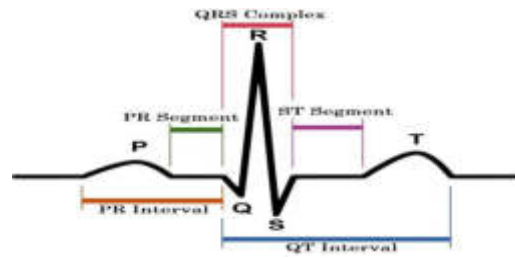


Fig. 1.1: ECG waveform depiction.

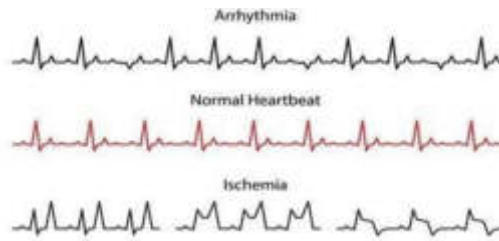


Fig. 1.2: Representation of Arrhythmia waveforms in comparison to Normal Heartbeat.

monitoring more accessible to the general population. However, the main issue remains: how can we effectively and precisely handle this overwhelming amount of data? [6]

Deep-Learning methods play a significant role in this context. Deep learning, a subfield of artificial intelligence, has experienced significant growth across multiple domains in recent decades. The system's strength lies in its capacity to analyze large volumes of data, acquiring complex patterns, and frequently surpassing human performance in certain tasks. Deep learning has the potential to not only offer computational efficiency but also demonstrate adaptability and scalability [8].

The goal of this work is to document the development of deep learning in arrhythmia classification, from its inception to the most advanced applications today. The several facets of architectures, approaches, accomplishments, and difficulties will all be covered in this study. The goal of this review is to provide a thorough understanding of how technology and healthcare specifically, cardiac care intersect. It will examine the field's technological developments, historical background, and clinical applications, emphasizing how it could transform cardiac care in the twenty-first century [12].

Like other technological advancements, the integration of DL into arrhythmia classification faces challenges. Researchers and clinicians encounter various challenges, including data privacy, diverse and representative datasets, model interpretability, and clinical validation.

The goal of this work is to document the development of deep learning in arrhythmia classification, from its inception to the most advanced applications today. The several facets of architectures, approaches, accomplishments, and difficulties will all be covered in this study. The goal of this review is to provide a thorough understanding of how technology and healthcare specifically, cardiac care intersect. It will examine the field's technological developments, historical background, and clinical applications, emphasizing how it could transform cardiac care in the twenty-first century [15].

1.1. Motivation. If left undiagnosed and untreated, arrhythmias can cause serious health issues, even potentially fatal situations. Only ECG data is used in the diagnosis process and technologies used today. As such, it ignores a number of additional physiological cues, such as heart rate and blood pressure. The combination of various physiological cues, artificial intelligence, and ECG data allows for the early identification and categorization of arrhythmias. Artificial intelligence can be used to create adaptive algorithms and real-

time monitoring due to the dynamic nature of arrhythmias. Therefore, the goal of this research is to use modern technologies for the early identification and categorization of arrhythmias. Convolution neural networks can be used to detect arrhythmias in an efficient manner because of its capacity to identify patterns and spatial hierarchies in the input data. CNN also provides the best result when the model containing diverse data set of arrhythmias are trained [16].

1.2. Major Contributions. As per the reports of the World Health Organization (WHO), the number one cause of death today is cardiovascular diseases (CVDs). As per their statistics, the number of deaths caused by CVDs are roughly 30 percent. Cardiac Arrhythmia is a condition in which the electrical activity of the heart is abnormal. The electrical activity is very irregular leading to the disruption in the cardiac rhythm. ECG data is very complex owing to different waveforms and their interpretation. With the advent of different computational paradigms, researchers have been very curious to explore the possibilities of leveraging different techniques like Machine Learning (ML), deep learning (DL) etc. to interpret the ECG like a cardiologist. It is very important to note that the accuracy if diagnosis is very critical as any deviation could be fatal [18].

Most of the conventional researchers so far have been focused around exploring the optimal computational paradigms for predicting and classifying the cardiac arrhythmia using a variety of different hardware and programming languages. Many research works range from leveraging techniques like SVM, PCA to feed-forward based neural networks and apply wavelet transform for feature extraction [20].

However, some of the downsides are a) achieving better performance without cross-validation, b) losing the beats due to filtering and feature extraction c) less number of arrhythmia type classification d) low accuracy and performance.

In our work, we novel deep learning-based framework to analyze the complex ECG data and develop a transferable representation of ECG signals. It is important to know that to realize such a framework it is very important to describe an architecture that offers scope for learning the signal representation. Once we build a model and train that model on a huge training data set, the model will be able to learn from the pattern and allow to use those representation to transfer the knowledge. We have further experimented the CNN by adding the batch normalization layer between subsequent layers thereby inhibiting the hidden / convolution layers from normalizing the values which facilitate in improving the efficiency. Also, the proposed algorithm employs a 2-D CNN with monochrome images of the ECG. One of the advantages of our approach is that conventional data pre-processing steps like feature extraction and noise removal and filtering are not required as the algorithm converts the 1-D Signal data to a 2-D image. Additionally, to improve the accuracy of the model we can augment the 2-D images and increase the size of the training data. Since our algorithm transforms a 1D signal to a 2D image, the model will automatically ignore the noise and extract the feature map. This allows the proposed model to be employed on heterogeneous signals and devices with different feature sets like sampling rate, amplitude etc. unlike the conventional models. Nonetheless, our approach can be implemented in an end-to-end clinical set up and that adds to the novelty of our work.

2. Background.

2.1. Arrhythmias: An Undeniable Obstacle. Derived from the Greek word "arrhythmias," which means "without rhythm," arrhythmias refer to a broad spectrum of illnesses marked by an irregular heartbeat. These disorders can manifest as a variety of heart rhythms, including bradycardias (slow heartbeats), atrial fibrillations (chaotic rhythm), and tachycardias (rapid heartbeats). These abnormalities might have a variety of causes, including extrinsic influences like stress or drug usage, congenital problems, and cardiomyopathies [21].

Early detection and management of arrhythmias are crucial due to the potential for severe complications, as some arrhythmias are benign while others are not. The electrocardiogram (ECG) is the primary tool used in this field. It is a non-invasive test that visually displays the heart's electrical activity [23].

2.2. Electrocardiograms (ECGs): The Diagnostic Mainstay. The introduction of the ECG in the early 20th century brought about a significant transformation in cardiac diagnostics. Clinicians can visualize the heart's electrical impulses by applying electrodes to the skin, which are then represented as waveforms on a graph. The various components of this waveform, including the P wave, QRS complex, and T wave, provide valuable information about different stages of the cardiac cycle. Variations in these waveforms and complexes

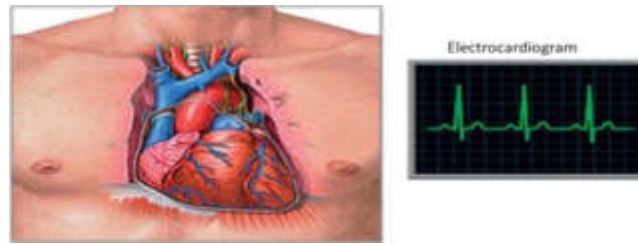


Fig. 2.1: Human Heart and ECG.

may suggest the presence of underlying arrhythmias or other cardiac pathologies. However, the analysis of electrocardiograms (ECGs) is not a simple process. Detecting subtle anomalies, such as distinguishing between a normal rhythm and a potentially dangerous arrhythmia, necessitates significant training, experience, and a discerning eye.

2.3. The Advent of Artificial Intelligence in Healthcare. The integration of technology into healthcare has significantly increased during the 21st century. The evolution of healthcare is evident through the digitization of medical records, utilization of advanced imaging modalities, and the increasing prevalence of telemedicine. Data has played a central role in this technological wave. The growing accessibility of extensive datasets has paved the way for the emergence of artificial intelligence (AI) in the field of medicine. AI, which refers to the simulation of human intelligence processes by machines, has been increasingly utilized in various medical fields. Machine learning, a subset of computer science, involves the use of statistical techniques by computers to learn from data. This has subsequently facilitated the development of more sophisticated techniques. Deep learning, a subfield of machine learning, has shown great promise as it utilizes neural networks algorithms. One of its notable strengths is its capacity to acquire knowledge and make informed choices based on data, frequently outperforming human abilities in certain tasks [26].

2.4. The Confluence of Deep Learning and ECG Analysis. Researchers started examining the combination of the two because of the difficulties involved in manually interpreting ECG data and the potential of deep learning in handling enormous volumes of data. Is it possible to train deep learning algorithms to identify arrhythmias as accurately as expert cardiologists, if not more so? Numerous studies, inventions, and discussions have resulted from the pursuit of an answer to this topic, which has set the stage for the information contained in this article.

3. Deep Learning Techniques for Arrhythmia Classification.

3.1. Convolutional Neural Networks (CNNs). Originally designed for image processing, convolutional neural networks (CNNs) have had a profound impact on fields involving the recognition of patterns in geographical data. These networks automatically identify hierarchical patterns in the data by using convolutional layers. By transforming ECG segments into time-frequency representations like spectrograms and treating them like image-like structures, CNNs can be utilized to identify arrhythmic patterns. Convolutional layers are very good at capturing the subtle fluctuations and anomalies that point to different kinds of arrhythmias because they are skilled at identifying spatial patterns in converted ECG data [29].

Multiple studies have shown that CNNs have proven to be effective in achieving high levels of accuracy in detecting arrhythmia. In some cases, CNNs have even demonstrated comparable performance to that of expert cardiologists [30].

3.2. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) Networks. Because they are made expressly to handle sequential data, recurrent neural networks or RNNs are ideal for evaluating time-series data like ECG sequences. Artificial neural networks have the ability to store data from previous calculations, which allows them to identify patterns and trends over time. LSTMs are a kind of RNN that effectively preserve patterns over lengthy sequences, hence resolving the vanishing gradient problem commonly observed in traditional RNNs [32].

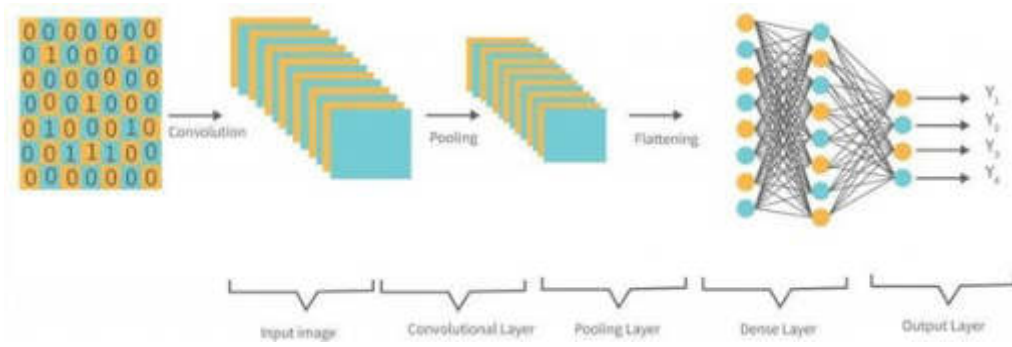


Fig. 3.1: Representation of Deep Learning CNN method.

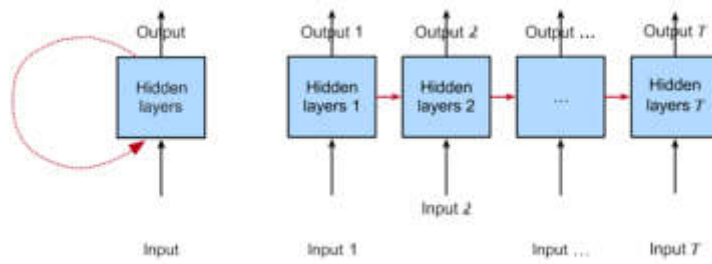


Fig. 3.2: Representation of Deep Learning CNN method.

RNNs and LSTMs are effective in analyzing raw ECG sequences due to their ability to capture temporal patterns which can serve as indicators of arrhythmias. Their capacity to identify long-term dependencies in the ECG sequence renders them highly valuable, particularly in instances where the arrhythmia’s distinctive pattern is distributed over an extended period.

3.3. Attention Mechanisms and Transformers. Attention mechanisms, derived from the domain of natural language processing, enable models to selectively concentrate on particular segments of the input data. Transformers, which rely solely on attention mechanisms, have demonstrated considerable potential across diverse domains. In the context of ECG data, attention mechanisms enable models to prioritize segments of the data that may be more indicative of an arrhythmia. By assigning weights to different components of an ECG sequence, these models have the potential to enhance accuracy by prioritizing the most pertinent signals [35].

3.4. Autoencoders. Autoencoders are a type of unsupervised neural network that aim to learn compact representations of input data. These models operate by compressing the input data into a condensed form and subsequently reconstructing the original input using this condensed representation. Autoencoders are a viable method for detecting anomalies in electrocardiogram (ECG) signals. Deviation or anomaly in the input signal, such as arrhythmic events, can lead to a high reconstruction error when trained on normal ECG data. This characteristic can be utilized to identify possible arrhythmias. Data, resulting in enhanced rates of arrhythmia detection.

Table 4.1: Overview of Arrhythmia Detection Studies

No	Paper Title	Authors	Year	Key Findings
1	Arrhythmia detection using deep convolutional neural network with long duration ECG signals	Ozal Yldrm, Pawe Pawiak	2018	This study proposes deep convolutional neural network (DCNN) to detect arrhythmia in long-duration ECG signals. The investigation of the model's ability to detect less common or subtle arrhythmias which is crucial for clinical applications is currently limited.
2	Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats	Shu Lih Oh, Eddie Y.K. Ng	2018	In this study, the authors employed the model's ability to accommodate heartbeats of varying lengths has not been thoroughly examined by means of the challenges or potential data misalignments it may pose. based on convolutional neural networks.
3	Multiclass Classification of Cardiac Arrhythmia Using Improved Feature Selection and SVM Invariants	Anam Mustaqeem, Syed Muhammad	2018	In this paper, the authors propose a novel approach for classifying cardiac arrhythmia. Although improved feature selection is suggested, it is possible that other advanced feature extraction or transformation techniques could provide a more effective representation and consequently improve classification outcomes.
4	Automated arrhythmia detection using novel hexadecimal local pattern and multilevel wavelet transform with ECG signals	Turker Tuncer, Sengul Dogan	2019	This paper reports the novel hexadecimal local pattern (HxLP) has been introduced but further research is needed to assess its robustness in the presence of noisy or artifact-laden ECG signals.
5	A Fast Machine Learning Model for ECG-Based Heartbeat Classification and Arrhythmia Detection	Miquel Alfaras, Silvia Ortín, Miguel Cornelles Soriano	2019	The research presents a machine learning model specifically designed for ECG signals with a primary focus on classifying heartbeats and detecting arrhythmias. Understanding the decision-making process of the model is crucial due to the critical nature of arrhythmia detection. The paper lacks in-depth analysis of model interpretability and the significance of the selected features.
6	Automated arrhythmia classification based on a combination network of CNN and LSTM	Chen Chen, Zhengchun Hua, Ruiqi Zhang	2020	This work presents arrhythmia classification through the use of hybrid model i.e CNN and LSTM. The model is more accurate and robust than the traditional model in terms of accuracy and robustness. The limitations of this study are that QRS detection is necessary which leads to additional computational cost. The second issue is that the data set used in this work is imbalanced.
7	Multirate Processing with Selective Sub bands and Machine Learning for Efficient Arrhythmia Classification	Saeed Mian Qaisar	2021	This study proposes a Multi-rate processing chain for the arrhythmia classification. Multi-rate processing feature selection were employed to decrease the information amount procedure thus reducing the complexity of the computational system. The performance results of model were varied by chosen various number of features.
8	Arrhythmia and Disease Classification Based on Deep Learning Techniques	Ramya G. Franklin, B. Muthukumar	2021	This work predicts converting raw ECG data to 2D pictures may cause information loss. Using straight 1D convolution on raw signals or different transformation methods may improve or accelerate results.
9	An Ensemble of Deep Learning-Based Multi-Model for ECG Heartbeats Arrhythmia Classification	Ehab Eesa, Xi-anghua Xie	2021	This research proposes multi-model system that was presented for the arrhythmia classification. It focusses on two models: CNN-LSTM to capture dynamics in temporal as well as local features for data ECG; RRHOS-LSTM that concatenated some features for classical i.e. RR intervals. This approach does not perform the feature extraction process.
10	ECG Heartbeat Classification Using Multimodal Fusion	Zeeshan Ahmad	2021	This work introduced two computationally effective multimodal feature fusion framework classification for ECG heart beat named Multimodal Image Fusion (MIF) and Multimodal feature fusion (MFF). This framework consumed much time for training and inference.

Continued on next page

Table 4.1 – continued from previous page

No	Paper Title	Authors	Year	Key Findings
11	Interpreting Arrhythmia Classification Using Deep Neural Network and CAM-Based Approach	Niken Prasasti Martono	2021	The work proposes an extension of CNN-based learning in detecting arrhythmia using recurrence plots from ECG signal and then the authors then conduct visualization using the Grad-CAM approach on the recurrence plot data to have a better interpretation of the classification process. In this work in the data preprocessing stage the appearance of R wave with irregular timing has been noted.
12	Arrhythmia Classification Techniques Using Deep Neural Network	Ali Haider Khan	2021	The research is focused on the latest trends in arrhythmia classification techniques and the system is constructed using deep neural networks. The study focused on understanding arrhythmia classification techniques to overcome their limitations. Time-series data was used by the authors to create the proposed automated system which is not applicable to different systems. For classification a balanced dataset is necessary.
13	Classification of Arrhythmia in Heartbeat Detection Using Deep Learning	Wusat Ullah	2021	The research developed a CNN model to classify ECG signals into eight categories. MIT-BIH arrhythmia database and PTB Diagnostic ECG database are used in this work. The study should concentrate on the development of denoising and data augmentation techniques.
14	Interpretation and Classification of Arrhythmia Using Deep Convolutional Network	Prateek Singh, and Ambalika Sharma	2021	The authors of this research trained a deep learning model and evaluated its classification performance Post-hoc explanation methods like SHapley Additive explanations (SHAP), local interpretable model-agnostic explanations (LIME), and Grad-CAM were used to interpret the decision rationale after interpreting the classification findings. This works lags in Interpretability.
15	Automatic cardiac arrhythmia classification based on hybrid 1-d CNN and bi-LSTM model	Jagdeep Rahul A , Lakhn Dev Sharma	2021	This work reports automatic classification system of ECG beats based on the multi-domain features derived from the ECG signals. This work reports overfitting where large dataset was used to remove.
16	Electrocardiogram based arrhythmia classification using wavelet transform with deep learning model	Shadhon Chandra Mohonta	2022	The study proposes a Deep Learning approach for the ECG based classification of the arrhythmia disease. The scalogram was acquired through the Continuous Wavelet Transform (CWT) was classified by the network based on signature according to arrhythmia. This approach was only suitable for the smaller segments of the signal
17	Inter-patient arrhythmia classification with improved deep residual convolutional neural network	Yuanlu Li	2022	The research paper presents enhanced Deep Residual Convolutional Neural Network (DRCNN) for automatic classification of arrhythmias. This approach was ability to effectively classified the arrhythmias without heartbeats extraction. This method had poor directionality as well as lack of phase information
18	A Hybrid Deep Learning Approach for ECG-Based Arrhythmia Classification	Parul Madan, Vijay Singh	2022	This work reports a hybrid model 2D-CNN-LSTM for the automation of the detection and process of classification. The dimension of the data was insufficient in classification, and the attributes has irrelevant data. This caused leads to inaccurate classification results in cardiac analysis.
19	Local-Global Temporal Fusion Network with an Attention Mechanism for Multiple and Multiclass Arrhythmia Classification	Yun Kwan Kim	2022	This research developed a new framework for an automatic classification that combined the residual network with squeeze-and-excitation (SE) block and bi-directional LSTM. This method designed to extract the features from original ECG data to acquire a unique intersubject attributes. The augmentation effect could be reduced due to baseline wander as well as noise couldbe extended to rhythm data.
20	An End-to-End Cardiac Arrhythmia Recognition Method with an Effective DenseNet Model on Imbalanced Datasets Using ECG Signal	Hadaate Ullah ,Md Belal Bin Heyat , Fajjan Akhtar	2022	This research proposed a 2D CNN Method to recognize arrhythmia from ECG automatically. This approach uses two datasets. The proposed model lags with real-time monitoring and end-to-end clinical study.

Continued on next page

Table 4.1 – continued from previous page

No	Paper Title	Authors	Year	Key Findings
21	Detection of heart arrhythmia based on UCMFB and deep learning technique	B MOHAN RAO and AMAN KUMAR	2022	This research presents Resnet50 model that classifies healthy people and patients with 4 types of cardiovascular diseases (CVD) based on ECG abnormalities. The study was done based on short and long segments of ECG databases. This has not deployed for Specific ECG multiclass classification
22	Arrhythmic Heartbeat Classification Using 2D Convolutional Neural Networks	M.Degirmenci, M.A.Ozdemir	2022	This study presents deep learning approach CNN to identify arrhythmias in ECG signals trained by two-dimensional (2D) ECG beat images. This work lags with real time monitoring.
23	Lightweight Shufflenet Based CNN for Arrhythmia Classification	HURUY TESFAI	2022	This research work proposes a lightweight Convolution Neural Network (CNN) model based on the ShuffleNet architecture targeting arrhythmia classification with a 9x reduction factor in the number of trainable parameters. In this work feature extraction capabilities on convolution block should be improved.
24	A novel automated CNN arrhythmia classifier with memory-enhanced artificial hummingbird algorithm.	Evren Kymaç, Yasin Kaya	2023	This work presents a novel method for the hyperparameter optimization (HPO) of a convolutional neural network (CNN) arrhythmia classifier using a metaheuristic (MH) algorithm. The proposed approach has not applied in different datasets for arrhythmia classification.
25	Classifying Cardiac Arrhythmia from ECG Signal Using 1DCNN Deep Learning Model	AdelA, Ahmed Waleed Ali	2023	The study proposes a deep learning model, specially convolutional neural network (1D-CNN), for the classification of arrhythmias. Limitations in this approach are dataset is imbalanced and it requires large dataset to train the model.
26	Electrocardiogram Heartbeat Classification for Arrhythmias and Myocardial Infarction	Bach-Tung Pham	2023	The research presents novel approach for ECG heartbeat classification. It uses MIT-BH and PTB datasets. The limitations of this approach is effectiveness of the model needs to be checked for additional datasets.
27	Cardiac arrhythmia detection using deep learning approach and time frequency representation of ECG signals	Yared Daniel Daydulo, Bheema Lingaiah D	2023	This research proposed an automated deep learning model capable of accurately classifying ECG signals into three categories. The model was trained on ECG data from the MIT-BIH and BIDMC database. Authors of this research concentrated on classifying ECG signals into three classes. It's better to collect more data for this work.
28	A novel deep learning approach for arrhythmia prediction on ECG classification using recurrent CNN with GWO	Prem Narayan Singh, Rajendra Prasad Mahapatra	2023	This research proposes a method called recurrent convolutional neural network (RCNN) and Grey Wolf Optimization (GWO) for predicting arrhythmia. The proposed method is evaluated by using two publicly available datasets PTB diagnostic ECG and Grey Wolf Optimization (GWO). The proposed method has been compared with other ML techniques. Improvement is needed in terms of adapting metamodel approach and identifying different arrhythmia types.
29	A novel deep neural network heartbeats classifier for heart health monitoring	Velagapudi, Swapna Sindhu, Kavuri Jaya Lakshmi	2023	This presents one-dimensional convolutional neural network (1D CNN) for the classification of heart arrhythmia. Hyperparameter tuning was not adapted in order to improve accuracy of the model.
30	Automated inter-patient arrhythmia classification with dual attention neural network	He Lyua,, Xiangkui Li b, Jian Zhang b	2023	This work presents a dual attention mechanism with hybrid network (DA-net) for arrhythmia classification. DA-net is based on modified convolutional networks with channel attention (MCC-Net) and sequence-to-sequence network with global attention (Seq2Seq). Improvement is needed in terms of adapting data augmentation techniques.

4. Comparative Study.

5. Proposed Methodology.

1. Obtain ECG datasets from various sources in order to guarantee a wide range of cardiac conditions.
2. Prepare the ECG signals for analysis by means of normalization, filtering, and segmentation.
3. Incorporate deep learning alongside conventional signal processing methods to autonomously extract pertinent features from electrocardiogram (ECG) data.

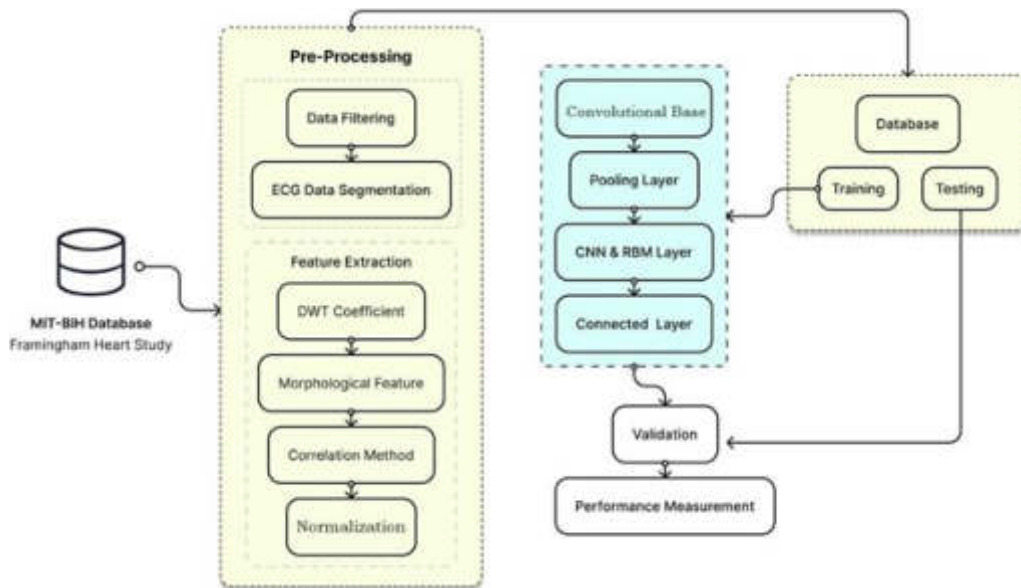


Fig. 5.1: Proposed work flow diagram for arrhythmia detection and classification.

4. Construct and evaluate a deep learning architecture that incorporates the gathered attributes to identify arrhythmias.
5. Performance can be enhanced and overfitting prevented by implementing regularization techniques and fine-tuning hyperparameters.
6. Evaluate the accuracy and robustness of the model using data obtained from multiple ECG devices.
7. Incorporate explainable AI methodologies in order to comprehend and represent the decision-making process of the model.
8. Evaluate the model on a test dataset utilizing metrics such as accuracy, sensitivity, specificity, and F1 score.
9. Develop a prototype system to demonstrate the arrhythmia detection capabilities.
10. Continuously refine the model using feedback and additional data for ongoing improvement.

6. Results and Discussion. In our work, we novel deep learning-based framework to analyze the complex ECG data and develop a transferable representation of ECG signals. It is important to know that to realize such a framework it is very important to describe an architecture that offers scope for learning the signal representation. Once we build a model and train that model on a huge training data set, the model will be able to learn from the pattern and allow to use those representation to transfer the knowledge. We have further experimented the CNN by adding the batch normalization layer between subsequent layers thereby inhibiting the hidden / convolution layers from normalizing the values which facilitate in improving the efficiency.

Also, the proposed algorithm employs a 2-D CNN with monochrome images of the ECG. One of the advantages of our approach is that conventional data pre-processing steps like feature extraction and noise removal and filtering are not required as the algorithm converts the 1-D Signal data to a 2-D image. Additionally, to improve the accuracy of the model we can augment the 2-D images and increase the size of the training data. Since our algorithm transforms a 1D signal to a 2D image, the model will automatically ignore the noise and extract the feature map. This allows the proposed model to be employed on heterogeneous signals and devices with different feature sets like sampling rate, amplitude etc. unlike the conventional models. Nonetheless, our approach can be implemented in an end-to-end clinical set up and that adds to the novelty of our work.

As shown in Figure 6.1, Individual cardiologist performance is indicated by the red crosses and averaged cardiologist performance is indicated by the green dot. The line represents the ROC curve of model performance.

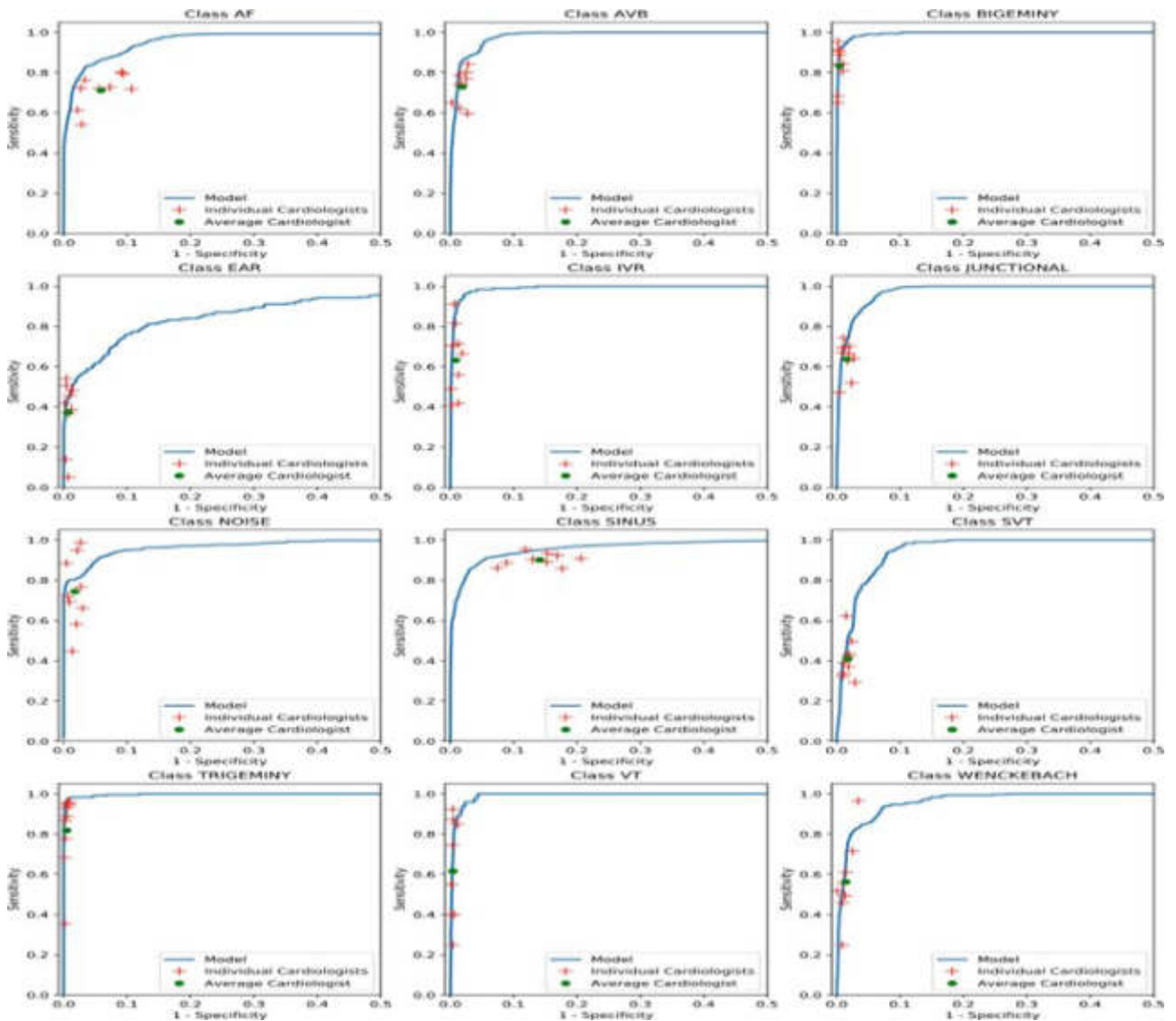


Fig. 6.1: Specificity and Sensitivity of Different classes of Arrhythmia.

AF-atrial fibrillation/atrial flutter; AVB- atrioventricular block; EAR-ectopic atrial rhythm; IVR-idioventricular rhythm; SVT supraventricular tachycardia; VT-ventricular tachycardia. $n = 7,544$ where each of the 328 30-second ECGs received 23 sequence-level predictions.

Fixing the specificity at the average specificity level achieved by cardiologists, the sensitivity of the DNN exceeded the average cardiologist sensitivity for all rhythm classes. Fixing the specificity at the average specificity level achieved by cardiologists, the sensitivity of the DNN exceeded the average cardiologist sensitivity for all rhythm classes as shown in Table 6.1. And the overall accuracy of the model (AUC) is around 0.97 as shown in Figure 6.2.

Our work demonstrates that the accuracy of the model is around 0.97. Our study demonstrates how employing Deep Learning Methods can improve the accuracy and open new avenues for research. Figure 6.2 shows the AUC of the rhythm classes and we can note that the accuracy is elevated compared to the annotations

When we started experimenting with the dataset, we realized that the data was quite imbalanced as shown

Table 6.1: Sensitivity comparison of our model vs. avg cardiologist

Condition	Specificity	Avg Cardiologist Sensitivity	Our Model's Sensitivity
Atrial fibrillation and flutter	0.941	0.71	0.861
AVB	0.981	0.731	0.858
Bigeminy	0.996	0.829	0.921
EAR	0.993	0.380	0.445
IVR	0.991	0.611	0.867
Junctional Rhythm	0.984	0.634	0.729
Noise	0.983	0.749	0.803
Sinus rhythm	0.859	0.901	0.950
SVT	0.983	0.408	0.487
Ventricular Tachycardia	0.996	0.652	0.702
Wenckebach	0.986	0.541	0.651

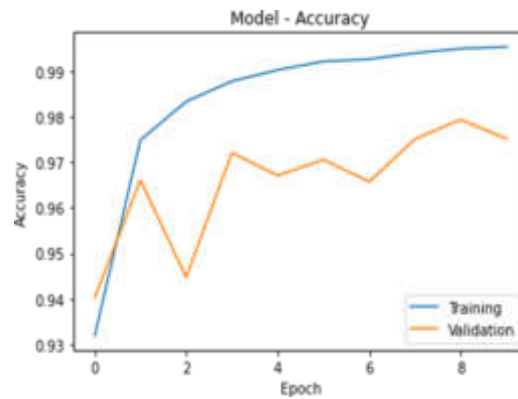


Fig. 6.2: Model Performance.

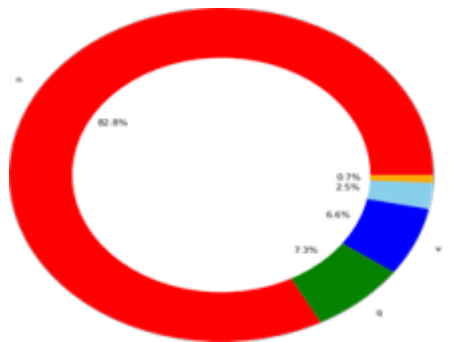


Fig. 6.3: Imbalanced Data.

in Figure 6.3 and after employing the resampling techniques we got a perfectly distributed data as shown in Figure 6.5.

Figure 6.6 as shown below, helps to visualize 1 ECG beat per category in the Time vs Amplitude format. This shows how different arrhythmic beats have different waveforms and how much do they vary from the normal beats shown in blue colour.

Figure 6.5 shows the results derived using the transformation method, where the 1-D signal data was transformed into 2-D 128 x 128 greyscale image. We have tested the classifier on 4,000 beats that we not a

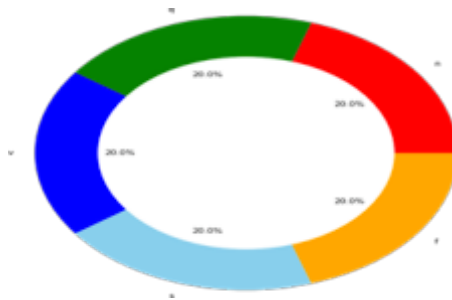


Fig. 6.4: Balanced Data after sampling.

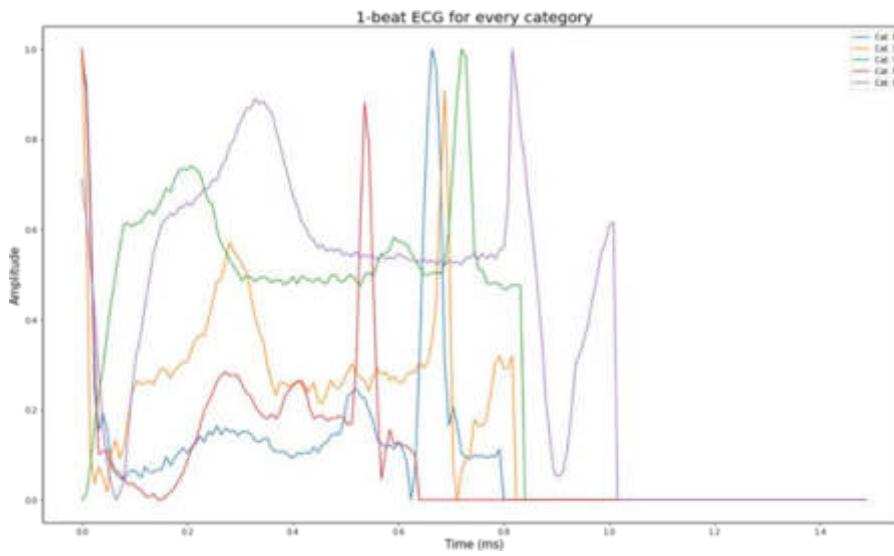


Fig. 6.5: ECG beats visualization.

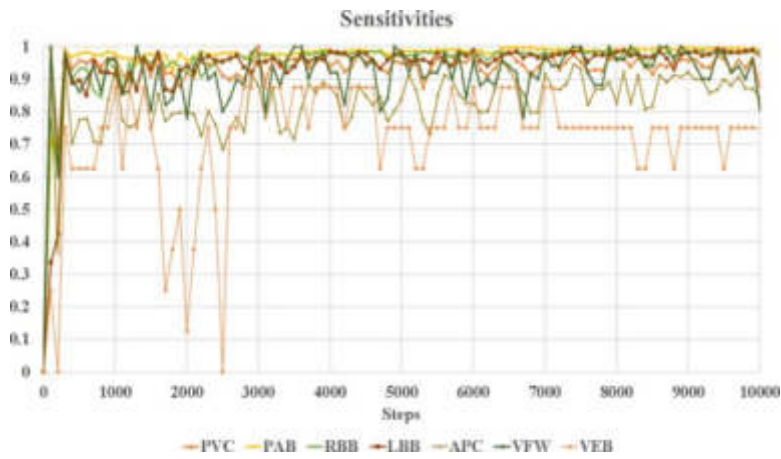


Fig. 6.6: Sensitivity of each type of arrhythmia class using transformation method.

part of training data. Figure 6.6 shows the confusion matrix of the classifier on the test set and we can infer that the model is making accurate predictions and also distinguish various arrhythmia classes.

Table 6.1 shows the avg accuracy of the proposed method. From the results we can infer that the proposed model has a very high accuracy and this has been characterized by the residual connections in the network which allows better learning in the networks compared to conventional methods.

7. Conclusion. This work develops a hybrid model for the automatic feature extraction and classification of various arrhythmias. Nevertheless, there are obstacles that need to be addressed, particularly regarding the accuracy of data, the establishment of standards, and the comprehensibility of these models. Future research should prioritize creation of models that combine features of machine learning and deep learning. These models have the potential to enhance robustness and improve generalization capabilities. Furthermore, it is crucial to develop a unified framework that can effectively integrate with current healthcare systems. In conclusion, although the use of ML and DL for arrhythmia classification is still in its early stages, it holds significant potential for improving patient care. The adoption of these technologies has the potential to significantly transform cardiac care, establishing early and precise diagnosis as the prevailing standard.

8. Future Scope. Our proposed model focusses on detecting and classifying arrhythmia using ECG and bio signals which is a game changer.in Future more focus will be on real time monitoring of patients using the required dataset, where we can achieve more accuracy.

Credit authorship contribution statement. Manjesh B N: Methodology, Software, Data curation, Writing original draft, Writing review and editing. Dr.Raja Praveen N-Investigation, Writing review and editing.

Declaration of Competing Interest. Declaration statement by co-authors for the manuscript, A Comprehensive Review on Detection of Arrhythmia using Deep Learning Methods with Deep Learning Model. We declare the following:

1. Data: The public database from MIT-BIH has been used in this study.
2. Ethics: The research protocol was approved by Jain University, Bangalore.
3. Conflict of Interest: Nothing to declare.
4. Financial support: Nothing to declare.

Acknowledgement. First and foremost, I wish to record my sincere gratitude to Management of Jyothy Institute of Technology and to my beloved Principal, Dr. Gopalakrishna, JIT, Bengaluru for his constant support and encouragement. My sincere thanks to Dr. Raja Praveen N, Associate professor ,Jain University for his valuable support.

REFERENCES

- [1] Rajpurkar, P., Hannun, A. Y., Haghpanahi, M., Bourn, C., and Ng, A. Y. (2017). Cardiologist-level arrhythmia detection with convolutional neural networks. *Nature Communications*, 9(1), 1-9.
- [2] Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P., and Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25(1), 65-69.
- [3] Attia, Z. I., Noseworthy, P. A., Lopez-Jimenez, F., Asirvatham, S. J., Deshmukh, A. J., Gersh, B. J., ... and Friedman, P. A. (2019). An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *The Lancet*, 394(10201), 861-867.
- [4] Ribeiro, A. H., Ribeiro, M. H., Paixão, G. M. M., Oliveira, D. M., Gomes, P. R., Canazart, J. A., ... and Meira Jr, W. (2020). Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, 11(1), 1-9.
- [5] Oster, J., and Clifford, G. D. (2020). A deep learning approach to arrhythmia classification using the ECG and non-ECG physiological signals. *IEEE Journal of Biomedical and Health Informatics*, 24(9), 2536-2545.
- [6] Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., and Acharya, U. R. (2018). Deep learning for healthcare applications based on physiological signals: A review. *Computer Methods and Programs in Biomedicine*, 161, 1-13.
- [7] Tison, G. H., Sanchez, J. M., Ballinger, B., Singh, A., Olgin, J. E., Pletcher, M. J., ... and Smuck, M. (2018). Passive detection of atrial fibrillation using a commercially available smartwatch. *JAMA Cardiology*, 3(5), 409- 416.
- [8] Xia, Y., Wulan, N., Wang, K., and Zhang, H. (2020). Detecting atrial fibrillation by deep convolutional neural networks. *Computers in Biology and Medicine*, 116, 103345

- [9] Jamil, S. and Rahman, M., 2022. A novel deep-learning-based framework for the classification of cardiac arrhythmia. *Journal of Imaging*, 8(3), p.70.
- [10] Reegu, F.A., Abas, H., Gulzar, Y., Xin, Q., Alwan, A.A., Jabbari, A., Sonkamble, R.G. and Dziyauddin, R.A.,(2023). Blockchain-Based Framework for Interoperable Electronic Health Records for an Improved Healthcare System. *Sustainability*, 15(8), p.6337.
- [11] Aseeri, A.O.(2021). Uncertainty-aware deep learning-based cardiac arrhythmias classification model of electrocardiogram signals. *Computers*, 10(6), p.82.
- [12] Tesfai, H., Saleh, H., Al-Qutayri, M., Mohammad, M.B., Tekeste, T., Khandoker, A. and Mohammad, B., (2022). Lightweight Shufflenet Based CNN for Arrhythmia Classification. *IEEE Access*, 10, pp.111842-111854.
- [13] Siddiqui, H.U.R., Saleem, A.A., Bashir, I., Zafar, K., Rustam, F., Diez, I.D.L.T., Dudley, S. and Ashraf, I., (2022). Respiration-based COPD detection using UWB radar incorporation with machine learning. *Electronics*, 11(18), p.2875.
- [14] Dang, H., Sun, M., Zhang, G., Qi, X., Zhou, X. and Chang, Q.(2019). A novel deep arrhythmia-diagnosis network for atrial fibrillation classification using electrocardiogram signals. *IEEE Access*, 7, pp.75577-75590.
- [15] Li, J., Zhang, Y., Gao, L. and Li, X., 2021. Arrhythmia classification using biased dropout and morphology- rhythm feature with incremental broad learning. *IEEE Access*, 9, pp.66132-66140.
- [16] Khan, A.H., Hussain, M. and Malik, M.K.(2021). Arrhythmia classification techniques using deep neural network. *Complexity*, 2021, pp.1-10.
- [17] Shafi, I., Din, S., Khan, A., Díez, I.D.L.T., Casanova, R.D.J.P., Pifarre, K.T. and Ashraf, I.(2022). An effective method for lung cancer diagnosis from ct scan using deep learning-based support vector network. *Cancers*, 14(21), p.5457.
- [18] Satti, F.A., Hussain, M., Hussain, J., Ali, S.I., Ali, T., Bilal, H.S.M., Chung, T. and Lee, S.(2021). Unsupervised semantic mapping for healthcare data storage schema. *IEEE Access*, 9, pp.107267-107278
- [19] Mohonta, S.C., Motin, M.A. and Kumar, D.K.(2022). Electrocardiogram based arrhythmia classification using wavelet transform with deep learning model. *Sensing and Bio-Sensing Research*, 37, p.100502
- [20] Li, Y., Qian, R. and Li, K.(2022). Inter-patient arrhythmia classification with improved deep residual convolutional neural network. *Computer Methods and Programs in Biomedicine*, 214, p.106582.
- [21] Essa, E. and Xie, X., (2021). An ensemble of deep learning-based multi-model for ECG heartbeats arrhythmia classification. *IEEE Access*, 9, pp.103452-103464.
- [22] G Sannino, G De Pietro(2018). A deep learning approach for ECG-based heartbeat classification for arrhythmia detection. *Future Generation computer System*, 86, p 446-455.
- [23] Özal Yldrma., Pawe Pawiakb, Ru-San Tanc, d , U. Rajendra Acharya (2018). Arrhythmia detection using deep convolutional neural network with long duration ECG signals. *Computers in Biology and Medicine*, 102 p. 411-420.
- [24] Fatin A. Elhaj, Naomie Salima, Arief.R Harris(2016). Arrhythmia recognition and classification using combined linear and nonlinear features of ECG signals. *Computer Methods and Programs in Biomedicine*, 127 p.52-63.
- [25] Turker Tuncer, Sengul Dogan, Pawe Pawiak, U. Rajendra Acharya(2019). Automated arrhythmia detection using novel hexadecimal local pattern and multilevel wavelet transform with ECG signals. *Knowledge -based systems*, 186 p.104923.
- [26] Anam Mustaqeem, Syed Muhammad Anwar, Muahammad Majid(2018). Multiclass Classification of Cardiac Arrhythmia Using Improved Feature Selection and SVM Invariants. *Hindawi Computational and Mathematical Methods in Medicine*, 7310496 p.10.
- [27] Miquel Alfaras, Silvia Ortín, Miguel Cornelles Soriano(2019). A Fast Machine Learning Model for ECG- Based Heartbeat Classification and Arrhythmia Detection. *frontier in physics*, doi.org/10.3389/fphy.2019.00103.
- [28] Sonain Jamil, MuhibUr Rahman(2022). A Novel Deep Learning-Based framework for the Classification of Cardiac Arrhythmia. <https://doi.org/10.3390/jimaging8030070>. Turker Tuncer, Sengul Dogan, Pawe Pawiak, U. Rajendra Acharya(2019). Automated arrhythmia detection using novel hexadecimal local pattern and multilevel wavelet transform with ECG signals. *Knowledge -based systems*, 186 p.104923.
- [29] Saroj Kumar Pandey, Rekh Ram Janghel, Aditya Vikram Dev, Pankaj Kumar Mishra[2021] Automated arrhythmia detection from electrocardiogram signal using stacked restricted Boltzmann machine model s42452-021-04621-5
- [30] Kishore G R[2021] Heartbeat Classification and Arrhythmia Detection using Deep Learning *Turkish Journal of Computer and Mathematics Education* PP 1457-1464
- [31] Ramya G. Franklin and B. Muthukumar [2021] Arrhythmia and Disease Classification Based on Deep Learning Techniques *Intelligent Automation and Soft Computing* PP 1-12.
- [32] Shu Lih Oha , Eddie Y.K. Ngb , Ru San Tanc , U. Rajendra Acharyaa, d, e, [2018] Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats *computers in biology and medicine* p278-287.
- [33] Chen Chen, Zhengchun Hua, Ruiqi Zhang, Guangyuan Liu, Wanhui Wen [2020] Automated arrhythmia classification based on a combination network of CNN and LSTM, *Biomedical Signal Processing and Control* 1-13.
- [34] Saeed Mian Qaisar Saeed Mian Qaisar , Alaeddine Mihoub , Moez Krichen and Humaira Nisar[2021] Multirate Processing with Selective Subbands and Machine Learning for Efficient Arrhythmia Classification PP 1-12.
- [35] EHAB ESSA AND XIANGHUA XIE[2021] An Ensemble of Deep Learning-Based Multi-Model for ECG Heartbeats Arrhythmia Classification .PP 1-10.
- [36] ZEESHAN AHMAD, LING GUAN, AND NAIMUL MEFRAZ KHAN[2021] ECG Heartbeat Classification Using Multimodal Fusion ,pp 100615-100626.
- [37] Niken Prasasti Martono , Toru Nishiguchi , Hayato Ohwada[2021] Interpreting Arrhythmia Classification Using Deep Neural Network and CAM-Based Approach, PP 35-40
- [38] Ali Haider Khan , Muzammil Hussain , and Muhammad Kamran Malik[2021] Arrhythmia Classification Techniques Using Deep Neural Network Article ID 9919588, PP 1-12.
- [39] Wusat Ullah, Imran Siddique , Rana Muhammad Zulqarnain , Mohammad Mahtab Alam , Irfan Ahmad, and Usman Ahmad

- Raza[2022], Classification of Arrhythmia in Heartbeat Detection Using Deep Learning 1-10
- [40] Prateek Singh and Ambalika Sharma[2022], Interpretation and Classification of Arrhythmia Using Deep Convolutional Network, PP1-12.
- [41] Jagdeep Rahul A , Lakhan Dev Sharma[2022], Automatic cardiac arrhythmia classification based on hybrid 1-D CNN and Bi-LSTM model PP 312-324.
- [42] Yuanlu Li , Renfei Qiana , Kun Li[2022], Inter-patient arrhythmia classification with improved deep residual convolutional neural network PP 1-10.
- [43] Parul Madan , Vijay Singh , Devesh Pratap Singh , Manoj Diwakar , Bhaskar Pant and Avadh Kisho[2022], A Hybrid Deep Learning Approach for ECG-Based Arrhythmia Classification PP 1- 26.
- [44] Satheesh Kumar, J., Vinoth Kumar, V., Mahesh, T.R. et al. Detection of Marchiafava Bignami disease using distinct deep learning techniques in medical diagnostics. *BMC Med Imaging* 24, 100 (2024). <https://doi.org/10.1186/s12880-024-01283-8>
- [45] Hadaate Ullah , Md Belal Bin Heyat , Faijan Akhtar , Sumbul , Abdullah Y. Muaad , 7Md. Sajjatul Islam, 8Zia Abbas, 3 Taisong Pan, 1 Min Gao, 1 Yuan Lin , 1,9 and Dakun Lai[2022], An End-to-End Cardiac Arrhythmia Recognition Method with an Effective DenseNet Model on Imbalanced Datasets Using ECG Signal PP 1-23.
- [46] B MOHAN RAO and AMAN KUMAR[2022], Detection of heart arrhythmia based on UCMFB and deep learning technique PP 1-15.
- [47] M. Degirmenci , M.A. Ozdemir , E. Izci , A. Akan[2022], Arrhythmic Heartbeat Classification Using 2D Convolutional Neural Networks PP 1-12.
- [48] HURUY TESFAI , HANI SALEH[2022], Lightweight Shufflenet Based CNN for Arrhythmia Classification PP 1-13.
- [49] Evren Kymaç, Yasin Kaya[2023], A novel automated CNN arrhythmia classifier with memory- enhanced artificial hummingbird algorithm PP 1-11.
- [50] Adel A. Ahmed , Waleed Ali , Talal A. A. Abdullah and Sharaf J. Malebar[2023], Classifying Cardiac Arrhythmia from ECG Signal Using 1D CNN Deep Learning Model PP 1-17.
- [51] Zubair Rahman, A.M.J., Gupta, M., Aarathi, S. et al. Advanced AI-driven approach for enhanced brain tumor detection from MRI images utilizing EfficientNetB2 with equalization and homomorphic filtering. *BMC Med Inform Decis Mak* 24, 113 (2024). <https://doi.org/10.1186/s12911-024-02519-x>
- [52] M, M.M., T. R, M., V, V.K. et al. Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50. *BMC Med Imaging* 24, 107 (2024). <https://doi.org/10.1186/s12880-024-01292-7>.
- [53] Albalawi, E., T.R., M., Thakur, A. et al. Integrated approach of federated learning with transfer learning for classification and diagnosis of brain tumor. *BMC Med Imaging* 24, 110 (2024). <https://doi.org/10.1186/s12880-024-01261-0>.
- [54] Velagapudi Swapna Sindhu, Kavuri Jaya Lakshmi[2023], A novel deep neural network heartbeats classifier for heart health monitoring PP 1-10.
- [55] Alshuhail, A., Thakur, A., Chandramma, R. et al. Refining neural network algorithms for accurate brain tumor classification in MRI imagery. *BMC Med Imaging* 24, 118 (2024). <https://doi.org/10.1186/s12880-024-01285-6>
- [56] Machine learning approach for COVID-19 crisis using the clinical data (2020). *Indian Journal of Biochemistry and Biophysics*. <https://doi.org/10.56042/ijbb.v57i5.40803>
- [57] Mahmoud, L., Praveen, R. (2020, December 8). Network Security Evaluation Using Deep Neural Network. 2020 15th International Conference for Internet Technology and Secured Transactions (ICITST). <https://doi.org/10.23919/icitst51030.2020.9351326>
- [58] K N, R. P., Pasumarty, R. (2021, November 11). Recognition of Bird Species Using Multistage Training with Transmission Learning. 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC). <https://doi.org/10.1109/i-smac52330.2021.9640676>
- [59] Smitha B A and Raja Praveen K N, ORDSAENet: Outlier Resilient Semantic Featured Deep Driven Sentiment Analysis Model for Education Domain, *Journal of Machine and Computing*, vol.3, no.4, pp. 408- 430, October 2023. doi: 10.53759/7669/jmc202303034

Edited by: Dhilip Kumar V

Special issue on: Unleashing the power of Edge AI for Scalable Image and Video Processing

Received: Jan 6, 2024

Accepted: Jun 12, 2024



OPTIMIZING WASTE REDUCTION IN MANUFACTURING PROCESSES UTILIZING IOT DATA WITH MACHINE LEARNING APPROACH FOR SUSTAINABLE PRODUCTION

FAISAL ALTARAZI*

Abstract. Sustainable manufacturing with the Internet of Things (IoT) reduces environmental impacts, conserves natural resources, saves energy, and improves worker, community, and consumer safety while maintaining economic viability. IoT's network of sensors and intelligent devices collects and analyzes data throughout the production lifecycle, enabling organizations to fulfil sustainability objectives and adopt more efficient, less wasteful operations. Waste management and reduction measures are the focus of sustainable manufacturing research. Improvements are needed to simplify waste management and reduce production waste. Thus, in this study, we introduce an innovative machine learning technology called "EcoEfficientNet", developed to tackle this problem. Our study addresses the issue of waste in manufacturing processes. EcoEfficientNet employs sophisticated deep learning algorithms to analyze complex production data, allowing it to identify significant patterns and determine specific areas where waste can be significantly minimized. EcoEfficientNet's approach to waste reduction in manufacturing processes revolves around three main strategies: data-driven analysis, optimization recommendations, and adaptable learning for continual enhancement. EcoEfficientNet's success lies in its capacity for perpetual learning, enabling it to adapt to novel data and evolve alongside production settings. An extensive case study of a particular manufacturing process is carried out to assess the efficiency of EcoEfficientNet and provide helpful perspectives into the model's effectiveness. By incorporating this method into the manufacturing process, organizations have seen a decrease in waste generation of up to 30%, demonstrating the applicability and efficacy of machine learning in improving sustainable manufacturing processes. The key to EcoEfficientNet's success is its ability to engage in continuous learning, allowing it to adjust to new data and develop in tandem with operational environments.

Key words: Waste reduction, IoT Sensed data, deep learning, decision processing, operational efficiency, manufacturing, sustainability.

1. Introduction. Sustainable manufacturing [1] is a crucial concept in the manufacturing sector that focuses on reducing environmental consequences, conserving energy and natural resources, ensuring worker safety, and maintaining financial viability. Although there have been notable progressions, the industry still faces challenges in managing and minimizing waste, which presents a promising opportunity for innovation. In response to this identified deficiency, the present study proposes "EcoEfficientNet," an advanced machine learning (ML) network specifically developed to address the shortcomings in waste management in industrial processes.

Integrating real-time data from various sensors and devices across the factory floor, including IoT data in "EcoEfficientNet" for evaluation purpose, significantly enhances its capacity to revolutionize sustainable manufacturing. This convergence allows for accurate monitoring of resource use, operational variables, and waste generation, providing the machine learning network with highly accurate data essential for detecting inefficiencies and forecasting opportunities for waste reduction. EcoEfficientNet utilizes the constant stream of IoT data to acquire knowledge and enhance operations actively, leading to improvements in waste management and the development of a more environmentally friendly production system.

Moreover, the need for technological intervention arises from growing ecological issues and strict rules designed to promote sustainable activities. Conventional waste management systems must be more robust because they cannot adjust to intricate production settings and optimize operations in real-time [2]. Hence, implementing intelligent systems with the capacity to analyze data and optimize processes in real time is beneficial and essential for advancing manufacturing towards increased sustainability.

Thus, in this study, we introduce "EcoEfficientNet", which is at the forefront of this transformation. By using sophisticated deep learning algorithms, this system examines the complexities of production data, revealing trends and identifying crucial areas for minimizing waste [3]. The study is essential as it can completely alter

*University of Jeddah, Jeddah, Saudi Arabia, (fmaltarazi@uj.edu.sa)

waste management by converting extensive data into practical and valuable information. This will contribute to the area's existing knowledge and provide a model that can be replicated to promote sustainable practices.

The initial objective is to showcase the effectiveness of "EcoEfficientNet" in substantially reducing waste via its analytical capabilities. This objective is accomplished by thoroughly examining both past and current production data, allowing for a complete comprehension of trends in waste formation. The second goal is to verify the flexibility and ongoing learning capacities of "EcoEfficientNet." The research intends to demonstrate the model's capacity to smoothly integrate into current production processes and adapt to changes, assuring long-term sustainability and efficacy. This will be achieved via empirical experiments conducted throughout the study.

This study addresses a significant need for sustainable manufacturing and advances the industry by demonstrating the tangible advantages of ML techniques. The expected result is a substantial drop in waste, with first deployments demonstrating a reduction of up to 30%, highlighting the revolutionary potential of "EcoEfficientNet." This study is positioned to establish a standard in sustainable manufacturing, providing a solid basis for future progress and strengthening the importance of technological advances in attaining sustainable and effective manufacturing procedures.

2. Related Work. Artificial neural network (ANN) approaches have become commonplace across several academic disciplines owing to their inherent capacity to acquire knowledge from provided instances. ANNs are extensively used and considered the predominant machine learning algorithms [4]. Additionally, they have been proposed for several manufacturing applications, particularly in the context of predictive automation. [5] examines explicitly the use of Artificial Neural Networks (ANNs) to classify the condition of tools in CNC (Computer Numerical Control) milling machines. The distinctiveness of this technique is in its retrofitting strategy, which allows older equipment to conform to the norms of Industry 4.0. The research showcases the successful implementation of tool wear monitoring using integrated detectors on a customizable prototyping platform. The ANN model effectively enables the modernization of outdated equipment and surpasses the performance of Support Vector Machine (SVM) and k-nearest Neighbors (KNN) approaches. [6] introduced a technique in which vibration information from a hypothetical motor unit is used to train an ANN to forecast equipment malfunctions. The method is distinguished by its use of frequency and amplitude data to predict the exact moment at which the vibrating system would break. The Multilayer Perceptron (MLP) approach was selected because of its simplicity in implementation and ability to generalize. The research demonstrates that the ANN outperforms Random Forest (RF), Regression Tree (RT), and SVM in making predictions over medium and long time periods. However, its performance is comparable to these methods in the short term. [7] use ANNs and SVMs to forecast the deterioration of gauges in train tracks.

The study concentrates explicitly on both straight and curved sections. The ANN model has a substantial coefficient of determination, which signifies its robust prediction capability. Although both SVM and ANN models provide excellent outcomes, the ANN model is marginally superior at forecasting gauge variation for linear segments. [8] constructed a test apparatus to replicate the functioning of a wind turbine, with a specific emphasis on observing its state employing vibration evaluation. The created ANN model, designed to identify the health status of essential components, has a remarkable accuracy score of 92.6%. This study highlights the possibility of ANNs in predicting and preventing maintenance issues in the field of green energy. [9] conduct a comparative analysis of physics-based models as well as models built on neural networks (NN) to assess the deterioration of instances in Auxiliary Power Units (APUs). This approach emphasizes a universal modeling strategy to tackle the difficulty of varying component features. The results indicate that the physics-based method is more dependable for deteriorated starts, but the NN model performs very well with starters in optimal circumstances. [10] presented a system that utilizes data to diagnose and predict the performance of machinery and maintenance expenses. Furthermore, a precise data labeling mechanism is devised for supervised learning by contrasting the serial numbers of target components on consecutive dates. The research used actual data from vending machines to verify the concept architecture using three distinct classifiers: SVM, RF, and Gradient Boosting Machines (GBM). The outcomes of the cross-validated simulated events demonstrate that the diagnostic approach can reach an accuracy of over 80%. Therefore, the proposed GBM model can effectively diagnose and monitor complicated machine types. The prognosis approach surpasses one-stage traditional forecasting techniques. Symbolic Regression (SR) has been used to estimate the state of well-functioning industrial

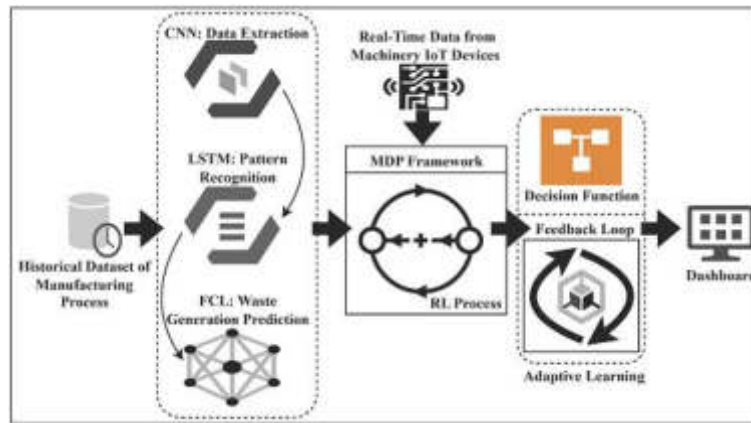


Fig. 3.1: Framework of EcoEfficientNet

equipment [11]. The work introduced a mechanism for handling idea drifts in persistent information streams. In addition, a practical case study was shown with industrial radial fans. The findings from the computerized information indicate that concept drift diagnosis and prognosis were highly effective. [12] addresses the crucial problem of inadequate productivity in an industrial setting, specifically emphasizing a tire manufacturing firm in Peru. The study's primary aim is to combine and design various tools to improve effectiveness, thereby decreasing the expensive upkeep of manufacturing machinery and the significant expenditures of adopting new systems. The primary focus of this work is on the creation of a waste-management strategy. This model is specifically designed to minimize the time required to set up and implement a viable operational control system, with the ultimate objective of enhancing the Overall Equipment Effectiveness (OEE) score. The model's assessment in an actual production setting yielded remarkable results, including a 13% enhancement in the OEE score and a substantial 22.5% decrease in the setup period.

Previous research has shown that ANNs can effectively monitor tool conditions and expected equipment malfunctions and perform scheduled upkeep. However, there has been less emphasis on using such strategies to improve waste management and decrease wastage in industrial environments. It is crucial to address this shortcoming to promote the progress of sustainable manufacturing methods. Suppliers can optimize resource utilization, mitigate environmental consequences, and improve their general productivity by incorporating ML techniques specifically designed for waste reduction.

3. Methodology. The EcoEfficientNet framework in Fig. 3.1 depicts a novel strategy in manufacturing that focuses on reducing waste by using sophisticated Deep Learning (DL) and Reinforcement Learning (RL) models. The two-stage procedure starts by using an advanced DL model that integrates Convolutional Neural Networks (CNN), Fully Connected Layers (FCL), and Long Short-Term Memory (LSTM) to detect and examine patterns in the manufacturing process. The second step utilizes these insights using a Reinforcement Learning (RL) model based on a Markov Decision Process (MDP) to implement strategic adjustments based on the real-time data from IoT devices during the manufacturing process. EcoEfficientNet optimizes efficiency and minimizes waste by constantly adjusting activities, aligning manufacturing operations with sustainable principles.

3.1. First Phase. Initially, a sophisticated DL model is used to identify patterns in the backdrop of minimizing waste in a manufacturing procedure. A fusion of CNN, LSTM, and FCL is implemented to achieve this. The first phase of the EcoEfficientNet Model has three primary components. CNN stands for Convolutional Neural Network, LSTM stands for Long Short-Term Memory, and FCL stands for Fully Connected Layer. Input data that contains visuals or spatial trends (like sensor heatmaps) can be extracted using the CNN layers, which deal with spatial features. IoT devices served as the primary sources for continuous, real-time data feeding into the EcoEfficientNet system. LSTM layers excel in processing time-series data by obtaining

temporal relationships and sequences of events, such as consumption of resource patterns over time. On the other hand, FCL layers act as the final decision-making layers, interpreting the features extracted by the CNN and LSTM. They are responsible for predicting waste generation in structured data, such as equipment logs and production data.

Computation at CNN Layers. The CNN layer utilizes several filters, k , to generate feature maps from the input visual (if applicable) I , which has dimensions HCEWØED (height, width, depth) [13]. Such process can be analytically defined as Equation (3.1),

$$f_{ij}^k = \text{ReLU} \left[\sum_{r=0}^{R-1} \sum_{c=0}^{C-1} \sum_{d=0}^{D-1} F_{(m \cdot n \cdot d)}^k \cdot I_{(i+r),(j+c),d} + e^k \right] \quad (3.1)$$

In Equation (3.1), f_{ij}^k represents the essential feature at (i, j) at the k th feature map, $F_{(m \cdot n \cdot d)}^k$ denotes the k^{th} filter employed at i^{th} input, and e^k indicates the bias at k^{th} filter.

Computation at LSTM. The acquired attributes are smoothed and then potentially processed via additional substantial layers before inputting into LSTM cells [14]. An LSTM cell sequentially analyzes time-series information, updating and preserving both a cell state (Z_t) and hidden state (h_t) at each time step. At each successive step t , the LSTM modifies its states in the following manner [15], Equation (3.2) to (3.7) :

$$\text{Forgetgate} : F_t = \sigma (w_F \cdot [h_{(t-1)}, I_t] + e_f) \quad (3.2)$$

$$\text{Inputgate} : I_t = \sigma (w_i \cdot [h_{(t-1)}, I_t] + e_l) \quad (3.3)$$

$$\text{Cellcandidate} : \tilde{Z}_t = \tanh (w_Z \cdot [h_{(t-1)}, I_t] + e_Z) \quad (3.4)$$

$$\text{NewerCellstate} : Z_t = F_t * Z_{(t-1)} + I_t * \tilde{Z}_t \quad (3.5)$$

$$\text{Outputgate} : o_t = \sigma (w_o \cdot [h_{(t-1)}, I_t] + e_o) \quad (3.6)$$

$$\text{NewerHiddenstate} : h_t = o_t * \tan h [Z_t] \quad (3.7)$$

where $*$ indicates the element-wise multiplication, σ denotes sigmoidal function, I_t indicates the input, e and w denotes the bias and weight for each gate, respectively.

Computation of FCL. The LSTM's output, denoted as h_t , is then fed into a FCL for the intent of classifying the states of operation into categories like normal, under-efficient, over-efficient (classifying the level of waste production). The FCL does the following operation [16]:

$$\delta = \sigma (w_{FCL} \cdot h_t + e_{FCL}) \quad (3.8)$$

In Equation (3.8), e_{FCL} and w_{FCL} are the biases and weights of the FCL.

Backpropagation [17] is used to optimize the parameters associated with the model throughout training. The loss function is used to quantify the discrepancy between the actual waste level and the projected waste level for each batch of data.

$$f(L) = \frac{1}{B} \sum_1^B [\delta_i - \tilde{\delta}_i]^2 \quad (3.9)$$

where B denotes the batch size, δ_I represents the true value, and $\tilde{\delta}_I$ is the predicted value by the model in Equation (3.9).

The derivatives of the loss function concerning the model's parameters are then calculated and used to adjust the parameters employing the Adam optimizer. In the first phase, EcoEfficientNet combines CNN, LSTM, and FCL. This enables the model to comprehend the manufacturing data's temporal and spatial patterns.

Consequently, EcoEfficientNet can make precise predictions about waste emergence, which in turn can be utilized to improve the manufacturing process and minimize such waste.

3.2. Second Phase. The second stage involves incorporating a reinforcement learning (RL) model that will execute actions to enhance the manufacturing process by leveraging the predictions generated by the DL model. This phase has two primary components: the Markov Decision Process (MDP) [18] and the learning process. In this context, the issue of waste reduction is conceptualized as a MDP, whereby the state corresponds to the existing condition of the manufacturing process, actions denote potential modifications, and rewards are allocated for actions that minimize waste.

Decision-Making Process. The MDP is a conceptual framework used to represent decision-making scenarios in which outcomes are influenced by both random factors and the management-maker's regulation. MDPs are valuable tools for analyzing optimization issues addressed using adaptive programming and RL techniques [18].

The RL model operates within the framework of an MDP and is characterized by the tuple (A, S, T, R, ϕ) . Here, the state signifies the existing condition of the production process, actions denote potential modifications, and incentives are granted for actions that minimize waste. Thus, MDP is characterized by the tuple (A, S, T, R, ϕ) , where:

A is a collection of activities that symbolize potential modifications to the process.

S is a collection of states that represents the present state of the manufacturing process.

The state transition potential matrix, denoted as T, represents the likelihood of moving from state s_t to state s_{t+1} after performing action a_t .

The reward function, denoted as $R[a_t, s_t]$, determines the reward obtained when a_t is taken in s_t .

The discount factor ϕ is used to strike a balance between present and potential rewards in the future.

An MDP aims to identify a strategy π that prescribes the optimal action 'a' to be taken in each S to maximize the overall expected reward. Q-learning facilitates [19] sophisticated learning processes, enabling the operation of intricate state spaces and acquiring optimum strategies via time. At first, the model randomly investigates several strategies inside safe operational boundaries to comprehend their influence on waste production. The gradual transition towards optimal strategies as the system gains knowledge from the results of its activities.

Learning Process. The DL model's predictions are integrated with the RL model's action-value estimates to facilitate informed decision-making. Furthermore, the DL model enhances the RL model by conveying information about the probable outcomes of various actions, enriching the state representation. This research used a widely used RL approach known as the Q-learning mechanism. The Q-learning update step at each t employs the Bellman formulation in the following manner [20]:

$$q^{new}[\alpha_t, s_t] = q[\alpha_t, s_t] + \alpha \{ \phi \max_a [\alpha_t, s_{t+1}] + R[\alpha_t, s_t] - \alpha[\alpha_t, s_t] \} \quad (3.10)$$

In Equation (3.10), α denotes the learning rate, $R[\alpha_t, s_t]$ represents the immediate reward received after taking a_t in s_t , $\max_a [\alpha_t, s_{t+1}]$ signifies the estimate of optimal future value.

To balance exploitation and exploration, a method known as ϵ -greedy is applied [21]. This approach involves the model randomly selecting a_t (exploration) with a probability of ϵ and selecting a_t with the greatest Q-value (exploitation) with a probability of $1 - \epsilon$. Table 3.1 represents the working mechanism of action-value function optimization [22].

Further, for real time data integration, let's assume D_t be the data received at time t. The data stream is fed into the system continuously, which is expressed as in Equation (3.11):

$$a(D_t) = \{ I_{(1t)}, I_{(2t)}, I_{(3t)}, \dots, I_{(nt)} \} \quad (3.11)$$

where $I_{(it)}$ denotes varying features of deployed machines in the production unit.

Implementation of Strategic Decisions and Continuous Learning: At this stage, the EcoEfficientNet framework is evaluated via received continuous, real-time data from IoT devices as its primary sources. This part of EcoEfficientNet involves decision function and feedback looping processes [23]. In the case of decision function, $a(D_t)$ considers the current state data as $I_{(it)}$ and suggests adjustments.

From Equation (3.12), q denotes the learned action-value function via RL process and A represents the possible set of actions.

$$\alpha(D_t) = \arg \max_{a \in A} q[\alpha, D_t] \quad (3.12)$$

Table 3.1: Working Mechanism of Action-Value Function Optimization

<p>Input: s (state), a (action), α (learning rate), φ (discount factor), ϵ (exploration rate), ϵ_{\min} (minimum exploration rate), decayrate (rate at which exploration rate decays), E (number of episodes to run the algorithm), R (reward function), q (action-value function), and initialized arbitrarily is zeros.</p> <p>Output: Optimized action-value function q.</p>
<p>For each E from 1 to N :</p> <p>Initialize s to the starting state.</p> <p>While terminal state is not reached:</p> <p>Choose a from s using policy derived from q :</p> <p>Execute a, observe r, and s'</p> <p>Update q-value for s and a :</p> $q[a, s] = q[a, s] + \alpha \{ \max_{a'} q[a', s'] + R - q[a, s] \}$ <p>$s \leftarrow s'$ (move to the new state)</p> <p>Decay ϵ where ($\epsilon \geq \epsilon_{\min}$)</p> $\epsilon \leftarrow \max(\epsilon, \epsilon_{\min} \cdot \text{decay_rate})$

Simultaneously, on the other hand, the model updates via action outcome recordings which is possible through feedback looping. For action outcome recordings, $R[\alpha(D_t), D_t]$ indicate the reward function obtained due to the outcome of $a(D_t)$ for the given D_t which is expressed as Equation (3.13),

$$R[\alpha(D_t), D_t] = \text{Efficiency}_{\text{gain}}(\alpha, D_t) | \text{Efficiency}_{\text{loss}}(\alpha, D_t) \mapsto \alpha(D_t) \quad (3.13)$$

Thus, the current q is updated based on the new incoming data and the $R[\alpha(D_t), D_t]$, which can be expressed as Equation (3.14),

$$q[a(D_t), D_t] \leftarrow \{q[a(D_t), D_t] + \varphi \max_{a'} q[a', D_{t+1}] + \alpha R[a(D_t), D_t] - q[a(D_t), D_t]\} \quad (3.14)$$

In addition, the dashboard is regularly updated with the essential metrics, as in Equation (3.15).

$$\text{dashboard}_t = \{(\mathbf{m}_t | \mathbf{m}) \in M\} \quad (3.15)$$

The system operates cyclically, incorporating actual information, using ML models for making decisions, and continuously refining these models through feedback concerning performance. The system is meant to be adaptable and continually enhance its performance by assimilating fresh data and analyzing the results of its operations. The EcoEfficientNet, developed by integrating hybridized sophisticated ML principles, emerges as a formidable instrument for minimizing waste. It can acquire knowledge and adjust to the unique circumstances and obstacles encountered in a manufacturing process. This leads to an intelligent and data-oriented strategy for sustainable manufacturing.

4. Performance Evaluation and Analysis.

4.1. Dataset. The dataset must comprehensively cover various aspects of the manufacturing process to effectively train and validate the machine learning model. So we have chosen an appropriate dataset from IEEE Dataport [24] that is meticulously structured to encapsulate a broad spectrum of key metrics and data sources, tailored to address specific needs of the manufacturing process. Few crucial metrics of the dataset are listed and described as follows:

1. Resource consumption in the dataset is the tracking of resources consumed during manufacturing. This encompasses the energy utilized, often quantified in kilowatt-hours (kWh), as well as the raw materials used, typically measured in kilograms or similar units. Accurately monitoring these inputs is pivotal for understanding and optimizing resource utilization.

Table 4.1: Vital Features Considered Process to Optimize Waste Reduction during the Manufacturing Process

Attribute	Description	Range/Type
Timestamp	Date and time of data entry	Last quarter of 2023 in the timestamp of 00 : 00 – 23 : 59.
MachineID	Identifier for the machine	[List of Machines] (e.g., Machine_01, Machine_02,..)
Resource Consumption	Energy and materials used	[Min Consumption - Max Consumption] (e.g., 50–500kWh for energy)
Production Output	Volume of products made	[Min Output - Max Output] (e.g., 100 - 1000 units)
Waste Generated	Waste material produced	[Min Waste - Max Waste] (e.g., 10 – 100 kg)
Operational Efficiency	Efficiency of the operation	[Lower Efficiency - Higher Efficiency] (e.g., 0.5–1.5 ratio)
Machine Temperature	Average operational temperature	[Min Temp - Max Temp] (e.g., 20 – 100°C)
Machine Vibration	Level of machine vibration	[Min Vibration - Max Vibration] (e.g., 0.5 - 2.5 mm/s)
Maintenance Status	Indicates maintenance activity	[0 (No), 1(Yes)]
Quality Control (QC) Failures	Number of failed QC checks	[Min Failures - Max Failures] (e.g., 0 - 5 failures)
Operator Shift	Shift during which data was recorded	[Shift A, Shift B, Shift C]

2. Production output is another vital metric is the volume of finished products yielded within a given time frame. This output can be measured in various units such as the number of items produced or their total weight or volume, offering a direct insight into the productivity of the manufacturing process.
3. Integral to sustainable manufacturing practices, waste generation metric quantifies the waste produced, which includes material scraps, defective products, and any form of emissions. Tracking this in terms of weight or volume is crucial for environmental impact assessment and for formulating strategies to minimize waste.
4. Operational parameters includes a range of data reflecting the operational health and efficiency of manufacturing equipment, such as machine operating temperatures, vibration levels, operational speed, and instances of downtime. These parameters are key indicators of machine performance and maintenance needs.

The granularity of data collection is meticulously chosen based on the specific nature of the manufacturing process. In high-pace environments like assembly lines, data is often collected at an hourly rate to capture the dynamic nature of operations. Conversely, in slower-paced manufacturing processes such as in chemical production, a daily data collection regime might suffice to provide meaningful insights. In addition, two major data source identifiers are incorporated, Internal and external manufacturing data, which includes machine logs, production records, quality control reports, environmental data, and other industry benchmarks. The complete list of attributes of the dataset is represented in Table 4.1 (Karthick Raghunath, 2024). Table 2 serves as a guide for setting up data collection protocols and designing machine learning models for sustainable manufacturing.

4.2. Empirical Setup. Table 4.2 presents the essential requirements for empirically evaluating the EcoEfficientNet model’s effectiveness in reducing waste in manufacturing. To assess the efficacy of the EcoEfficientNet, we conduct a comparison study with other established ML techniques such as GBM, SR, MLP, and RF.

The evaluation of the EcoEfficientNet model’s accomplishment in waste reduction optimization across manufacturing processes, as well as its comparison with other ML models such as GBM, SR, MLP, and RF, includes incorporating the following four performance criteria. The following metrics are used to measure the efficacy, efficiency, and precision of the models in the particular context: Overall Equipment Effectiveness (OEE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Waste Reduction Percentage (WRP).

Table 4.2: Requirement Components for the Empirical Analysis

Component	Hyperparameter	Optimal Value
CNN Layer	Number of Layers	3
	Number of Filters	64
	Filter Size	3×3
	Stride	1
	Activation Function	ReLU
	Pooling Type	Max
LSTM Layer	Number of Layers	2
	Units per Layer	100
	Dropout Rate	0.3
	Recurrent Dropout Rate	0.3
FCL	Number of Layers	2
	Units per Layer	128
	Activation Function	ReLU
RL Model	Learning Rate	0.01
	Discount Factor (γ)	0.95
	Exploration Rate (ϵ)	0.2
	Replay Memory Size	10000
	Batch Size	64
	Target Network Update Frequency	Every 5000 steps
Software Requirement	Packages & Versions	
Programming Language	Python V3.8	
Deep Learning Libraries	PyTorch V1.1	
Machine Learning Libraries	Scikit-learn V0.24	
High-Level Neural Network API	Keras V2.4	
Numerical Computation	NumPy V 1.20	

OEE is employed in factory settings to quantify the efficiency of a manufacturing procedure. It consolidates several aspects of business operations into a unified and all-encompassing measure. OEE is computed as [25],

$$OEE = (AvailabilityPerformanceQuality) \quad (4.1)$$

where, Availability is the proportion of run time parted over the intended production time. Performance is calculated as the proportion of Ideal Cycle Time parted by the proportion of run time divided by total components in Equation (4.1). Quality is determined by the proportion of good components parted by total components.

Fig. 4.1 presents a concise graphical representation of the OEE (Overall Equipment Efficiency) for several techniques, such as EcoEfficientNet, GBM, SR, MLP, and RF. The investigation reveals that EcoEfficientNet achieves an outstanding OEE score of 0.85, indicating its exceptional effectiveness in the manufacturing procedure. The improved efficacy can be ascribed to the model's sophisticated use of CNN, LSTM, and FCL, which excel in recognizing patterns and enhancing processes, particularly in waste reduction.

GBM, while it has an OEE of 0.75, performs well compared to other models but is not as capable as EcoEfficientNet. GBM has high prediction accuracy, although it may need to be more proficient in managing the temporal and spatial data patterns crucial in industrial environments. SR with an accuracy of 0.65, and MLP, with an accuracy of 0.70, while valuable in certain situations, demonstrate lower proficiency in effectively managing the intricacies of industrial data compared to EcoEfficientNet. With a score of 0.68, the RF model has modest efficacy but is often surpassed by models that provide more advanced skills for integrating and analyzing data, such as EcoEfficientNet.

The dominance of EcoEfficientNet in this scenario may be attributed to its customized structure, specifically designed to enhance industrial processes by monitoring several data points and operational efficiency. By using

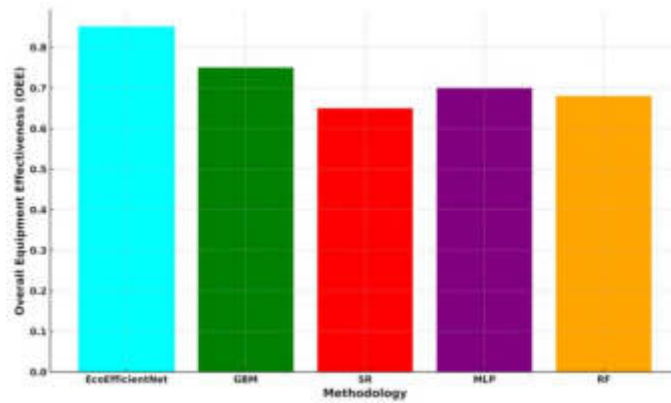


Fig. 4.1: Analysis of OEE in the Manufacturing Process

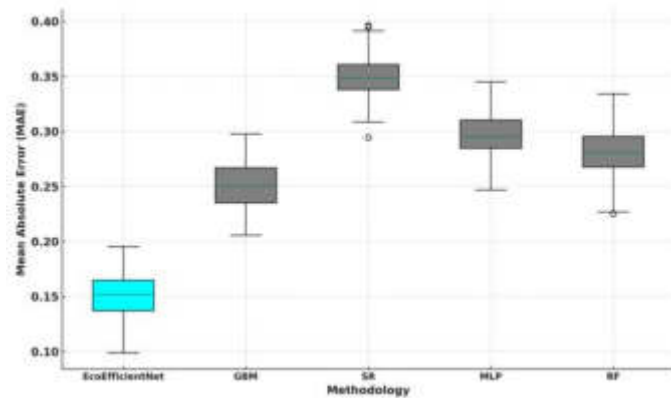


Fig. 4.2: Evaluation of MAE

this holistic approach, it is possible to develop a more refined and efficient optimization plan, resulting in increased OEE values.

MAE quantifies the level of accuracy in predicting continuous information [26].

$$MAE = \left(\frac{1}{n}\right) \times \sum |y_i - \hat{y}_i| \quad (4.2)$$

In Equation (4.2), y_i is the true value, \hat{y}_i is the predicted value, and n is the number of observations.

Fig. 4.2 compares the MAE across several techniques, such as EcoEfficientNet, GBM, SR, MLP, and RF. The most notable aspect of this plot is the exceptional performance of EcoEfficientNet, which is highlighted by its distinctive coloration. EcoEfficientNet has superior accuracy and consistency in predictions compared to the other techniques, as seen by its lower median MAE and narrower interquartile range. The exceptional performance of EcoEfficientNet is in line with its innovative deep learning architecture, which seamlessly combines CNN, LSTM, and FCL to identify complex patterns accurately and optimize industrial processes.

On the other hand, GBM, SR, MLP, and RF exhibit more variability in MAE, as seen by their broader box ranges and higher median values. This implies that while these strategies are successful in some instances, they may not be as proficient as EcoEfficientNet in dealing with intricate, uninterrupted data that is unique to reducing waste in manufacturing. Higher MAE levels indicate less precision in forecasts, resulting in less efficient results in real-world scenarios. The lower and more constant MAE of EcoEfficientNet highlights its

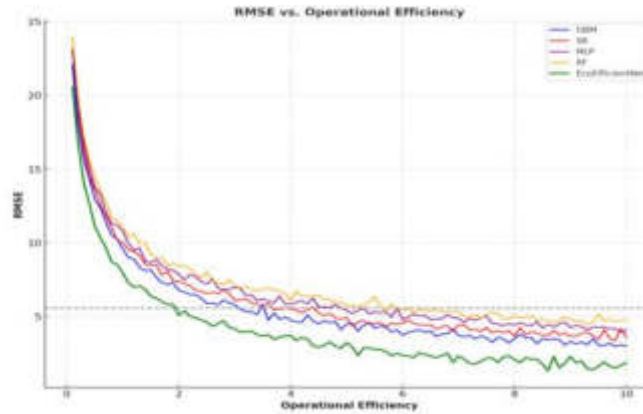


Fig. 4.3: Evaluation of RMSE

appropriateness for complex and ever-changing settings such as sustainable manufacturing. Accuracy and dependability are essential for making informed decisions and optimizing processes in these situations.

RMSE in Equation (4.3) quantifies the magnitude of errors by calculating the square root of the mean of the squared discrepancies between expected and actual outcomes [27].

$$RMSE = \sqrt{\left(\frac{1}{n}\right) \times \sum (|y_i - \hat{y}_i|)^2} \tag{4.3}$$

Based on the observed RMSE values in Fig. 4.3, it can be concluded that EcoEfficientNet demonstrates superior performance compared to the other models (GBM, SR, MLP, and RF) in terms of operational efficiency over the whole range. The EcoEfficientNet regularly exhibits a lower error rate than the other approaches, indicating superior prediction accuracy. This is consistent with the previously mentioned idea that EcoEfficientNet, a model created for environmentally friendly production, utilizes sophisticated DL algorithms such as CNNs, FCLs, and LSTMs to reduce waste by detecting trends and inefficiencies in manufacturing procedures.

The resultant demonstrates the effectiveness of EcoEfficientNet, which can be credited to its advanced design and ability to learn, adapt, and evolve continuously with updated information. This attribute is essential for sustainable manufacturing since adjusting to ever-changing production settings and minimizing waste is necessary. The higher technical performance of EcoEfficientNet, as seen by the reduced RMSE values, validates its usefulness in promoting sustainable manufacturing. It does this by offering data-driven insights that facilitate operational enhancements.

A relevant indicator called Waste Reduction Percentage [28] is used to confirm the extent of waste reduction achieved via the use of EcoEfficientNet in the course of production. The ML model’s impact on waste reduction can be quantitatively measured by computing the decrease in waste production Equation (4.4).

$$WRP = \frac{\text{waste before} - \text{waste after}}{\text{waste before}} \times 100 \tag{4.4}$$

Fig. 4.4 illustrates the extent of waste reduction in a manufacturing environment before and after adopting several ML techniques. The 'Before' bars represent the original quantity of trash produced, while the 'After' bars display the decreased amount after implementation, with the disparity between them indicating the effectiveness of each machine learning approach in waste reduction.

Upon examining the outcome, it is apparent that all ML approaches have a role in reducing waste. However, EcoEfficientNet has the most effect, decreasing waste from about 91 units to 57.5 units. This is consistent with the prior conversations where EcoEfficientNet, with advanced deep learning techniques, was mainly created to address waste in industrial processes. The model’s sophisticated algorithms, such as CNNs, FCLs, and LSTMs, allow it to recognize and respond to patterns that result in waste, thus reducing it.

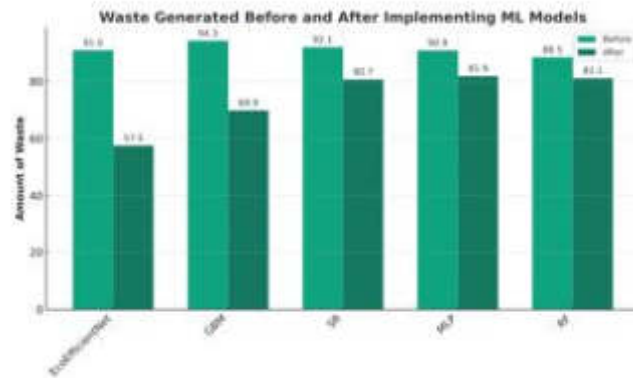


Fig. 4.4: Analysis of Waste Reduction for Various Models

Table 4.3: Sample Segment of the Outcome Showcasing the Optimal Performance for Sustainable Manufacturing over a 24-Hour Period

Timestamp	Machine	Resource Consumption (kWh)	Production Output (units)	Waste Generated (kg)	Operational Efficiency (ratio)	Machine Temperature (C)	Machine Vibration (mm/s)	Maintenance Status	Quality Control Failures	Operator Shift
2023-10-01 00:00:00	Machine_01	50.00	100.00	10.00	0.50	20.00	0.50	0	0.00	Shift A
2023-10-01 00:00:00	Machine_01	69.57	139.13	13.91	0.54	23.48	0.59	0	0.22	Shift A
2023-10-01 00:00:00	Machine_02	89.13	178.26	17.83	0.59	26.96	0.67	0	0.43	Shift A
2023-10-01 00:00:00	Machine_03	108.70	217.39	21.74	0.63	30.43	0.76	0	0.65	Shift A
...
2023-10-01 00:00:00	Machine_05	500.00	1000.00	100.00	1.50	100.00	2.50	1	5.00	Shift A

The consequences of reducing such waste are significant for the production ecosystem in an industrial setting. EcoEfficientNet’s substantial reduction in waste output decreases the manufacturing process’s environmental impact and leads to cost savings and improved resource use. This is especially crucial in sectors where materials disposal leads to environmental deterioration and operational inefficiency.

By integrating EcoEfficientNet into the production process, as shown in Fig. 4.4, it is possible to reduce waste output by about 37%. This demonstrates the practicality and effectiveness of machine learning in enhancing sustainable manufacturing practices. Such waste reduction may lead to a series of beneficial outcomes, such as decreased consumption of raw materials, reduced energy use, and less environmental contamination. These outcomes are essential elements of sustainable industrial operations. The result represents progress towards environmentally friendly production, highlighting the importance of modern technology such as EcoEfficientNet in promoting sustainability in the sector.

Table 4.3 displays the measured data at different time points throughout the production process during 24 hours (sample). These optimal values demonstrate the equilibrium between elevated productivity (increased manufacturing output), effectiveness (enhanced operational efficiency and reduced resource consumption), and sustainability (limited waste generation and minimum machine strain shown by vibration and temperature levels). The observed result directly reflects the critical performance indicators in a manufacturing setting. For example, the Resource Consumption metric represents the equilibrium between using energy and materials and generating manufacturing output. An improved process is shown by a decrease in consumption coupled with an increase in production. The waste-generated feature directly impacts the sustainability element since a smaller amount of trash is associated with improved environmental and economic results.

5. Conclusion and Future Work. The thorough examination of sophisticated ML models in manufacturing, namely the implementation of EcoEfficientNet, has uncovered a significant improvement in sustainable manufacturing processes. The DL framework of EcoEfficientNet, which combines CNN, LSTM, and FCL, has shown remarkable effectiveness in waste reduction. Its exceptional OEE score and minimum RMSE values support this, showcasing its superior prediction accuracy and operational efficiency. Compared to other models such as GBM, SR, MLP, and RF, EcoEfficientNet surpasses them due to its specialized skills in handling intricate industrial datasets. The tabulated data from IoT devices for 24 hours provides more evidence of how EcoEfficientNet enhances essential performance parameters. It achieves a harmonious combination of high productivity and sustainability by minimizing resource use and waste production, all while ensuring the machine's well-being. The empirical findings, which demonstrate a substantial decrease in waste before and after adopting EcoEfficientNet, provide evidence of the model's strength in promoting an environmentally friendly, efficient, and economically sustainable IoT-based industrial setting. ML in this paradigm shift is crucial for enterprises that want optimal efficiency while maintaining environmental integrity. This advancement sets the stage for an eventuality wherein sustainable manufacturing becomes the standard.

Subsequent investigations in this field aim to combine diverse IoT datasets [29] with EcoEfficientNet to enhance the agility and reactivity of industrial processes. Investigating the integration of blockchain technology for reliable and transparent monitoring of supply chains, together with AI-powered predictive maintenance, has the potential to improve productivity and sustainability.

REFERENCES

- [1] HARIYANI, D., MISHRA, S., HARIYANI, P., & SHARMA, M. K., *Drivers and motives for sustainable manufacturing system*. Innovation and Green Development, 2(1), 2023, 100031.
- [2] SALEM, K. S., CLAYSON, K., SALAS, M., HAQUE, N., RAO, R., AGATE, S., ... & PAL, L., *A critical review of existing and emerging technologies and systems to optimize solid waste management for feedstocks and energy conversion*. Matter, 2023.
- [3] CLANCY, R., O'SULLIVAN, D., & BRUTON, K., *Data-driven quality improvement approach to reducing waste in manufacturing*. The TQM Journal, 35(1), 51-72, 2023.
- [4] CARVALHO, T. P., SOARES, F. A., VITA, R., FRANCISCO, R. D. P., BASTO, J. P., & ALCALÁ, S. G., *A systematic literature review of machine learning methods applied to predictive maintenance*. Computers & Industrial Engineering, 137, 106024, 2019.
- [5] HESSER, D. F., & MARKERT, B. (2019). TOOL WEAR MONITORING OF A RETROFITTED CNC MILLING MACHINE USING ARTIFICIAL NEURAL NETWORKS. Manufacturing letters, 19, 1-4.
- [6] SCALABRINI SAMPAIO, G., VALLIM FILHO, A. R. D. A., SANTOS DA SILVA, L., & AUGUSTO DA SILVA, L., *Prediction of motor failure time using an artificial neural network*. Sensors, 19(19), 4342, 2019.
- [7] FALAMARZI, A., MORIDPOUR, S., NAZEM, M., & CHERAGHI, S., *Prediction of tram track gauge deviation using artificial neural network and support vector regression*. Australian Journal of Civil Engineering, 17(1), 63-71, 2019.
- [8] BISWAL, S., & SABAREESH, G. R., *Design and development of a wind turbine test rig for condition monitoring studies*. In 2015 international conference on industrial instrumentation and control (icic) (pp. 891-896). IEEE, 2015.
- [9] ZHANG, Y., LIU, J., HANACHI, H., YU, X., & YANG, Y., . *Physics-based model and neural network model for monitoring starter degradation of APU*. In 2018 IEEE international conference on prognostics and health management (ICPHM) (pp. 1-7). IEEE, 2018.
- [10] XIANG, S., HUANG, D., & LI, X., *A generalized predictive framework for data driven prognostics and diagnostics using machine logs*. In TENCON 2018-2018 IEEE region 10 conference (pp. 0695-0700). IEEE, 2018.
- [11] ZENISEK, J., HOLZINGER, F., & AFFENZELLER, M., *Machine learning based concept drift detection for predictive maintenance*. Computers & Industrial Engineering, 137, 106031, 2019.
- [12] AUCASIME-GONZALES, P., TREMOLADA-CRUZ, S., CHAVEZ-SORIANO, P., DOMINGUEZ, F., & RAYMUNDO, C., *Waste Elimination Model Based on Lean Manufacturing and Lean Maintenance to Increase Efficiency in the Manufacturing Industry*. IOP Conference Series: Materials Science and Engineering, 999, 012013, 2020. <https://doi.org/10.1088/1757-899x/999/1/012013>
- [13] YANG, X., ZHU, K., TANG, X., WANG, M., ZHAN, M., LU, N., ... & SUN, N., *An in-memory-computing charge-domain ternary CNN classifier*. IEEE Journal of Solid-State Circuits, 2023.
- [14] LIU, C., ZHU, H., TANG, D., NIE, Q., LI, S., ZHANG, Y., & LIU, X., *A transfer learning CNN-LSTM network-based production progress prediction approach in IIoT-enabled manufacturing*. International Journal of Production Research, 61(12), 4045-4068, 2023.
- [15] YANG, C. L., YILMA, A. A., SUTRISNO, H., WOLDEGIORGIS, B. H., & NGUYEN, T. P. Q., *LSTM-based framework with metaheuristic optimizer for manufacturing process monitoring*. Alexandria Engineering Journal, 83, 43-52.
- [16] SCABINI, L. F., & BRUNO, O. M., *Structure and performance of fully connected neural networks: Emerging complex network properties*. Physica A: Statistical Mechanics and its Applications, 615, 128585, 2023.

- [17] BOUGHAMMOURA, A., *A two-step rule for backpropagation*. International Journal of Informatics and Applied Mathematics, 6(1), 57-69, 2023.
- [18] HE, J., ZHAO, H., ZHOU, D., & GU, Q., *Nearly minimax optimal reinforcement learning for linear markov decision processes*. In International Conference on Machine Learning (pp. 12790-12822). PMLR, 2023.
- [19] JIA, Y., & ZHOU, X. Y., *q-Learning in continuous time*. Journal of Machine Learning Research, 24(161), 1-61, 2023.
- [20] BAYRAKTAR, E., & KARA, A. D., *Approximate q learning for controlled diffusion processes and its near optimality*. SIAM Journal on Mathematics of Data Science, 5(3), 615-638, 2023.
- [21] DING, W., JIANG, S., CHEN, H. W., & CHEN, M. S., *Incremental reinforcement learning with dual-adaptive ϵ -greedy exploration*. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 6, pp. 7387-7395), 2023.
- [22] KIM, S., JANG, M. G., & KIM, J. K., *Process design and optimization of single mixed-refrigerant processes with the application of deep reinforcement learning*. Applied Thermal Engineering, 223, 120038, 2023.
- [23] PAGAN, N., BAUMANN, J., ELOKDA, E., DE PASQUALE, G., BOLOGNANI, S., & HANNÁK, A., *A Classification of Feedback Loops and Their Relation to Biases in Automated Decision-Making Systems*. arXiv preprint arXiv:2305.06055, 2023.
- [24] <https://iee-dataport.org/documents/automativemanufacturingdataset>
- [25] SATHLER, K. P. B., SALONITIS, K., & KOLIOS, A., *Overall equipment effectiveness as a metric for assessing operational losses in wind farms: a critical review of literature*. International Journal of Sustainable Energy, 42(1), 374-396, 2023.
- [26] AHMAR, A. S., *Forecast Error Calculation with Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE)*. JINAV: Journal of Information and Visualization, 1(2), 9496, 2020.
- [27] NAZAR, S., YANG, J., AMIN, M. N., KHAN, K., JAVED, M. F., & ALTHOEY, F., *Formulation of estimation models for the compressive strength of concrete mixed with nanosilica and carbon nanotubes*. Developments in the Built Environment, 13, 100113, 2023.
- [28] HELMY, S. H., TAHWIA, A. M., MAHDY, M. G., & ABD ELRAHMAN, M., *Development and characterization of sustainable concrete incorporating a high volume of industrial waste materials*. Construction and Building Materials, 365, 130160, 2023.
- [29] CHANDAN K. SAHU, CRYSTAL YOUNG & RAHUL RAI, *Artificial intelligence (AI) in augmented reality (AR)-assisted manufacturing applications: a review*, International Journal of Production Research, 59:16, 4903-4959,2021. DOI: 10.1080/00207543.2020.1859636

Edited by: Dhilip Kumar V

Special issue on: Unleashing the power of Edge AI for Scalable Image and Video Processing

Received: Jan 16, 2024

Accepted: Jul 4, 2024



A VISION-BASED ANALOG METER READING METHOD FOR INSPECTION ROBOTS

JIACHENG LI^{*} HONGLEI WANG[†] XISHUO ZHU[‡] SIJIAN LIU[§] AND JUNSHENG ZHANG[¶]

Abstract. Computer vision technology has been widely applied in reading recognition of analog meters. However, it is still a challenge to quickly and accurately read various types of analog meters under different environmental conditions. We propose a fast-reading method for analog meters based on keypoint detection, which is applied to inspection robots. First, we use the YOLOv5s network to locate the analog meter. Second, the HRNet network is used to detect the keypoints of the pointer and scale on the dial. Third, an objective image quality assessment method that includes multiple indicators is established to select the optimal image for reading recognition. Finally, we calculate the reading of the analog meter based on the deflection angle of the pointer. The experiment shows that our method can accurately read the readings of analog meters, with an average reading error of 3.81%. It can be effectively applied to inspection robots to read analog meter readings.

Key words: analog meter; object detection; keypoint detection; inspection robot; reading recognition

1. Introduction. Many places that require accurate and reliable measurements, such as substations, chemical factories, water pump houses, and other similar locations, still use analog meters instead of digital ones. This is because analog meters are less affected by the electromagnetic environment and can provide more stable readings. However, a drawback of these meters is that they usually cannot send analog signals to a remote location for monitoring or analysis. Therefore, they depend on manual inspection by workers who have to read the analog meters in person [1]. With the developments of computer vision technology [2], it has become the mainstream method to use an inspection robot equipped with a camera for instrument recognition. Using computer vision to automatically read analog meters is more efficient and convenient than manual inspection. However, the images collected by the visible light cameras are easily affected by different poses, illumination changes, and complex backgrounds, which brings difficulties to reading recognition. Many researchers have explored vision-based methods for fast and stable instrument recognition by inspection robots [3-6].

Traditional methods for machine vision reading [7-8] have some drawbacks, which prevent them from being stably applied. These methods usually identify the dial area in an image by using template matching or Hough circle detection. Then they recognize the scale and pointer in the dial by using image segmentation or line detection. Finally, they read the meter based on the angle or distance between the pointer and the scale [9]. However, these methods are not suitable for inspection robots that work in complex and changing environments [10]. For example, they cannot adapt to different angles and lighting conditions because they rely on manually set parameters in the algorithm. This makes their detection robustness very poor.

In recent years, deep learning has been applied to analog meter recognition with remarkable results [11-14]. One of the key steps in analog meter recognition is instrument detection, which involves locating the instrument region in an image [15]. Huang et al. [17] employed an improved YOLOv3 network to locate the analog meter and a monocular-vision pointer reconstruction algorithm was used to read the instrument. In [18], a Faster Region-based Convolutional Network (Faster R-CNN) was used to detect the target meter and guide the camera

^{*}Beijing Technology Research Branch of Tiandi Science & Technology Co., Ltd., Beijing 100013, China; Intelligent Mine Research Institute, Chinese Institute of Coal Science (CICS), Beijing, 100013, China

[†]Beijing Technology Research Branch of Tiandi Science & Technology Co., Ltd., Beijing 100013, China; Intelligent Mine Research Institute, Chinese Institute of Coal Science (CICS), Beijing, 100013, China (Corresponding author, wanghonglei@mail.ccric.cteg.cn)

[‡]Beijing Technology Research Branch of Tiandi Science & Technology Co., Ltd., Beijing 100013, China; Intelligent Mine Research Institute, Chinese Institute of Coal Science (CICS), Beijing, 100013, China

[§]Beijing PINS Medical Co., Ltd., Beijing, 102200, China

[¶]Beijing Technology Research Branch of Tiandi Science & Technology Co., Ltd., Beijing 100013, China; Intelligent Mine Research Institute, Chinese Institute of Coal Science (CICS), Beijing, 100013, China

alignment. Then reading could be obtained after the pointer is recognized by Hough Transform. Salomon et al. [19] used a YOLOv4 target detector and proposed a new regression method (AngReg) to achieve automatic meter reading in unconstrained scenarios. Fan et al. [20] proposed an adaptive anchor and global context (GC) module for meter detection, which combines deep learning methods and traditional computer vision methods to achieve power equipment meter recognition. The above methods adopt the deep learning method in the instrument detection stage but are not applied to the identification of pointer and scale, resulting in limited improvement of reading robustness.

Compared to meter detection, reading recognition is a more complex task. Especially for such tasks as patrol robot automatic meter reading, the scene of the image to be identified usually has a large difference due to background changes, robot movement, and other factors. The focus of research is to stably recognize the dial information of meters for reading under different conditions. The deep learning-based image segmentation technology has been applied to detect scales and pointers, leading to a significant improvement in the robustness of the reading system. Alexeev et al. [21] proposed a three-level model to regress the coordinates of the keypoints in the panel but did not investigate the pointer recognition and reading strategies. Zuo et al. [22] improved Mask-RCNN to extract binary marks of the instrument dial and pointer. In [23], a modified RFB-Net network was used to detect the keypoint of the pointer image. Dong et al. [24] developed a vector detection network to recognize the direction of the pointer. Compared to traditional methods, these CNN-based pointer and dial detection methods greatly reduced the probability of missed detection and false detection, enabling the recognition of various analog meters in dynamic scenes. However, no optimal general reading method exists for the many types of analog meters for electric power and pressure measurement. Deng et al. [25] proposed a meter reading method that combines YOLOv5s and DeeplabV3+ networks. The segmented dashboard area is unfolded, and the reading is calculated by calculating the distance between the pointer and the scale. Zhou et al. [26] proposed an end-to-end pointer meter reading method based on deep learning. Simultaneously locate the pointer and extract the pointer object and achieve meter reading recognition through the local angle method. These reading methods, which also apply deep learning technology in the reading stage after the dial detection and positioning, have achieved good results in terms of reading accuracy and robustness. But training a model that can segment all the scale and pointer information on the dial requires a large amount of annotated data. And it requires a more complex post-processing process to determine the relationship between the pointer and the scale.

The acquired meter images may be blurred and distorted due to environmental factors and the robots motion factors [27]. It is better to use image quality assessment to filter out the images with good imaging effects for reading recognition. Image quality assessment methods are categorized into subjective and objective assessment. Objective assessment methods give a quantitative result based on a mathematical model. A well-designed objective assessment method can be very close to the subjective judgment of a human being. It can effectively screen out images with more distinctive features for further processing before detection.

In this study, a fast-reading method based on object and keypoint detection is proposed to read analog meters for inspection robots in motion. Our reading method consists of a target detection model and a keypoint detection model. First, a YOLOv5s [28] target detection network is used to detect analog meters in complex patrol scenes. Second, inspired by the human posture estimation algorithm, the High-Resolution Net (HRNet) [29] is used to extract the scale and pointer information of the detected analog meter panel. Third, the objective image quality assessment method is used to evaluate the meter images collected by inspection robots, and the best quality images are selected for reading calculation. Finally, based on the results of instrument target detection and keypoint detection, the reading is calculated by the angle of the pointer.

The primary contributions of our study are as follows.

1. The YOLOv5s network is used for instrument detection, which is suitable for the complex working scene of patrol robots.
2. A keypoint detection method is used for the detection of analog meter range and pointer, which improves the reading recognition effect.
3. A meter recognition method for inspection robots in motion is proposed, which selects the best quality image from consecutive frames of the video stream and accurately recognizes circular simulated meter readings.

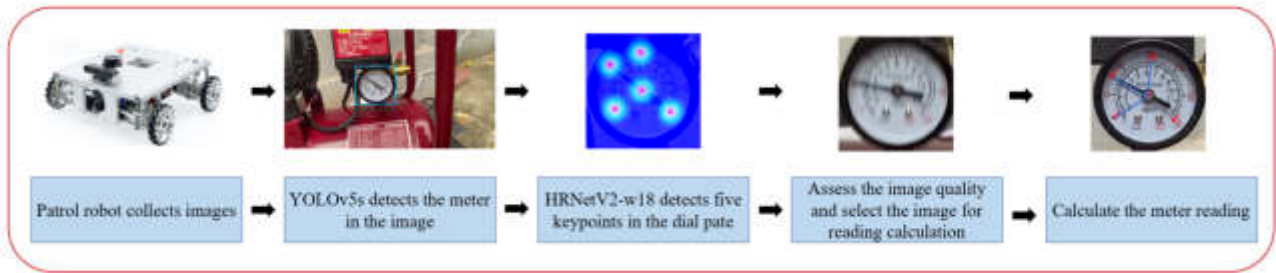


Fig. 2.1: Flowchart of the proposed reading method.

4. The effect of the proposed reading method is verified on the real inspection robot platform.

2. Proposed Method. We propose a deep learning-based method for fast meter reading. Fig. 2.1 illustrates its flowchart. First, the inspection robot performs analog meter target detection under the moving state. Real-time video data can be processed directly through the Nvidia Jetson edge computing device mounted on the robot. Or transmit it to the computer for processing through network streams such as RTSP and RTMP. The YOLOv5s detection model is used to detect and mark the pointer meters in the field of view of the inspection robot in real-time. Second, inspired by the method of human posture estimation, the HRNet network is modified to detect five keypoints in the dashboard: the start and end of the range, the midpoint of the range, the center of the dial, and the end of the pointer. Third, an objective image quality assessment method that combines information entropy, standard deviation, and the mean gradient is used to select the optimal quality image for reading recognition. Finally, a method based on pointer deflection angle calculation was used for meter reading recognition.

2.1. Analog Meter Detector Based on Yolov5s. The key to accurately reading a pointer instrument is to identify and locate the instrument panel area in the image. It is almost impossible for the traditional target detection method to recognize the instrument panel stably and accurately under the complex and changeable image background conditions collected by the inspection robot. The target detection method based on deep learning makes this application possible. The target detection algorithm can be divided into two types: one-stage method and two-stage method. The one-stage method represented by YOLO [30] has high real-time performance, but it is usually slightly inferior to the recognition accuracy of the two-stage target detection method.

YOLO (You Look Only Once) is a one-stage target detection method that predicts bounding boxes and class probabilities directly from an input image. YOLOv5s has the smallest network depth and the smallest feature map width among four variants: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. It is mainly composed of input, backbone, neck, and prediction. The input part enriches the data set by splicing training data to achieve a better detection effect with smaller samples. The backbone part is mainly composed of a Cross Stage Partial (CSP) module for feature extraction. Feature Pyramid Network (FPN) and Path Aggregation Network (PANet) are used in the neck part to aggregate the image features preliminarily extracted from the backbone part. The prediction part performs target prediction and outputs the prediction results.

The target detection of the analog meter using YOLOv5s mainly includes four steps. First, collect and label the training images with bounding boxes around pointers. Second, build the training set, validation set, and test set of YOLOv5s network according to a certain ratio. Third, train the target detection model using a deep learning framework with appropriate hyperparameters. Finally, test the model recognition effect by running inference on test set images and evaluating metrics such as precision and recall.

We use the mosaic data augmentation method to improve the accuracy of YOLOv5s target detection model with fewer labeled samples. This method randomly crops, rotates, scales, and color transforms four images, and then combines them into one image as a training input sample. It makes the scale and color space of the sample more variety, and makes the limited data generate more value. Figure 2.3 illustrates how mosaic data augmentation method improves the pointer meter data. This method helps the model learn from different

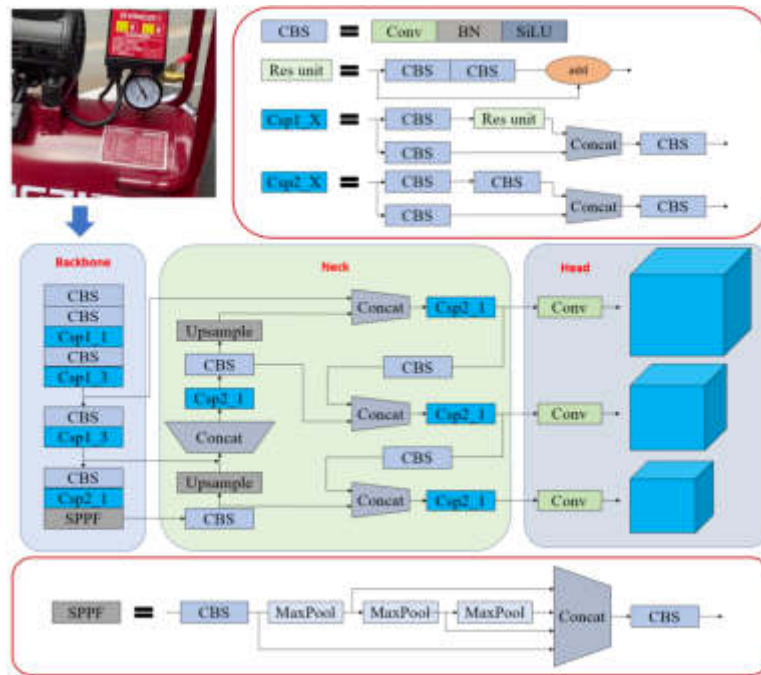


Fig. 2.2: The YOLOv5s framework.



Fig. 2.3: Mosaic data augmentation.

contexts and scales of the pointer meter images. It has been verified that this method can effectively enhance the model’s ability to generalize to unseen data.

For the robot that checks along the preset route, the pointer meter usually appears from one end of the captured image and disappears from the field of vision from the other end as the robot moves. Obviously, for the reading calculation of the pointer meter, the image collected when the camera is facing the instrument can get a more accurate reading. For each detected target, YOLOv5s outputs the coordinates of the upper left

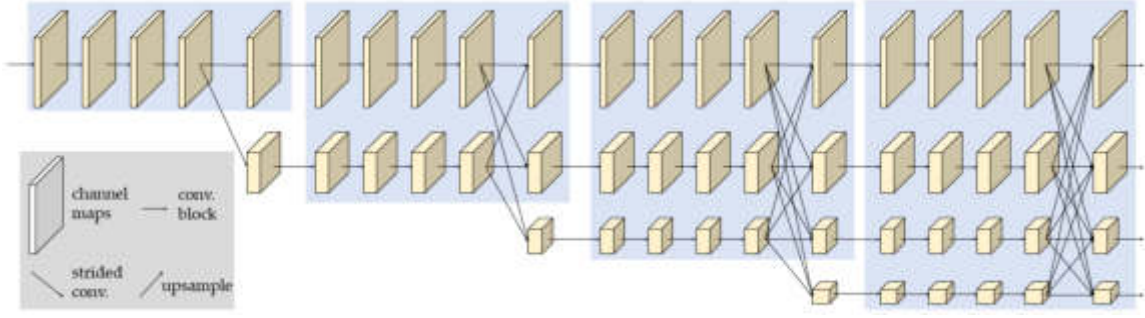


Fig. 2.4: Backbone structure of HRNet.

corner (x_1, y_1) and the lower right corner (x_2, y_2) of the bounding box to achieve the positioning of the pointer meter. R_p is calculated as shown in Equation (2.1). The relative position R_p of the center point of the bounding box in the whole image is used to determine the angle between the camera and the pointer instrument. The closer the R_p is to 0.5, the higher the alignment between the camera and the target pointer gauge.

$$R_p = \frac{x_2 + x_1}{2w} \quad (2.1)$$

where w represents the width of the whole image.

2.2. Keypoint Detector Based on HRNetV2-w18. When the instrument in the picture is detected, the most critical step is to identify the scale and pointer. We can use angle calculation or distance calculation methods for reading recognition of pointer instrument. These methods have different pros and cons. The distance calculation method needs higher accuracy of scale and pointer position recognition and gives higher reading accuracy, but it also demands higher quality of collected images. The angle calculation method is more robust and can be more precise even with unstable image quality.

HRNet is designed to preserve high-resolution features throughout the network by using multiple parallel branches with different resolutions. The structure of HRNet is shown in Figure 2.4. The branches are connected by repeated multi-scale fusion modules that allow information exchange across resolutions. HRNet has several variants with different depths and widths to suit different applications and resources. HRNetV2-w18 is a variant that has a smaller width than other versions. For inspection robots that need to read pointer meters while moving, detection speed and accuracy are crucial. The HRNetV2-w18 network can achieve a good balance between these two factors.

In the keypoint detection task, the position learning of key points can be divided into two categories: heatmap-based and region-based. The keypoints in the image cannot be represented by a single pixel, which may be composed of the labeled coordinates and some nearby pixels. If all the pixels except the marker coordinates are defined as negative samples, the model may be difficult to converge. We use the heatmap method to obtain higher coordinate prediction accuracy. The generation method of heatmap is shown in Equation (2.2).

$$Heatmap(x, y) = e^{-\frac{(x-\mu_x)^2 + (y-\mu_y)^2}{2\sigma^2}} \quad (2.2)$$

where μ_x and μ_y represent the true keypoint coordinates obtained from annotation, and $Heatmap(x, y)$ is the value of a point (x, y) on the heatmap.

2.3. Objective Image Quality Assessment. We propose a reference-free image quality assessment method for images acquired by inspection robots. The calculated metrics include entropy, standard deviation and mean gradient. Entropy measures the amount of information contained in an image; higher information entropy indicates that the image contains more information, and the image is usually more detailed. The image



Fig. 2.5: Reading calculation methods in two cases: (a) The pointer is in the first half of the range; (b) The pointer is in the second half of the range.

is first converted to a gray level image and the probability of occurrence of each gray level is counted to finally get the information entropy. The entropy is calculated as shown in Equation (2.3).

$$H = - \sum_{i=0}^{255} p_i \log_2 p_i \quad (2.3)$$

where p_i is the probability of occurrence of the corresponding gray level.

The standard deviation reflects the degree of dispersion between the pixel values and the mean value of the image, and a larger standard deviation usually indicates a better image quality. The standard deviation is calculated as shown in Equation (2.4).

$$\sigma = \sqrt{\frac{1}{w * h} \sum_{i=1}^w \sum_{j=1}^h (p_{ij} - \mu)^2} \quad (2.4)$$

where w and h are the width and height of the image, p_{ij} is the pixel value of the corresponding coordinate and μ is the mean value of the image.

The mean gradient reflects the rate of change of the gray values on both sides of the image edges, this data can be used to measure the fineness of the image details, the larger the mean gradient is, the clearer the image is in general. The mean gradient is calculated as shown in Equation (2.5).

$$G = \frac{1}{w * h} \sum_{i=1}^w \sum_{j=1}^h \sqrt{\frac{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2}{2}} \quad (2.5)$$

where w and h are the width and height of the image, $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ are the gradients of the image horizontally and vertically.

We normalize the above three metrics by mapping them between 0 and 1. The number of our image dataset is small, and the normalization is performed by using between the maximum and minimum values. After actual testing, here we no longer assign different weights to the above three metrics, and use the score obtained by Equation (2.6) as the final assessment of image quality. In the application, images with meter alignment values R_p greater than 0.35 and less than 0.65 will be calculated for image quality, and the reading with the highest image quality score will be used as the reading for that meter.

$$Score = H + \sigma + G \quad (2.6)$$

2.4. Reading Calculation Based on Keypoints. Upon acquisition of precise dial scale and pointer end coordinates, determination of the rotation angle of the pointer enables computation of the reading of the meter. Figure 2.5 illustrates two distinct methods for computing the reading of the meter, which depend on the position of the pointer relative to the scale. If the pointer falls between the starting point and the midpoint of



Fig. 3.1: Examples of pointer meter annotations. (a) Instrument A with a maximum range of 180 Lb/inš; (b) Instrument B with a maximum range of 0.6 MPa; (c) Instrument C with a maximum range of 1.6 MPa.

the scale, the reading is determined using these two reference points. In contrast, if the pointer has surpassed the midpoint and is in closer proximity to the endpoint, the reading is calculated using the midpoint and endpoint of the scale.

In Figure 2.5(a), the position of the pointer falls within the initial half of the meter’s range, whereby the vectors \overrightarrow{CS} and \overrightarrow{CM} , represented by points 4 and 1 and points 4 and 2 respectively, serve as the starting and ending points for the beginning and end of the range. To determine half of the full range of the meter, the angle α enclosed between the two vectors is computed. To calculate the meter reading, vector \overrightarrow{CP} is constructed using points 4 and 5 as the starting and ending points for representation of the pointer. The included angle β between vector \overrightarrow{CP} and \overrightarrow{CS} is then ascertained. The reading of the meter can be calculated using Equation (2.7).

$$Numf = \frac{90\alpha}{\beta} \quad (2.7)$$

In Figure 2.5(b), the pointer is situated in the second half of the range. The vector \overrightarrow{CM} with points 4 and 2 as the starting and ending points signifies the beginning range, whereas the vector \overrightarrow{CE} , represented by points 4 and 3, corresponds to the end of the range. Vector \overrightarrow{CP} , constructed with points 4 and 5, portrays the pointer. To determine the meter reading, the enclosed angles α and β between vector \overrightarrow{CP} and \overrightarrow{CM} , and \overrightarrow{CM} and \overrightarrow{CE} , respectively, are calculated. The reading of the meter can be calculated using Equation (2.8).

$$Nume = \frac{90(\alpha + \beta)}{\beta} \quad (2.8)$$

3. Experiments and Results.

3.1. Experimental Environment and Dataset. We ran the proposed algorithm on a computer with I7-13900H CPU, 32 GB RAM and NVIDIA GeForce RTX4060 8GB GPU. We used PyTorch deep learning framework, Python 3.10 programming language, CUDA 11.8 and cuDNN 8.9 NVIDIA acceleration tools.

We collected 1632 images of three kinds of air pressure pointer meters using an industrial camera with a resolution of 1920E1080 pixels. We labeled the images with LabelMe, an open-source annotation tool. The position of the pointer in the image and the five key points in the dial are marked at one time. The position of the pointer meter is marked with the bounding box. The start of the range is marked with 1, the midpoint of the range is marked with 2, the end of the range is marked with 3, the center of the dial is marked with 4, and the end of the pointer is marked with 5. Figure 3.1 shows examples of pointer meter annotations in the dataset.

The marked data is further processed to provide the YOLOv5s target detection model and HRNet key point detection model for training. The bounding box and category information marked by LabelMe software are extracted and stored in the format of VOC data set and provided to YOLOv5s model. The heatmap files, which are produced using the methodology outlined in Section 2.2, are provided to HRNetV2-w18 for training a keypoint detection model. Table 3.1 shows the number of different keypoints marked on the pointer meter.

Table 3.1: Definition of keypoints of the meter.

Label number	Position of the pointer meter
1	The start of the range.
2	The midpoint of the range.
3	The end of the range.
4	The center of the dial.
5	The end of the pointer.

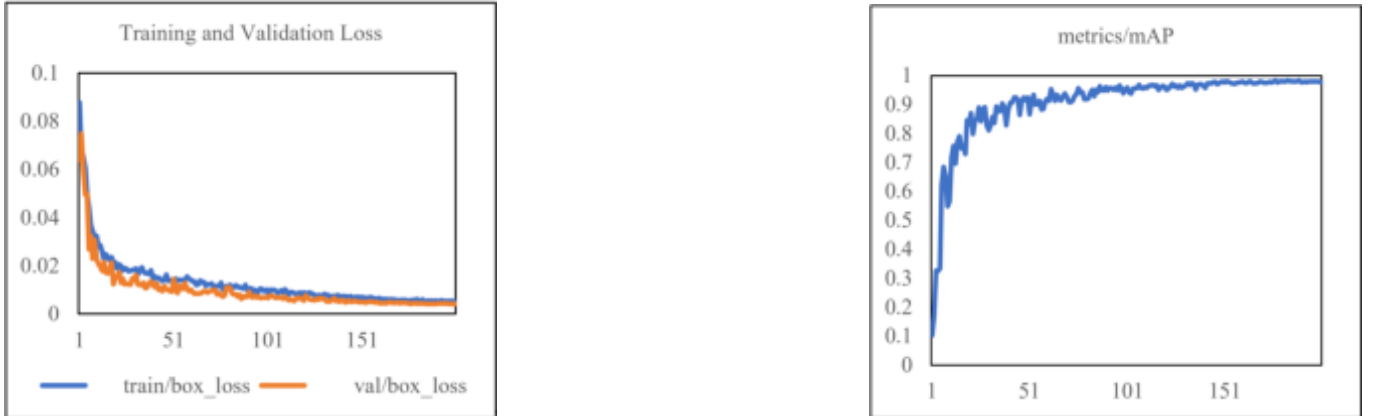


Fig. 3.2: Loss and mAP variation curves. (a) The loss variation curves for the training and validation sets; (b) mAP variation curve.

3.2. Experimental Results.

3.2.1. YOLOv5s Analog Meter Detector. The dataset was partitioned into three subsets, namely training, validation, and testing, in an 8:1:1 ratio. We trained YOLOv5s model with an input image resolution of 640×640 pixels, a batch size of 8, an initial learning rate of 0.001, and a training duration of 200 epochs. Due to the small size of the dataset, we utilized a pre-trained model on the COCO dataset provided by the official source to initialize the weights of our model, aiming to achieve better recognition performance. Figure 3.2(a) and Figure 3.2(b) illustrate the variation of loss and mean average precision (mAP) during the training process of the YOLOv5s object detection model. From Figure 3.2(a), it can be observed that the loss for both the training and validation datasets gradually decreases as the training progresses. After the 150th epoch, the loss stabilizes. Figure 3.2(b) presents that the model achieves a high mAP of 98.3% at last.

On the self-built pointer instrument dataset, YOLOv5s was compared with other common models. The training parameter settings remained unchanged from the previous text. The performance of the optimal weight of each model on the test set is shown in Table 3.2, and YOLOv5s achieves a balance between accuracy and detection speed. The average accuracy has reached 97.9%, and the detection speed can reach 33.78 FPS.

3.2.2. HRNetV2-w18 Keypoint Detector. The circular meter area in the image is obtained by the target detector and adjusted to 384×384 pixels before being input into the HRNetV2-w18 keypoints detection model. The batch size is set to 1, and stochastic gradient descent (SGD) is used for parameter optimization with an initial learning rate of 1e-5 over a total training epoch of 60. The network outputs are heatmaps, and the network’s keypoints detection performance is evaluated using root mean square error (RMSE) after extracting keypoints. Figure 3.3(a) and Figure 3.3(b) show the changes in loss on the training and validation sets during the HRNetV2-w18 training process. It can be seen that after the 100th epoch, the loss basically does not decrease, and the training is completed.

Table 3.2: Comparison of YOLOv5s with Faster R-CNN, SSD and PP-YOLOE-s on the test dataset.

Method	Epochs	mAP	FPS
Faster RCNN	200	0.975	6.71
SSD	200	0.925	26.96
PP-YOLOE-s	200	0.749	18.21
YOLOv5s	200	0.979	33.78

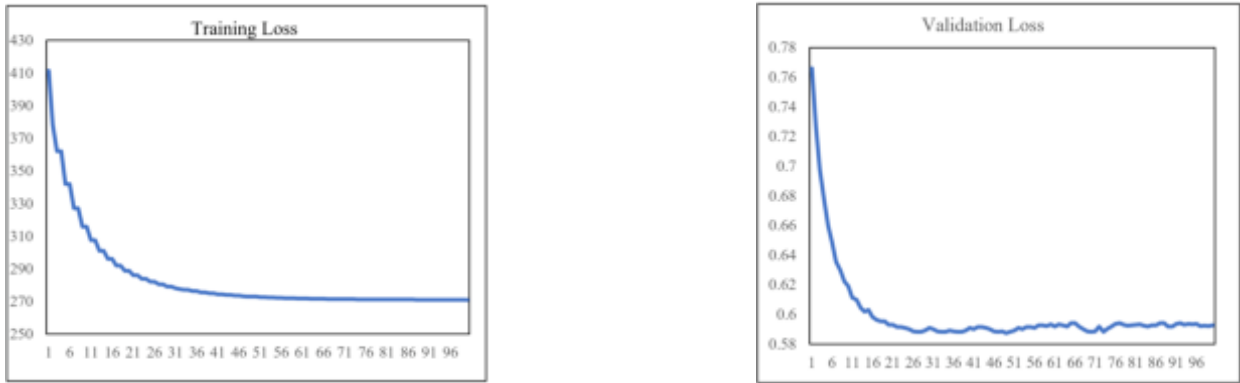


Fig. 3.3: The loss curve of the training and validation sets during the training process. (a) The loss variation curves for the training sets; (b) The loss variation curves for the validation sets.

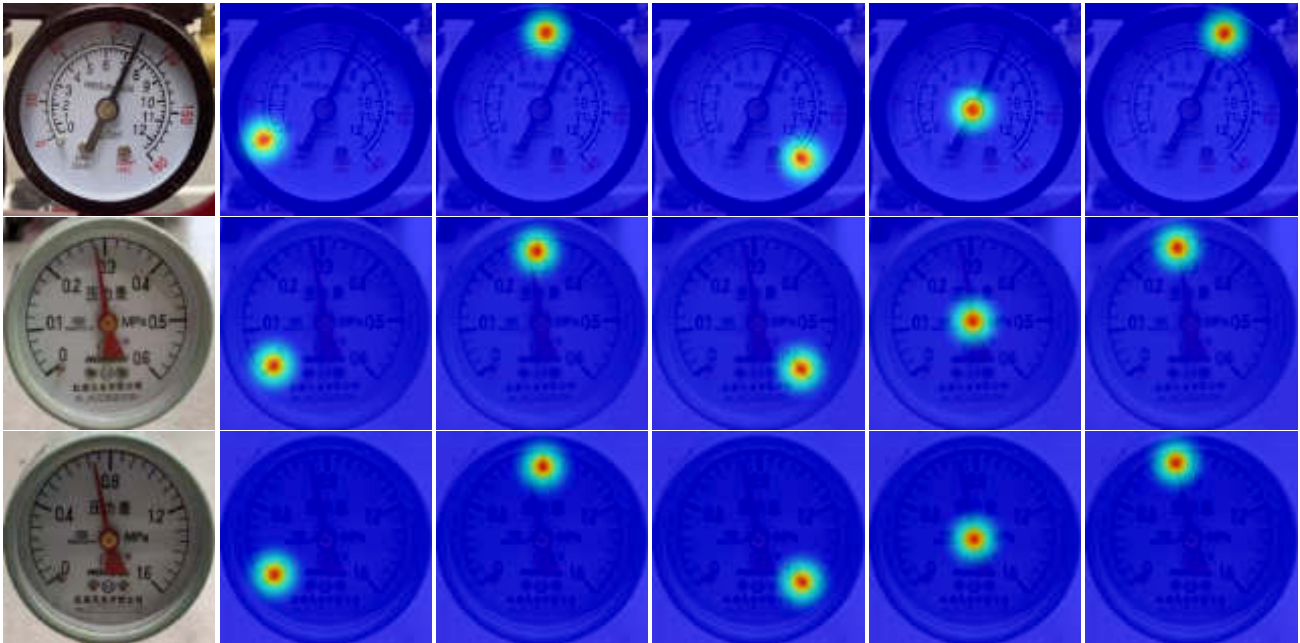


Fig. 3.4: Keypoint detection results. (a) Cropped and resized pointer instrument image detected by YOLOv5s;(b-f) Predicted heatmaps for 5 keypoints.

Table 3.3: Maximum and minimum values of image quality metrics for the dataset.

Image quality metric	Maximum value	Minimum value
Entropy	7.785	5.618
Standard deviation	78.967	13.184
Mean gradient	89.982	7.815

Table 3.4: Maximum and minimum values of image quality metrics for the dataset.





Num.	Image	Entropy	Standard deviation	Mean gradient	Score
1		0.458	0.228	0.126	0.812
2		0.470	0.274	0.190	0.934
3		0.486	0.200	0.310	0.996
4		0.563	0.352	0.359	1.274

Figure 3.4 shows the detection results output by the HRNetV2 w18 model, and the predicted results of the five key points are presented in heatmaps. For three different dial types of pressure gauges, the model can accurately locate the key points of the readings. The key point positions selected by this method have relatively universal characteristics.

3.2.3. Image Quality Assessment. The evaluation method established in the previous section is used to select the image with the best image quality for meter reading recognition. The three metrics of entropy, standard deviation and mean gradient are calculated for all images in the image dataset and their maximum and minimum values are shown in Table 3.3. It can be seen that the contribution of the entropy metrics to the final score may not be fully reflected if normalization is not performed. It can be seen that the contribution of the entropy to the final score may not be fully reflected if normalization is not performed. When the dataset is large enough, the weights of the three metrics contribution to the final score can be further assigned. Here we take the normalized three metrics and add them directly.

Table 3.4 gives the images captured while the inspection robot is moving, and their image quality is evaluated after locating and intercepting the meter dial area images. Higher values of the three image quality metrics and assessment scores indicate better image quality.

3.2.4. Meter Reading Recognition. After detecting the keypoint on the dial of the pointer meter, the reading of the meter can be calculated using the method proposed in Section 2.4. This angle-based reading calculation method has a significant impact on the calculation accuracy by the positioning accuracy of keypoints. Especially when the deviation degree of the center point detection of the dial is relatively high, it will bring significant errors to the reading calculation. However, compared to other pointer meter reading methods based on semantic segmentation, using reading methods based on key points and angles has lower data annotation costs. Adding half range key point annotation will not excessively increase the cost of data annotation but can to some extent improve reading accuracy.

Table 3.5: Reading error of different strategies.

Num.	Ground truth	Full range angle reading results	E_{RF}	Half range angle reading results	E_{RH}
1	30.0	26.8	10.67 %	31.8	6.00%
2	46.8	44.2	5.56%	44.5	4.91%
3	62.4	64.7	3.69%	65.2	4.49%
4	68.4	72.1	5.41%	71.1	3.95%
5	91.2	91.9	0.77%	90.4	0.88%
6	103.2	106.7	3.39%	101.1	2.03%
7	105.6	102.0	3.41%	107.9	2.18%
8	124.8	131.5	5.37%	119.6	4.17%
9	0.168	0.181	7.74%	0.151	10.12%
10	0.204	0.213	4.41%	0.199	2.45%
11	0.284	0.288	1.41%	0.290	2.11%
12	0.340	0.365	7.35%	0.358	5.29%
13	0.392	0.385	1.79%	0.393	0.26%
14	0.466	0.483	3.65%	0.480	3.00%
15	0.26	0.29	11.54%	0.25	3.85%
16	0.56	0.62	10.71%	0.59	5.36%
17	0.70	0.71	1.43%	0.73	4.29%
18	0.92	0.98	6.52%	0.96	4.35%
19	0.99	0.95	4.04%	1.02	3.03%
20	1.18	1.11	5.93%	1.22	3.39%

The comparison between the reading results and the true values using the method presented in this article is shown in Table 3.3. The true value is obtained by manually observing the instrument horizontally from the front. Evaluate the effectiveness of the proposed method by calculating the error between the automatic reading and the actual reading. The calculation formula for reading error E_R is shown in Equation (3.1).

$$E_R = \frac{|T - L|}{T} \times 100\% \quad (3.1)$$

where L is the value obtained by the proposed automatic reading method, and T is the instrument reading obtained through manual observation.

We calculated the readings obtained based on the full range key points and the readings obtained based on the half range keypoints. Then evaluate the accuracy of the readings by calculating their errors using true values. The decimal places of manually observed values are calculated based on 1/5 of the pressure gauge graduation value. Table 3.5 shows the reading error results obtained by different strategies under the same keypoint detection results.

The different angles at which inspection robots collect analog meter images can have an impact on the reading results. The instrument images used for calculating readings in the experiment are collected with the camera horizontally centered on the analog meters. From the results shown in Table 3.3, it can be seen that the reading strategy using half range key points has higher accuracy. From the results shown in Table 3.3, it can be seen that the reading strategy using half range key points has higher accuracy.

4. Discussion. This article proposes a pointer instrument reading method based on key point recognition, inspired by the top-down human posture estimation method. It consists of a target detector and a keypoint detector. This detection method has high accuracy, but its computational complexity increases linearly with the increase of multiple targets in the image. For the application of patrol robots, the accuracy advantage of this method is more prominent. The light target detection network based on YOLOv5s achieves the balance between detection speed and accuracy. It can be easily transplanted to the edge computing equipment carried by the patrol robot. A reading calculation method based on five key point detection can achieve high reading accuracy through finetune with a small amount of annotated data.

However, this is not an end-to-end recognition method. In the future, pointer instrument reading recognition may use multitask learning methods. It can use the network to simultaneously perform object detection and keypoint detection, share the same feature extraction backbone, and improve computational efficiency. To recognize readings from different kinds of meters in the future, the proposed method should also account for pointer meters' range recognition. The meter's measuring range can be either pre-entered as prior knowledge for different types or detected from the digits on the dial.

REFERENCES

- [1] Jaffery, Z.A.; Dubey, A.K. Architecture of noninvasive real time visual monitoring system for dial type measuring instrument. *IEEE Sensors Journal* **2012**, *13*, 1236-1244.
- [2] Zhang, J.; Ge, K.; Xun, L.; Sun, X.; Xiong, W.; Zou, M.; Zhong, J.; Li, T. MFCD-Net: Cross Attention Based Multimodal Fusion Network for DPC Imagery Cloud Detection. *Remote Sensing* **2022**, *14*, 3905.
- [3] Gao, H.; Yi, M.; Yu, J.; Li, J.; Yu, X. Character segmentation-based coarse-fine approach for automobile dashboard detection. *IEEE Transactions on Industrial Informatics* **2019**, *15*, 5413-5424.
- [4] Liu, L.; Qiao, X.; Liang, W.-z.; Oboamah, J.; Wang, J.; Rudnick, D.R.; Yang, H.; Katimbo, A.; Shi, Y. An Edge-computing flow meter reading recognition algorithm optimized for agricultural IoT network. *Smart Agricultural Technology* **2023**, *5*, 100236.
- [5] Chen, Y.-S.; Wang, J.-Y. Computer Vision-Based Approach for Reading Analog Multimeter. *Applied Sciences* **2018**, *8*, 1268.
- [6] Zou, L.; Wang, K.; Wang, X.; Zhang, J.; Li, R.; Wu, Z. Automatic Recognition Reading Method of Pointer Meter Based on YOLOv5-MR Model. *Sensors* **2023**, *23*, 6644.
- [7] Guo, X.; Zhu, Y.; Zhang, J.; Hai, Y.; Ma, X.; Lv, C.; Liu, S. Intelligent pointer meter interconnection solution for data collection in farmlands. *Computers and Electronics in Agriculture* **2021**, *182*, 105985.
- [8] Alegria, E.C.; Serra, A.C. Automatic calibration of analog and digital measuring instruments using computer vision. *IEEE transactions on instrumentation and measurement* **2000**, *49*, 94-99.
- [9] Hou, L.; Qu, H. Automatic recognition system of pointer meters based on lightweight CNN and WSNs with on-sensor image processing. *Measurement* **2021**, *183*, 109819.
- [10] Li, D.; Li, W.; Yu, X.; Gao, Q.; Song, Y. Automatic Reading Algorithm of Substation Dial Gauges Based on Coordinate Positioning. *Applied Sciences* **2021**, *11*, 6059.
- [11] Li, Z.; Zhou, Y.; Sheng, Q.; Chen, K.; Huang, J. A high-robust automatic reading algorithm of pointer meters based on text detection. *Sensors* **2020**, *20*, 5946.
- [12] Cai, W.; Ma, B.; Zhang, L.; Han, Y. A pointer meter recognition method based on virtual sample generation technology. *Measurement* **2020**, *163*, 107962.
- [13] Laroca, R.; Barroso, V.; Diniz, M.A.; Gonçalves, G.R.; Schwartz, W.R.; Menotti, D. Convolutional neural networks for automatic meter reading. *Journal of Electronic Imaging* **2019**, *28*, 013023-013023.
- [14] Li, T.; Meng, Z.; Ni, B.; Shen, J.; Wang, M. Robust geometric ℓ_p -norm feature pooling for image classification and action recognition. *Image and Vision Computing* **2016**, *55*, 64-76.
- [15] Zhang, C.; Shi, L.; Zhang, D.; Ke, T.; Li, J. Pointer Meter Recognition Method Based on Yolov7 and Hough Transform. *Applied Sciences* **2023**, *13*, 8722.
- [16] Zhang, Z.; Hua, Z.; Tang, Y.; Zhang, Y.; Lu, W.; Dai, C. Recognition method of digital meter readings in substation based on connected domain analysis algorithm. *Actuators* **2021**, *10*, 170.
- [17] Huang, J.; Wang, J.; Tan, Y.; Wu, D.; Cao, Y. An automatic analog instrument reading system using computer vision and inspection robot. *IEEE Transactions on Instrumentation and Measurement* **2020**, *69*, 6322-6335.
- [18] Liu, Y.; Liu, J.; Ke, Y. A detection and recognition system of pointer meters in substations based on computer vision. *Measurement* **2020**, *152*, 107333.
- [19] Salomon, G.; Laroca, R.; Menotti, D. Image-based automatic dial meter reading in unconstrained scenarios. *Measurement* **2022**, *204*, 112025.
- [20] Fan, Z.; Shi, L.; Xi, C.; Wang, H.; Wang, S.; Wu, G. Real time power equipment meter recognition based on deep learning. *IEEE Transactions on Instrumentation and Measurement* **2022**, *71*, 1-15.
- [21] Alexeev, A.; Kukharev, G.; Matveev, Y.; Matveev, A. A highly efficient neural network solution for automated detection of pointer meters with different analog scales operating in different conditions. *Mathematics* **2020**, *8*, 1104.
- [22] Zuo, L.; He, P.; Zhang, C.; Zhang, Z. A robust approach to reading recognition of pointer meters based on improved mask-RCNN. *Neurocomputing* **2020**, *388*, 90-101.
- [23] Zhang, Q.; Bao, X.; Wu, B.; Tu, X.; Jin, Y.; Luo, Y.; Zhang, N. Water meter pointer reading recognition method based on target-key point detection. *Flow Measurement and Instrumentation* **2021**, *81*, 102012.
- [24] Dong, Z.; Gao, Y.; Yan, Y.; Chen, F. Vector detection network: An application study on robots reading analog meters in the wild. *IEEE Transactions on Artificial Intelligence* **2021**, *2*, 394-403.
- [25] Deng, G.; Huang, T.; Lin, B.; Liu, H.; Yang, R.; Jing, W. Automatic meter reading from UAV inspection photos in the substation by combining YOLOv5s and DeepLabv3+. *Sensors* **2022**, *22*, 7090.
- [26] Zhou, D.; Yang, Y.; Zhu, J.; Wang, K. Intelligent reading recognition method of a pointer meter based on deep learning in a real environment. *Measurement Science and Technology* **2022**, *33*, 055021.
- [27] Wu, X.; Shi, X.; Jiang, Y.; Gong, J. A high-precision automatic pointer meter reading system in low-light environment.

- Sensors* **2021**, *21*, 4891.
- [28] Jocher, G.; Stoken, A.; Borovec, J.; Changyu, L.; Hogan, A.; Chaurasia, A.; Diaconu, L.; Ingham, F.; Colmagro, A.; Ye, H. ultralytics/yolov5: v4.0-mn. SiLU () activations, Weights & Biases logging, PyTorch Hub integration. Zenodo **2021**.
- [29] Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, **2019**; pp. 5693-5703.
- [30] Yin, H.; Chen, M.; Fan, W.; Jin, Y.; Hassan, S.G.; Liu, S. Efficient Smoke Detection Based on YOLO v5s. *Mathematics* **2022**, *10*, 3493.

Edited by: Bradha Madhavan

Special issue on: High-performance Computing Algorithms for Material Sciences

Received: Dec 6, 2023

Accepted: Feb 16, 2024



PREDICTION METHOD OF RATE OF PENETRATION BASED ON FUZZY SUPPORT VECTOR REGRESSION

LI YANG*, LISHEN WANG†, LILI BAI‡ AND WENFENG SUN§

Abstract. Predicting the rate of penetration (ROP) is important for optimizing drilling parameters, improving drilling efficiency, and optimizing economic benefits throughout the drilling process. The current prediction model of ROP based on machine learning algorithms does not consider the interference of outliers. Therefore, in this study, we propose a method to predict ROP based on fuzzy support vector regression (FSVR). First, appropriate input parameters were selected from the controllable parameters. Second, based on the local outlier factor, a fuzzy membership degree was assigned to each sample. Finally, the sample with the fuzzy membership value was input into the model for ROP prediction. The results demonstrated that the goodness of fit (R2) of the improved FSVR model is 0.9634, and the mean absolute error is 0.1974. Compared with standard SVR and other models, the improved FSVR model has a stronger anti-interference ability, smaller prediction error for normal samples, and higher accuracy.

Key words: drilling; rate of penetration; support vector regression; local outlier factor; fuzzy membership.

1. Introduction. The rate of penetration (ROP) is an important factor affecting drilling duration and drilling cost. Accurate prediction of ROP can provide support for drilling parameter optimization, thus enhancing drilling efficiency while minimizing drilling expenses. There are three types of ROP prediction models [1]: traditional mathematical equations, which consider a few small numbers of ROP influencing parameters [2], statistical models represented by multiple regression, which can have increased solving accuracy for the traditional ROP coefficient [3]; and models employing machine learning algorithms. Wu et al [4] established a linear prediction model using principal component analysis (PCA); however, in some cases, the element obtained by this method is not optimal. Su et al. [5] applied a gradient boosting decision tree algorithm to predict ROP, which is more accurate than the existing method of predicting ROP; however, the characteristics involved were limited. Kahraman et al. [6] proposed an ROP prediction model based on backpropagation (BP) neural networks; however, BP neural networks have low convergence speed and can easily fall into local extrema. Yang et al. [7] established a fuzzy neural network model by combining fuzzy control with a BP neural network, and the level of prediction accuracy was higher than that of ordinary BP neural networks. Zhao et al. [8] proposed an ROP prediction model based on extreme learning machines, achieving better generalization performance than the common BP neural network. Li et al. [9] combined the beetle antennae search (BAS) algorithm with a BP neural network to establish the BAS-BP model. Compared with standard BP and genetic-algorithm BP models, the error is lower and the prediction accuracy is higher; however, the generalization ability of the model should be improved. Xu et al. [10] established an ROP prediction model based on integrated learning, utilizing a variety of machine learning algorithms, and the results show that the prediction accuracy of the integrated model is higher than that of the single model; however, the model is more complex. Sabah et al. [11] used a multilayer perceptron particle swarm optimization (PSO-MLP) model for ROP prediction, achieving favourable performance. As a representative of traditional machine learning, support vector regression (SVR) has been demonstrated to have fitting performance compared to alternative machine learning methods (e.g., k-nearest neighbours, linear regression, polynomial regression, and decision trees) in experiments involving a drilling process that is characterized by high nonlinearity and complexity [12, 13]. The above two models are

*College of Electrical Information Engineering, Northeast Petroleum University, Daqing 163318, China

†College of Electrical Information Engineering, Northeast Petroleum University, Daqing 163318, China

‡Shanghai Technical Institute of Electronics & Information, STIEI Shanghai, 201411, China

§Shanghai Technical Institute of Electronics & Information, STIEI Shanghai, 201411, China (Corresponding author, 380780857@qq.com)

compared with the proposed model. Hamid et al. [14] used an imperialist competitive algorithm to optimize SVR. They established the model, compared three kernel functions, and found that the Gaussian radial basis kernel function had the best effect. Therefore, the Gaussian radial basis kernel function was also applied in this study.

Recently, entropy-based improved support vector machine models have been proposed to solve the classification problem [15, 16, 17]. Asadolahi et al. [18] and Xue et al. [19] improved the loss function by combining the fuzzy response with robust SVR to improve the accuracy of the fuzzy regression model. Chakravarty et al. [20] and Liu et al. [21] improved the fuzzy function and the robustness of the model against outliers and verified its effectiveness. Fuzzy control has been continuously applied to complex system processing [22]. Successful applications have been demonstrated in wind turbine control [23], concrete component analysis [24], and life prediction of slewing bearings [25]. Nevertheless, applications in drilling engineering remain scarce.

During drilling operations, signals are disturbed to produce outliers or slip drilling and other situations occasionally occur in drilling may lead to sudden increase in the values of certain parameters, which interferes with the training of the model and consequently affects its performance. A method of elimination has been adopted to process outlier data; however, it may delete misjudged data. When processing new data sets, there will be outliers of new data sets, which should be eliminated again. Therefore, we focused on reducing the influence of abnormal data on model prediction without deleting primary data. Current ROP prediction models based on machine learning algorithms do not consider interference from outlier samples; however, the fuzzy support vector regression (FSVR) model proposed in this study considers such interference. According to the outlier detection method based on the local outlier factor (LOF) algorithm, the LOF values of all data points are calculated, and then the membership values of all data points are calculated through the LOF values. A normal sample has a large membership value, and an outlier sample has a small membership value. The fuzzy training set with membership values is input into the training model. This can reduce the interference of outlier samples in the model, and the normal samples with a large contribution to the training model can play the greatest role. We demonstrated how this model has stronger anti-interference ability and higher prediction accuracy than other models.

2. Method.

2.1. Calculation of local outlier factor values. Outlier detection based on LOF is a density-based outlier detection method; it works on the principle that each data point is assigned a LOF that depends on neighbourhood density. For any data point in a particular dataset, if the number of surrounding points is high and its local neighbourhood is dense, the data point is considered as a normal data point. Contrarily, an outlier is a data point that is far from the nearest neighbour of a normal data point. Thus, LOF can be understood as the outlier degree of a sample relative to its neighbours.

Among the k nearest neighbours to P , the distance between the farthest neighbour and P is the k -distance of P . The k -distance neighbour of sample P is the neighbour set whose distance from P is not greater than the k -distance of P . It is defined as the k -distance neighbour set $N_k(P)$, provided $N_k(P)$ by equation (1), where P' is the neighbour of P and D is the sample set. $d(P, P')$ is the distance between P' and P , and $d_k(P)$ is the k -distance of P . Furthermore, $d_k(P, O)$ is the k -th reachable distance from sample O to sample P . It corresponds to the largest distances among the k -distance from O and the distance from P to O , and it is provided by the expression in equation (2), where $d_k(O)$ is the k -distance of O and $d(P, O)$ is the distance between P and O .

$$N_k(P) = \{P' | P' \in D \setminus \{P\} | d(P, P') \leq d_k(P)\} \quad (1)$$

$$d_k(P, O) = \max\{d_k(O), d(P, O)\} \quad (2)$$

The neighbours within a sample and the density relationship between the sample and its neighbours can be described by the concept of local reachable density. The local reachable density of a sample P is the number of elements in the k -nearest neighbour set of P divided by the sum of the relative reachable distances from all points within the data set to P , which is expressed as $\rho_k(P)$. The expression is as follows:

$$\rho_k(P) = \frac{|N_k(P)|}{\sum_{o \in N_k(P)} d_k(P, O)} \quad (3)$$

The LOF is obtained by calculating the ratio of the average local reachable density of k points around the sample point to the local reachable density of this point, which is expressed by equation (4).

$$lof_k(P) = \frac{\sum_{o \in N_k(P)} \frac{\rho_k(O)}{\rho_k(P)}}{|N_k(P)|} \tag{4}$$

2.2. Fuzzy support vector regression. To improve the anti-interference ability of SVR, FSVR is obtained by combining the fuzzy membership value with SVR. Several sample points are required to train the model; however, the contribution of different sample points to the final result is different. Therefore, we assign a fuzzy membership value, λ , to each sample based on the LOF to obtain the fuzzy training set $D' = \{(x_1, y_1, \lambda_1), (x_2, y_2, \lambda_2), \dots, (x_m, y_m, \lambda_m)\}$, $y_i \in \mathbb{R}, 0 \leq \lambda \leq 1$, where the specific setting method for λ_m is described in the subsection Calculating the value of fuzzy membership in Section 3 of this paper. The mathematical expression of the model is provided by equation (5). The fuzzy membership degree λ is the contribution of the sample point to the final regression results, that is, a larger λ is assigned to a normal sample or a sample with a larger contribution to the model training, and a smaller λ is assigned to an abnormal outlier data point or a sample with a smaller contribution to the final result. This can reduce the interference of some of the samples outlier points or meaningless samples on the model to ensure that the final prediction results are closer to the true value, resulting in increased robustness of the model.

$$\begin{aligned} & \min \left\{ \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \lambda_m (\xi_i + \xi_i^*) \right\} \\ & \text{s.t.} \begin{cases} y_i - f(x_i) \leq \varepsilon + \xi_i \\ f(x_i) - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0, i = 1, 2, \dots, n. \end{cases} \end{aligned} \tag{5}$$

The above formula is an optimization problem with constraints, where ε is an insensitive loss function; ξ_i and ξ_i^* are slack variables, which describe the loss caused by the sample away from the regression interval; and C is the penalty factor, which refers to the penalty for the sample whose regression function error is greater than ε . The smaller the value of C , the smaller the loss of the sample and, consequently, the smaller the loss of the objective function. Points far from the interval have larger errors; therefore, the penalty factor C will make the regression function more sensitive to such points to further fit these points. To calculate the corresponding loss, the absolute value should be greater than ε . In particular, considering $f(x)$ as the center, we set an interval band width of 2ε . When the training sample falls into the interval band, the prediction result is correct.

The Lagrange multiplier $a_i, a_i^*, \mu_i, \mu_i^*$ is introduced to obtain the Lagrange equation, and the partial derivative of the equation with respect to w, b, ξ_i, ξ_i^* is obtained, such that the partial derivative is 0. The results of partial derivatives are returned to the Lagrange equation, and the equation of the dual problem is obtained as follows:

$$\begin{aligned} & \max \sum_{i=1}^n y_i (a_i^* - a_i) - \varepsilon (a_i^* + a_i) - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (a_i^* - a_i) (a_j^* - a_j) \kappa(x_i, x_j) \\ & \text{s.t.} \begin{cases} \sum_{i=1}^n (a_i^* - a_i) = 0 \\ 0 \leq a_i, a_i^* \leq C \lambda_m \end{cases} \end{aligned} \tag{6}$$

When the above process satisfies the KarushKuhnTucker conditions, the value of b is obtained as follows:

$$b = y_i + \varepsilon - \sum_{i=1}^n (a_j^* - a_j) \kappa(x_i, x_j) \tag{7}$$

After feature mapping, w is expressed as follows:

$$w = \sum_{i=1}^n \phi(x_i) (a_i^* - a_i) \tag{8}$$

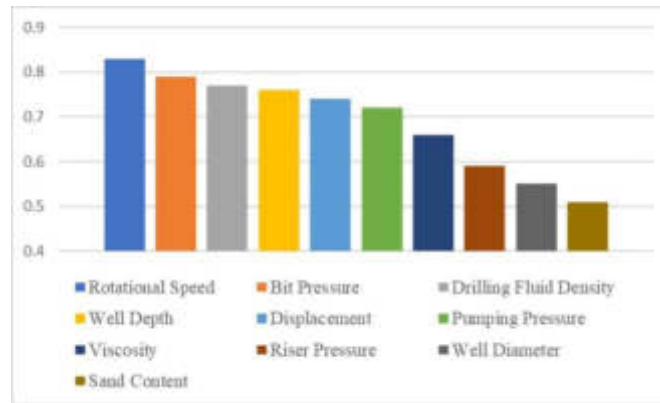


Fig. 3.1: Ranking of grey correlation degree

Table 3.1: Partial training sample - Part 1

rotational speed $/(r \cdot \text{min}^{-1})$	bit pressure /KN	density $/(g \cdot \text{cm}^{-3})$	well depth/m
120	20	1.10	214.10
160	30	1.14	365.03
176	50	1.19	1733.02
176	60	1.20	1876.11

By substituting w into $f(x) = w^T \phi(x) + b$, the final expression in equation (9) is obtained, where κ is the kernel function. Through this, the sample points can be mapped from low-dimensional space to high-dimensional space, which reduces the calculation difficulty. This model uses the Gaussian radial basis kernel function, which can better deal with high-dimensional samples and involves fewer parameters, consequently gaining good stability.

$$f(x) = \sum_{i=1}^n (a_i^* - a_i) \kappa(x, x_i) + b \quad (9)$$

3. Experiments and results.

3.1. Analysis of influencing factors on ROP. The parameters related to mechanical drilling speed can be divided into controllable and uncontrollable. Uncontrollable parameters are objective parameters related to formation conditions, such as pore pressure and fracture pressure. Controllable parameters are parameters that can be adjusted by technology and equipment and can be further divided into drilling fluid parameters, hydraulic parameters, and mechanical parameters. Drilling fluid parameters include density, rheological parameters, etc. Hydraulic parameters include bit pressure drop, etc. Mechanical parameters include drilling pressure, rotational speed, etc.

The data used in the model are from a block in the Shunbei Oilfield. There are a total of 11173 samples in the dataset. The mechanical drilling speed is assumed as the parent sequence, and its influencing parameters are assumed as the subsequence. The correlation degree ranking of the influencing parameters of ROP is obtained by grey correlation analysis and calculation, as shown in Figure 3.1.

Figure 3.1 shows that the correlation degree between each parameter and mechanical drilling rate is significantly different. The first seven parameters exhibiting a high correlation degree, namely, rotational speed, bit pressure, drilling fluid density, well depth, displacement, pumping pressure, and viscosity, are assumed to be the input of the model. Some training samples are shown in Table 3.1.

3.2. Preprocessing of data.

Table 3.2: Partial training sample - Part 2

displacement/($L \cdot s^{-1}$)	pumping pressure/ Mpa	viscosity/ m	ROP/($m \cdot h^{-1}$)
35	6	52	2.16
38	11	57	2.01
35	13	60	1.30
35	13	56	0.72

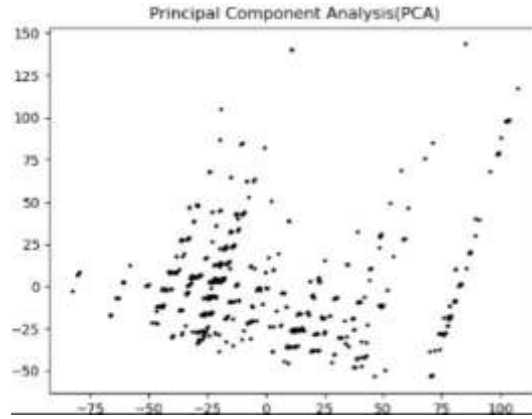


Fig. 3.2: Visualization results of 2D data

3.2.1. Detecting outliers. Owing to the introduction of the concept of nearest neighbours, the method based on LOF can better identify outliers located at the edge of the sample class and has better recognition effect. To display the results of LOF outlier detection conveniently, PCA dimension reduction was used to reduce the multidimensional data to 2D for visualization. The results are shown in Figure 3.2.

The model calculates the LOF for each data point and determines whether the point is an outlier based on the size of the LOF value. According to its definition, if the LOF value is close to 1, it indicates that the sample density is similar to its neighbourhood density, and the sample is classified as a normal data point. It is believed that a LOF value between 1 and 1.5 corresponds to normal data. If the LOF value is significantly greater than 1, it indicates that the sample density is less than its neighbourhood density, and the sample may be classified as an outlier. The 2D data obtained after dimensionality reduction are used for outlier detection, and the results are shown in Figure 3.3, where the red circles mark each data point. The size of the red circle represents the size of the LOF value, that is, the radius of the red circle determines the outlier degree of the data point. The larger the radius of the red circle, the lower the sample density; hence, the sample may be outliers. A smaller radius of the red circle indicates that the point is within the same cluster as the surrounding neighbourhood points and the degree of outliers is low, which indicates a normal data point. The outliers were labelled -1 and the normal points were labelled 1.

3.2.2. Normalizing the data. To eliminate the influence of different parameter ranges on the model and improve the generalization ability of the model data processing, the data should be normalized. The data normalization interval is $[-1,1]$, and the expression is as follows:

$$x_{norm} = \frac{(y_{max} - y_{min})(x - x_{min})}{x_{max} - x_{min}} + y_{min} \quad (10)$$

where x_{norm} and x represent the values after and before data normalization, respectively; x_{max} and x_{min} represent the maximum and minimum values before data normalization; and y_{max} and y_{min} represent the maximum and minimum values after normalization.

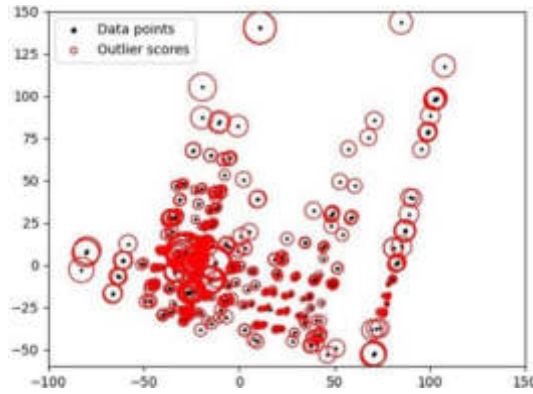


Fig. 3.3: LOF value of sample point

Table 3.3: Partial prediction results of SVR.

prediction of ROP for SVR	actual value of ROP	error	Is it a outlier (SVR)	LOF	membership value	Is it a outlier (LOF)
3.09	3.20	0.11	1	1.0251	0.8996	1
5.26	8.37	3.11	-1	1.7096	0.6203	-1
7.04	10.12	3.08	-1	1.8139	0.6059	-1
6.21	9.41	3.20	-1	1.7325	0.6147	-1

3.3. Training the fuzzy support vector regression model.

3.3.1. Training the standard SVR model. Normal training classifies data sets into training and test sets, which renders data sets not fully applicable to training. This can be avoided by cross-validation, wherein it is possible to use all data for training and testing. This model uses ten-fold cross-validation, wherein the data set is divided into ten sets, of which nine are training sets and one is a test set, which is used for an experiment. Subsequently, a new test set is chosen from the training set, and the original test set is used for training. Thus, ten experiments were conducted. Finally, the model with the smallest average absolute error in the ten experiments was assumed as the final model, and the average value of the ten average absolute errors was used to evaluate the model.

To verify the accuracy of the labelled outliers in the subsection Detecting outliers, the standard SVR model was used for ROP prediction. The data set whose absolute difference between the predicted ROP value and real value was greater than 3, which was temporarily recorded as an abnormal outlier and marked as -1, while other normal data were marked as 1. Some prediction results are listed in Table 3.3.

There are 81 groups of abnormal outlier data marked by the standard SVR model. Compared with the 72 groups of outlier data detected using the LOF, there are 61 groups of data consistent between both models, thereby indicating that the LOF outlier detection recognition rate reaches 84.72%.

3.3.2. Calculating the value of fuzzy membership. The membership function value λ_m in this study is determined by the value of the LOF, expressed as

$$\lambda_m = \begin{cases} (1 - \theta)^\mu + \sigma & \overline{lof} < lof_k(p) \leq lof_{max} \\ 1 - \theta & lof_{min} \leq lof_k(p) \leq \overline{lof} \end{cases}, \theta = \frac{lof_k(p) - lof_{min}}{lof_{max} - lof_{min}} \quad (11)$$

Here, $\mu \geq 2$ (in this study, n is taken to be 10); σ is a sufficiently small positive real number less than 1; and lof_{min}, lof_{max} , and \overline{lof} are the minimum, maximum, and average values in the sample LOF, respectively.

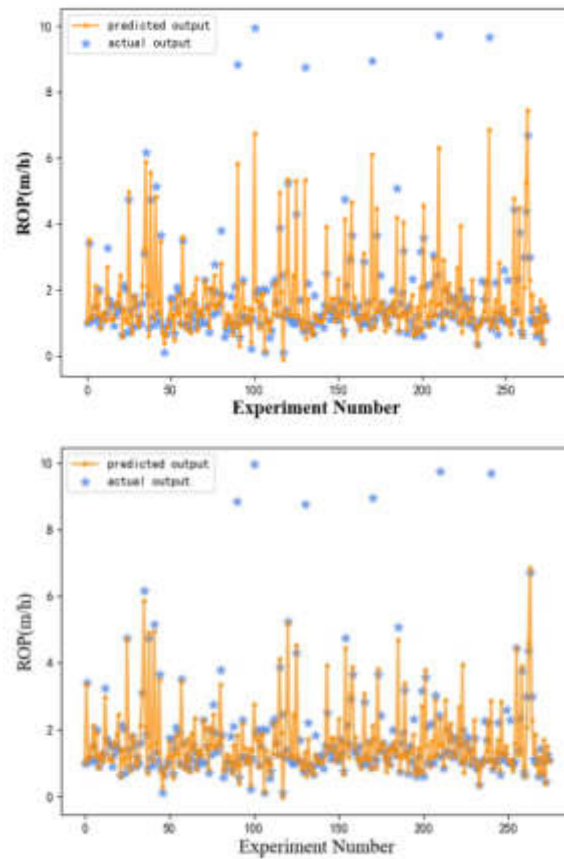


Fig. 3.4: (a) Results before model improvement (b) Results after model improvement

3.3.3. Training the fuzzy support vector regression model. The calculated fuzzy membership value is attached to each data point, and the fuzzy data set $D' = \{(x_1, y_1, \lambda_1), (x_2, y_2, \lambda_2), \dots, (x_m, y_m, \lambda_m)\}$ is sent to the FSVR model for training. The purpose is to control the samples with a large LOF and reduce their membership value to ensure that they have a smaller influence on the model. For the samples with a small LOF, the membership value is large to ensure that they have a significant contribution to the model. This training still uses ten-fold cross-validation. The results before and after improvement are shown in Figures 4 and 5, respectively. It can be observed that for the outliers above the general trend of the data, the standard SVR model (Figure 3.4) shifts the outliers significantly; thus, it is prone to overfitting and reduces generalization ability. However, the FSVR model with the fuzzy membership value added (Figure 5) is less affected by outliers and does not shift significantly toward them, thus demonstrating stronger anti-interference ability.

The four subplots within Figure 3.5 are the prediction results of ROP in different formations of four wells within the Shunbei Oilfield using the improved FSVR model. Figures 3.5(a)(d) utilize well Nos. 14, respectively. The abscissa shows different strata, and the ordinate is ROP. It can be observed from the figures that the error between the predicted and real values is controlled within 1, and the prediction effect is good.

3.4. Contrast experiment.

To evaluate the prediction accuracy of the model for normal data, the abnormal outliers of the dataset were removed, and three models mentioned in the Introduction section ordinary SVR, BP neural network, and PSO-MLP were used, in addition to the proposed FSVR model, to predict ROP. The errors of these four models were compared, and the results are listed in Table 3.4. It can be observed from Table 3.4 that the mean absolute

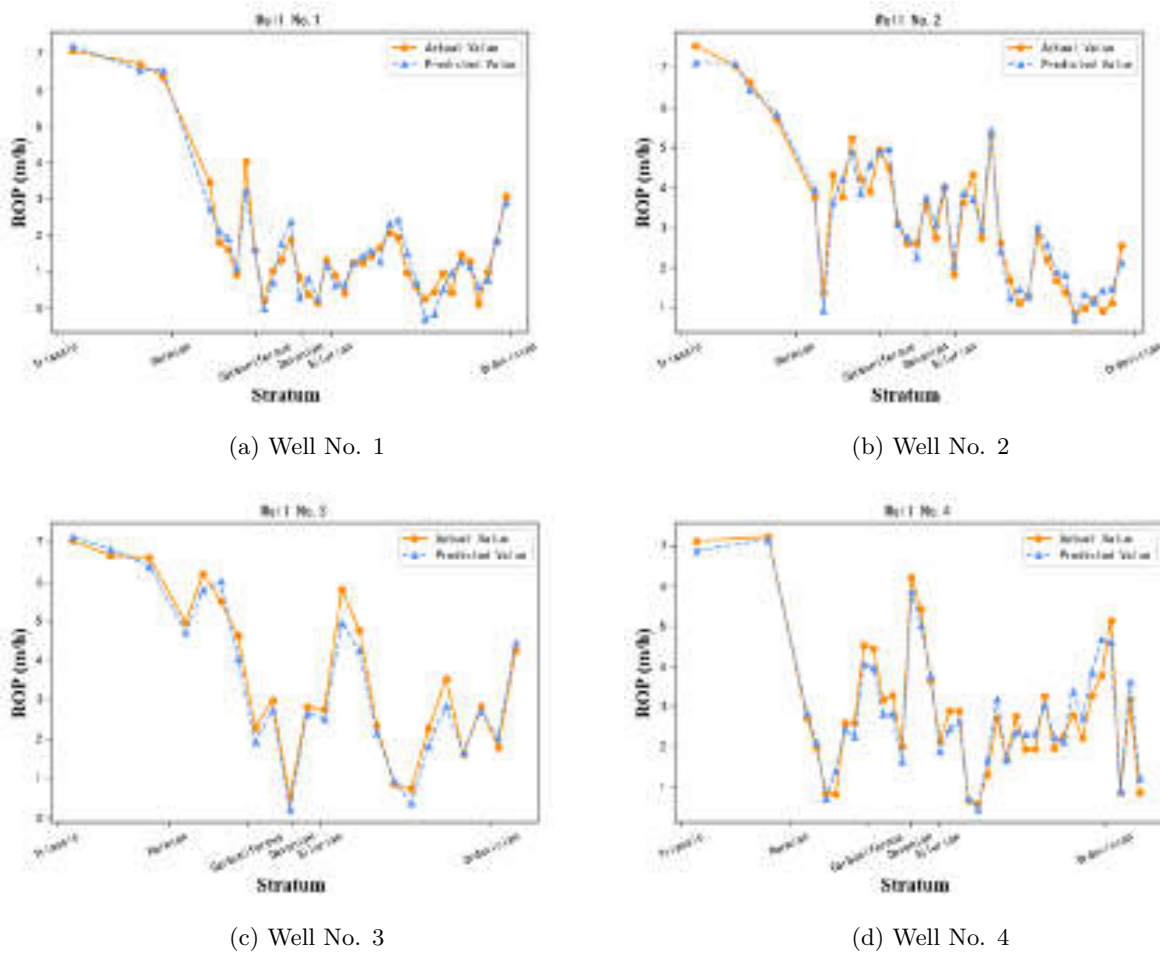


Fig. 3.5: Prediction results of FSVR

Table 3.4: Comparison of errors.

model	R^2	MAE	MSE
FSVR	0.9634	0.1974	0.0709
SVR	0.9354	0.2430	0.1252
PSO-MLP	0.9168	0.2446	0.1613
BPNN	0.8210	0.3784	0.4248

error (MAE) and mean squared error (MSE) of the FSVR model are smaller than those of the other models. The four graphs in Figure 3.6 show the goodness of fit (R^2) of the four models; R^2 of the FSVR model is closer to 1, thereby indicating that the predicted value of the FSVR model is closer to the real value; therefore, the regression fitting effect is enhanced.

According to the comparative experiments, the improved FSVR model is not sensitive to abnormal outlier data, which can reduce the interference of outliers within the model. Additionally, the model has higher prediction accuracy for normal data points.

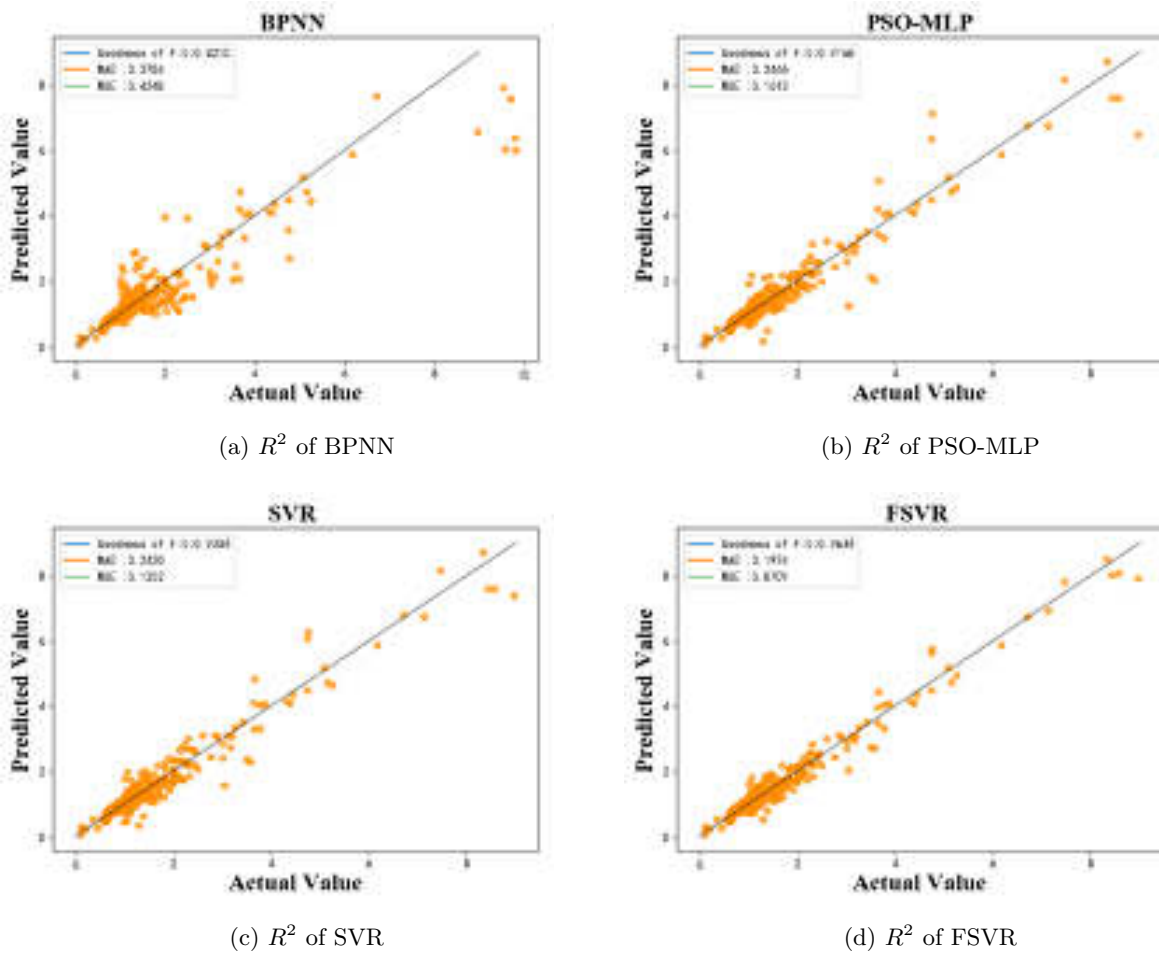


Fig. 3.6: Graphs showing goodness of fit

4. Conclusions. In this study, the FSVR model was established to address the problem of interference of outlier samples in a model, which are not considered by ROP prediction models based on machine learning algorithms. Parameters such as rotational speed and bit pressure of a block in the Shunbei Oilfield were selected as the input through grey correlation analysis. According to the LOF, a fuzzy membership value was assigned to each group of samples and was used to weight the penalty factor. The model was subsequently trained by the fuzzy training set. Attaching a smaller membership value to the outliers reduces the interference induced in the model. The normal data points were assigned a larger membership value to ensure that they have a greater influence. The results demonstrated the following.

1. The improved FSVR model has better anti-interference ability than other models, such as standard SVR, and is not prone to overfitting.
2. R^2 of the improved FSVR model is 0.9634, and the MAE is 0.1974. Therefore, the improved FSVR model has a smaller prediction error, higher accuracy, and better regression fitting effect for ROP of normal sample points.
3. The model has a good ability to predict ROP, which can provide support for the optimization of drilling parameters in the next step, thereby improving drilling efficiency and saving drilling costs.

Acknowledgments. The authors are grateful to High-level and scarce talent research Initiation Foundation of STIEI under Grant No. GCC2023007 and No. GCC2023034, and the National Natural Science Foundation of China under Grant No. 51974090.

REFERENCES

- [1] L.F.F.M Barbosa, A. Nascimento, M.H. Mathias et al. (2019) Machine learning methods applied to drilling rate of penetration prediction and optimization-A review, *Journal of Petroleum Science and Engineering*, 183, 106332.
- [2] C. Soares and K. Gray (2019) Real-time predictive capabilities of analytical and machine learning rate of penetration (ROP) models, *Journal of Petroleum Science and Engineering*, 172, 934-959.
- [3] C.S. Li (2013) Research on drilling speed prediction method based on multiple regression analysis, *Science Technology and Engineering*, 13(7), 1740-1744.
- [4] C.G. Wu, M. Zhao and Z.W. Guo (2015) Study on drilling rate prediction based on principal component analysis, *Mining Research and Development*, 35(10), 84-86.
- [5] X.H. Su, J.M. Sun, X. Gao et al. (2019) Prediction method of drilling rate of penetration based on GBDT algorithm, *Computer Applications and Software*, 36(12), 87-92.
- [6] S. Kahraman (2016) Estimating the penetration rate in diamond drilling in laboratory works using the regression and artificial neural network analysis, *Neural Processing Letters*, 43(2), 523535.
- [7] L. Yang, T.Y. Liu, W.J. Ren et al. (2021) Fuzzy Neural Network for Studying Coupling between Drilling Parameters, *ACS Omega*, 6(38), 24351-24361.
- [8] Y. Zhao, T. Sun, J. Yang et al. (2019) Extreme learning machine-based offshore drilling ROP monitoring and real-time optimization, *China Offshore Oil and Gas*, 31(6), 138-142.
- [9] Q. Li, F.T. Qu, J.B. He et al. (2021) Prediction model of mechanical ROP during drilling based on BAS-BP, *Journal of Xi'an Shiyou University (Natural Science Edition)*, 36(6), 89-95.
- [10] M.Z. Xu, M.H. Wei, S. Deng et al. (2021) Application of multi-model ensemble learning in prediction of mechanical drilling rate, *Computer Science*, 48, 619-622.
- [11] M. Sabah, M. Talebkeikhah, D.A. Wood et al. (2019) A machine learning approach to predict drilling rate using petrophysical and mud logging data, *Earth Science Informatics*, 12(3), 319339.
- [12] Y. Zhou, X. Chen, H. Zhao et al. (2021) A novel rate of penetration prediction model with identified condition for the complex geological drilling process, *Journal of Process Control*, 100, 3040.
- [13] O.S. Ahmed, A.A. Adeniran, and A. Samsuri (2019) Computational intelligence based prediction of drilling rate of penetration: A comparative study, *Journal of Petroleum Science and Engineering*, 172, 1-12.
- [14] R.A. Hamid, J.S.H. Mohammad, and A. Masoud (2017) Drilling rate of penetration prediction through committee support vector regression based on imperialist competitive algorithm, *Carbonates and Evaporites*, 32(2), 205-213.
- [15] S. Moslemnejad and J. Hamidzadeh (2021) Weighted support vector machine using fuzzy rough set theory, *Soft Computing*, 25(13), 8461-8481.
- [16] D. Gupta and B. Richhariya (2018) Entropy based fuzzy least squares twin support vector machine for class imbalance learning, *Applied Intelligence*, 48(11), 4212-4231.
- [17] S. Chen, J. Cao, and F. Chen (2020) Entropy-based fuzzy least squares twin support vector machine for pattern classification, *Neural Processing Letters*, Vol. 51(11), 41-66.
- [18] M. Asadolahi, M.G. Akbari, and G. Hesamian (2021) A robust support vector regression with exact predictors and fuzzy responses, *International Journal of Approximate Reasoning*, 132, 206-225.
- [19] Z. Xue, R. Zhang, and C. Qin (2020) An adaptive twin support vector regression machine based on rough and fuzzy set theories, *Neural Computing and Applications*, Vol. 32(9), 4709-4732.
- [20] S. Chakravarty, H. Demirhan, and F. Baser (2020) Fuzzy regression functions with a noise cluster and the impact of outliers on mainstream machine learning methods in the regression setting, *Applied Soft Computing*, 96, 106535.
- [21] J. Liu (2021) Fuzzy support vector machine for imbalanced data with borderline noise, *Fuzzy Sets and Systems*, 413, 64-73.
- [22] A.T. Nguyen, T. Taniguchi, and L. Eciolaza (2019) Fuzzy control systems: Past, present and future, *IEEE Computational Intelligence Magazine*, 14(1), 56-68.
- [23] T. Jeon and I. Paek (2021) Design and verification of the LQR controller based on fuzzy logic for large wind turbine, *Energies*, 14(1), 230.
- [24] Z. Fan, R. Chiong, and Z. Hu (2020) A fuzzy weighted relative error support vector machine for reverse prediction of concrete components, *Computers & Structures*, 230, 106171.
- [25] W. Bao, H. Wang, J. Chen et al. (2020) Life prediction of slewing bearing based on isometric mapping and fuzzy support vector regression, *Transactions of the Institute of Measurement and Control*, 42(1), 94-103.

Edited by: Bradha Madhavan

Special issue on: High-performance Computing Algorithms for Material Sciences

Received: Dec 25, 2023

Accepted: Mar 21, 2024



RESEARCH ON DISTRIBUTED SCHEDULING ALGORITHM FOR VIRTUAL CITY POWER PLANTS BASED ON BLOCKCHAIN TECHNOLOGY

QING ZHU^{*}, YUFENG ZHANG[†] AND JIZE SUN[‡]

Abstract. In order to solve the problem that allocating the whole virtual power station is not centralized and promoting the free access of power distribution, a research on distributed scheduling of urban power network based on blockchain technology is proposed. and a scheme is proposed. Based on the consistency of decentralized virtual power plant and blockchain in decentralized point-to-point interaction and decentralized cooperation, this paper proposes to use blockchain consensus mechanism to realize distributed scheduling of virtual power plant. According to the principle of continuous micro- increment cost, the optimal economic dispatch of virtual power system is realized by taking the micro- increment characteristic as the variable. and the optimal economic dispatch strategy is realized. As a node of the blockchain, each distributed energy in the virtual power plant has a complete backup of the key data of the whole network. When the load changes, each node uses the PBFT consensus algorithm to independently calculate the new power of each unit. The new power data is stored on the chain, and the global consistency of micro incremental features is maintained to realize the reasonable load distribution among units. The experimental results show that, when the initial $t = 1$, the λ the values of each unit are different. The system does not meet the principle of constant consumption micro-increase rate, and the system does not operate in the optimal state. After a PBFT consensus, λ variable reaches the same value at $t=8$, and the system reaches the optimal operating state. Conclusion: The simulation experiment verifies the effectiveness of the algorithm, realizes distributed scheduling by using blockchain, and provides a feasible reference scheme for the operation mode of decentralized virtual power plant.

Key words: Consensus mechanism, Distributed scheduling, Completely distributed, Virtual power plant, Blockchain

1. Introduction and examples. With the rapid development of renewable energy such as wind and light, the pattern of distribution of electricity entering the distribution network is increasingly in the direction of access go high and high speed. However, a large number of new energy sources are interconnected, and their randomness and uncertainty will have an important impact on the operation, dispatching and trading of the power grid. Virtual power plant technology is an important part of the integration of multiple power distribution sites to participate in the power market and promote new energy use [1]. Virtual power plant technology can effectively balance the supply and demand relationship of electricity, improve the reliability and stability of electricity, and promote the use and popularization of new energy by integrating and managing multiple small distributed energy resources. In addition, virtual power plant technology can also provide more flexible and efficient trading methods for power producers and consumers through the participation of the electricity market, while also reducing the cost and risk of electricity trading.

As a new form of energy aggregation, virtual power plant provides a feasible method for a large number of distributed energy sources to be reliably connected to the grid. Through advanced communication technology, virtual power plant aggregate control of distributed energy, and make distributed energy coordinated and optimized operation, so as to achieve mutual adjustment of output, reliable grid connection. Compared with traditional power plants, virtual power plants have more diversified resources, are more environmentally friendly, and are more competitive in the power market, which promotes the transformation of the power industry and the development of the whole power system [2].

^{*}Nari Technology Co., Ltd., Nanjing, 211106, China; NARI Nanjing Control System Co., Ltd., Nanjing, 211106, China (Corresponding author, QingZhu15@126.com)

[†]Nari Technology Co., Ltd., Nanjing, 211106, China; NARI Nanjing Control System Co., Ltd., Nanjing, 211106, China (YufengZhang5@163.com)

[‡]Nari Technology Co., Ltd., Nanjing, 211106, China; NARI Nanjing Control System Co., Ltd., Nanjing, 211106, China (JizeSun@126.com)

Block chain is a new type of data structure formed by a large number of blocks linked together in an orderly manner, and block refers to the collection of relevant data information, is the basic unit of block chain [3].

2. Literature review. It adopts advanced information and communication technology to realize DERs (distributed energy storage system, controllable load, electric vehicle, etc.) resources) for integrated and coordinated optimization. As a special power plant to participate in the power market and grid operation of power coordination management system, is considered to be the ultimate configuration of energy Internet. The Energy Internet is a new energy system mainly composed of renewable energy and supported by information and communication technology. It has the characteristics of decentralization, intelligence, interconnection, and efficiency, and can achieve large-scale application and sustainable development of clean energy.

As an important component of the energy internet, virtual power plants can integrate dispersed distributed energy resources to form a virtual power plant, participate in transactions in the electricity market, and achieve large-scale application of clean energy. At the same time, they can also bring more intelligent and efficient management methods to the operation and management of the power grid, providing strong support for the construction and development of the energy internet.

Virtual power plant integrates various DERs together to realize the stability and reliability of its overall output and provide efficient electric energy for the grid, so as to ensure the stability and safety of grid-connection. According to operation control mode, virtual power plants can be divided into three types: centralized central-decentralized and fully decentralized. The centralized architecture not only has high requirements for communication system, but also is difficult to cope with the management of numerous fit and forget distributed energy sources. Therefore, in order to better realize the management of DERs, the scheduling strategy of virtual power plant has been transformed from traditional centralized to distributed. Distributed virtual power plant has gained more attention and has better scalability and openness. In distributed virtual power plants, the multi-agent technology is usually used to realize the distributed communication of DERs. Through the information interaction between the agent and the neighbor node, it realizes self-regulation and eventually tends to the consistent variable. However, because multi-agent technology cannot obtain global information efficiently and accurately, it needs to approach consistent results through continuous iteration, which has problems such as inaccurate calculation and low efficiency [4,5].

To solve this problem, some optimization methods can be adopted, such as hierarchical control, local information sharing, predictive control, etc. These methods can help multi-agent system achieve consistent results more quickly and reduce computing and communication overhead.

Relevant scholars have studied the application of blockchain technology in the decentralized trading mechanism of distribution network, automatic demand response system, energy Internet, multi-module collaborative autonomous mode, distribution network, power trading and other fields. Blockchain and distributed virtual power plant are consistent in decentralized decentralized collaborative regional autonomy and other aspects. For example, in the distributed virtual power plant, there is no central control mechanism, and each DERs obtains global information through direct data exchange and then completes the regulation and operation of its own node independently [6,7].

To sum up as a whole, for distributing all virtual power plants, according to their status with blockchain in point-to-point cooperation and fair cooperation, this paper proposes to use blockchain technology to realize the time division of virtual power plants. Based on the constant consumption micro-increment rate criterion, the optimal economic dispatch of the virtual power plant is realized by using the micro-increment characteristic as a consistent variable. Combined with the blockchain recommendation algorithm, each power distribution node independently calculates the new power of each unit, and stores the new power information of the chain, while maintaining the global relationship of small features, to realize the reasonable distribution of load among units [8].

3. Method.

3.1. VPP distributed scheduling model. The VPP (Virtual Power Plant) distributed scheduling model is an energy management model based on virtual power plants. It integrates multiple distributed energy devices (such as solar, wind, energy storage, etc.) into a virtual power plant, and achieves centralized management and scheduling of these devices through communication methods such as the Internet, thereby achieving

participation in the electricity market and energy trading.

This paper takes the optimal distribution of active power load as an example. The so-called optimal distribution is to make the power generation equipment consume the least amount of energy per unit time in the process of generating electric energy on the premise of satisfying the continuous power supply of a certain amount of load. It usually adopts the principle of constant consumption rate increase to distribute load among units. In order to simplify the description, the optimal active power allocation scheme without network loss is adopted [9,10].

The implementation of optimal allocation can effectively improve the efficiency and reliability of the power system, reduce energy consumption and operating costs, and also reduce environmental pollution and carbon emissions, achieving sustainable utilization of clean energy. Therefore, optimal allocation is one of the important means for the operation and management of the power system, which is of great significance for achieving energy transformation and building a sustainable energy system.

Assuming that the cost function of the generator set in DERs is quadratic, the minimum generation cost of VPP can be achieved by expressing the cost function:

$$\min F = \min \sum_{i=1}^n F_i(P_i) \quad (3.1)$$

n indicates the number of group DERs in VPP, P_i represents the output power of unit i . The total cost of VPP is denoted as F , $F_i(P_i) = a_i P_i^2 + b_i P_i + c_i$, a_i , b_i and c_i represents the coefficient of the cost function.

All the DERs units inside the VPP meet the active power balance of the whole system when running, without considering the network loss, namely

$$\sum_{i=0}^n P_i = P_{LD} \quad (3.2)$$

P_{LD} represents the total load demand of all users.

According to the principle of constant consumption micro increase rate, the load is distributed among units.

$$\lambda = \frac{dF_i}{dP_i} = 2a_i P_i + b_i \quad (3.3)$$

As a result, λ can be used as a consistent variable between nodes in the blockchain, adjusting as the load changes, but remaining consistent across the network [11].

3.2. Consensus mechanism in alliance chain. The private chain is controlled by a single organization and is open only to its own organization; Consortiums are between public and private chains, open to a specific industry organization, and require that each new node be verified and audited. The alliance chain can adapt to the situation containing a small number of faulty or evil nodes and has certain fault tolerance characteristics [12].

Generally, different types of blockchain adopt different consensus mechanisms, and Byzantine Fault Tolerance (BFT) is one of the core issues to be solved in blockchain consensus algorithms. Bitcoin's POW and Ethereum's POS are public chain algorithms, which solve the BFT in the case of numerous consensus nodes. Compared with POW POS of the public chain, PBFT adopts the mechanism of each node voting to reach consensus, which can solve the bifurcation problem and improve efficiency. The operating environment of the PBFT requires a relatively closed cluster, and each consensus requires multiple p2-node communication. The traffic volume is $O(n^2)$, and n is the number of nodes in the cluster [13]. PBFT is suitable for alliance chain dominated by industry and government, and is an ideal choice for a system with limited number of nodes and no need for virtual currency incentive mechanism. Based on the operation characteristics of virtual power plant, this paper chooses alliance chain and adopts PBFT consensus algorithm [14].

The premise of PBFT algorithm is to ensure that message communication between nodes is immutable through cryptography technology. PBFT has certain fault tolerance characteristics, assuming that the total number of nodes in the system is $|n| = 3f + 1$, f is Number of tolerated PBFT faulty or malicious nodes. So in order for the whole system to work properly, you need to have $2f+1$ normal nodes. The classic representative project in the consortium chain is the Fabric project under the IBM Hyperledger organization, with Fabric0.6 using the PBFT algorithm [15].

Table 3.1: Blockchain and virtual power plant characteristics comparison

Feature	blockchain technology	Virtual Power Plant
Decentration	All nodes have equal rights and obligations	All power generation and power consumption subjects are equal, dispersed and coordinated
Cooperative autonomy	Using the consensus mechanism, all nodes jointly maintain	Each DERs realizes collaborative scheduling and supply-demand balance through data interaction
Credit system	No third party trust mechanism is required	Each DERs uses communication network to realize point-to-point direct interaction
Smart contract	The ability to automatically execute contracts	The power of each unit should be set reasonably according to the control command

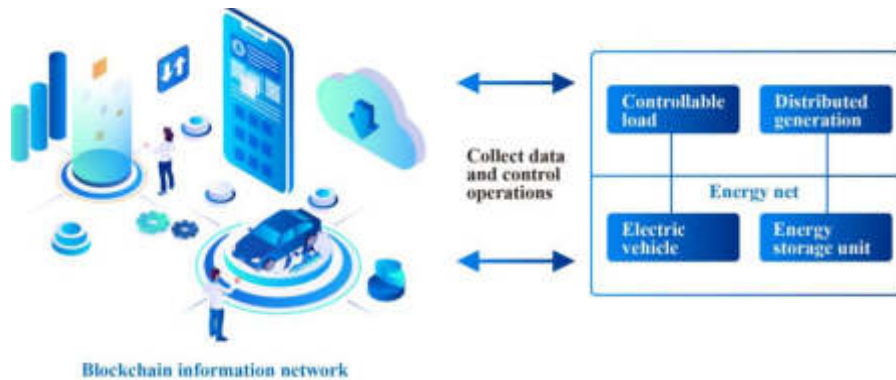


Fig. 3.1: Blockchain builds the underlying information architecture of decentralized virtual power plants

3.3. Blockchain distributed scheduling model. Blockchain distributed scheduling model is an energy management model based on blockchain technology, which realizes the security and reliability of the power market participation and energy transaction through the establishment of a distributed account. In the blockchain distributed scheduling model, each distributed energy device is regarded as a node in the network, and each node has an account and some data. These data include the equipment’s production capacity, storage capacity, energy consumption, and so on. Data from all nodes is encrypted and stored on the blockchain, with each node having access to and modify its own data. In a completely decentralized virtual power plant, there is no centralized control mechanism, and the decision-making process is completed through direct interaction and multiple iterations of each power generation and power consumption unit. Based on the consistency of the two sides’ ideas and needs, blockchain is expected to become one of the underlying architectures for building decentralized virtual power plants, and realize the regulation and operation of virtual power plants. The comparison of characteristics between virtual power plants and blockchain technology is shown in Table 3.1.

Based on the coincidence between blockchain and virtual power plant, blockchain can be used to build the underlying information architecture of decentralized virtual power plant. Each DERs forms a blockchain network, which can realize point-to-point communication. The consensus mechanism is used to calculate the consistency variables. The specific structure of decision control through intelligent contract execution is shown in Figure 3.1.

As shown in Figure 3.1, it consists of a physical energy transmission network and a blockchain information network, where the energy network is responsible for power transmission and connects each DERs. Each DERs

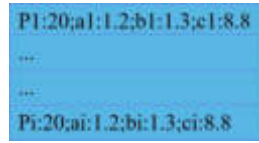


Fig. 3.2: Data stored on the chain

unit corresponds to a node in the blockchain information network. Data and control signals can be transmitted between blockchain nodes and DER, and a point-to-point network is formed between nodes. The control strategy of the entire virtual power plant is calculated by the blockchain network and delivered to the physical network for execution.

3.4. Distributed scheduling policy execution process.

3.4.1. On-chain data. When the optimal operation configuration is reached, the micro increment characteristic λ value of all units is the same. Therefore, λ is selected as the state information of the blockchain platform, and the calculation is completed through the consensus mechanism, and the consistency is maintained.

According to the characteristics of blockchain data storage, each node saves a complete backup of global status data and keeps synchronization with other node data. According to the scheduling policy, each generator set needs to save the power information of all the units P_i (unit: kW) in the whole network and the cost function coefficient of the corresponding unit (a_i, b_i, c_i). Usually, the cost function remains inconvenient. After each power adjustment, the new power on-link storage makes use of blockchain technology so that each unit can accurately obtain the global state information. The data storage structure of each node is roughly shown in Figure 3.2 [16].

3.4.2. Distributed scheduling execution process. According to the PBFT consensus execution process, combined with the isobaric micro-increment criterion, the distributed scheduling calculation process is shown in Figure 3.3. When the total load demand changes, the power adjustment demand broadcast to all units, after PBFT consensus, each unit separately calculate the new power and λ , and adjust the power of the unit and broadcast λ at the same time all units of the new power on the chain storage for the next adjustment application.

The specific process is as follows

1. When the total load changes, the power calculation request operation of primary node DER1 is activated, and the primary node selection is completed according to the agreed algorithm rules of PBFT.
2. After receiving the request, the primary node broadcasts the request to each DERs node in the network according to the three-phase protocol rules.

Different requests have different ordinals. A pre-prepare message is constructed using the ordinals and request operations and broadcast to each DERs node. In this stage, the PLD and the power calculation request of the total change load are sent to each node.

Then there is the interaction phase, where each DERs node receives the pre-prepare message and each node broadcasts the prepare message to other DERs nodes. If a node receives messages from $2f$ different nodes, it means that the prepare phase of the node has been completed and each node has obtained the total load PLD after the change. Conditions are available to calculate the new power.

Finally, there is the sequence number confirmation stage. After each node verifies the request and order in the view, if the node receives $2f+1$ commit message, it means that most nodes have entered the commit stage and a consensus has been reached in this stage. The following takes the calculation process of the I -th node in the blockchain as an example to illustrate the calculation process of power allocation according to the constant consumption micro-increment criterion.

According to the distributed scheduling criterion, all nodes are required to have the same λ , that is, the formula 3.4 is satisfied

$$\lambda = 2a_1P_1 + b_1 = \dots = 2a_iP_i + b_i \quad (3.4)$$

The power of all nodes is expressed in P_i

$$\begin{cases} P_1 = ((2a_1P_i + b_i) - b_1)/2a_1 \\ P_2 = ((2a_2P_i + b_i) - b_2)/2a_2 \\ \dots \\ P_n = ((2a_nP_i + b_i) - b_n)/2a_n \end{cases} \quad (3.5)$$

Substitute formula 3.5 into formula 3.2 to get formula 3.6

$$P_{LD} = ((2a_1P_i + b_i) - b_1)/2a_1 + ((2a_2P_i + b_i) - b_2)/2a_2 + \dots + ((2a_nP_i + b_i) - b_n)/2a_n \quad (3.6)$$

As P_{LD} is known, P_i^* of DERs's new power can be calculated, then, P_i^* is substituted into formula 3.5 to obtain the new power value of all nodes and update the new value to the blockchain;The value λ is calculated at the same time and sent to the task initiator node; The drive DERs unit works according to the new power value [17].

3. The task initiator receives responses from different nodes. If there are $2f+1$ responses with the same λ value, the response is the result calculated for this request and is the consistent variable of the whole network. The distributed power distribution adjustment is completed to achieve optimal economic scheduling.

Specifically, the task initiator will perform the following actions:

Collect responses: Wait for responses from different nodes and collect them into a list.

Verify Response: For each response, verify its completeness and correctness. Ensure that the response is sent by the correct node and that the calculation results in the response are correct.

Select Response: For each request, select with the same λ $2f+1$ response of the value. If there is not enough response, wait for more responses until sufficient responses are received.

Calculation result: Use the selected response to calculate the final result. This will include the optimal economic dispatch values calculated for each node, thereby achieving optimal economic dispatch for distributed distribution adjustment.

Distribute Results: Distribute the results to all nodes so that they can update their local status and perform the next calculation.

Because the distributed energy can be added or withdrawn at any time, when a new node is added, the node broadcasts its own parameters and the current power value, requesting that the new node be added and the power redistribution be completed. Based on the PBFT mechanism, each node adds new node parameters to the on-chain storage, and adjusts the power based on the criteria to ensure that the system runs in the optimal state. When a node exits, similar operations are adopted to update the data on the chain and adjust the power [18].

The execution process of distributed scheduling is a dynamic and real-time process that requires continuous monitoring and adjustment to adapt to the complex and ever-changing operating environment of the power system.

3.5. Experimental verification. To verify the effectiveness of the proposed algorithm, an experimental simulation is conducted to verify that four DERs are distributed in a VPP system of the type Micro Gas Generators (MGGs), which are connected through a blockchain network.Each MGG is a node in the network. It is assumed that the network has a certain network delay, but the operation parameters and initial power of the point-to-point direct communication unit are guaranteed as shown in Table 3.2.

4. Results and discussion. In the presence of network delay, the change of consistency variable λ is tested. As can be seen from Figure 4.1, when the initial $t=1$, the λ values of each unit are different, which does not meet the constant consumption micro rate increase criterion, and the system does not operate in the optimal state. After a PBFT consensus, λ variable reaches the same value at $t=8$, and the system reaches the optimal operating state[19].

Table 3.2: Unit operating parameters and initial power

electric generator	a_i	b_i	c_i	P_i /KW
MGG1	1.256	1.355	8.812	30
MGG2	1.099	1.293	4.876	35
MGG3	0.942	1.220	1.220	70
MGG4	1.079	1.276	8.831	40

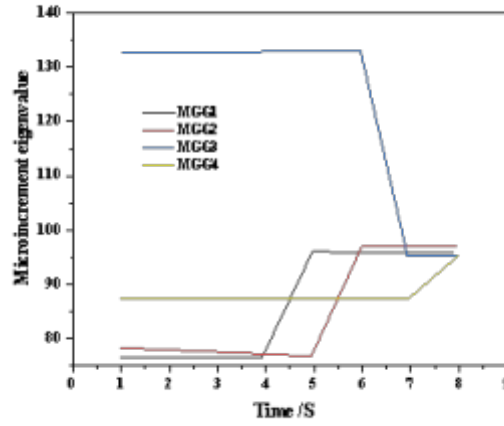


Fig. 4.1: The change of λ value of the microincrement feature

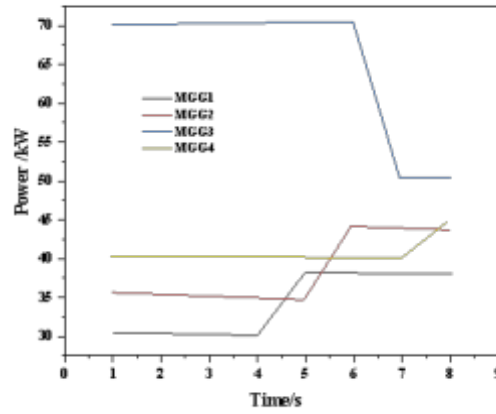


Fig. 4.3: Power variation of each unit

Figure 4.1 shows the comparison between the total load power P_{LD} and the total power $\sum(P_i)$ of the generating unit in the process of consensus.

Figure 4.3 shows the power adjustment of each mgg unit, and the final power value remains stable in the working state[20].

5. Conclusion. A research on distributed scheduling of virtual city power plants based on blockchain technology is proposed. The virtual power plant will be one of the last models of the future power of the internet, realize the reality of the agreement of resources in a wide range, and promote the use of electricity share. For the optimal operation of distributed virtual power plants, the fully distributed operation control of virtual power plants is realized by adopting the equal consumption and small increase principle and combining the distributed decentralized autonomous blockchain technology. In addition, the virtual power plant has some fault-tolerance capacity for situations such as node communication failure, which improves the safety level of the electric power. Specifically, virtual power plants typically adopt a distributed architecture, where multiple nodes form a network, and each node can transmit and process information. Therefore, in the event of a node failure or communication interruption, other nodes can still continue to operate and interact, thus ensuring the continuity and stability of the system.

6. Acknowledgement. Thanks for the project supported by the Science and technology projects from State Grid Corporation"Research on the interactive operation and transaction support technology of flexible resources on customer side" (5400-202019491A-0- 0-00)

REFERENCES

- [1] Huang, Y. , Jiang, Y. , & Wang, J. . (2021). Adaptability evaluation of distributed power sources connected to distribution network. *IEEE Access*, PP(99), 1-1.
- [2] Hu, J. , Liu, X. , Shahidehpour, M. , & Xia, S. . (2021). Optimal operation of energy hubs with large-scale distributed energy resources for distribution network congestion management. *IEEE Transactions on Sustainable Energy*, PP(99), 1-1.
- [3] Al-Haboobi, A. , & Kecskemeti, G. . (2023). Developing a workflow management system simulation for capturing internal iaas behavioural knowledge. *Journal of Grid Computing*, 21(1), 1-26.
- [4] Gradwohl, C. , Dimitrievska, V. , Pittino, F. , Muehleisen, W. , & Kienberger, T. . (2021). A combined approach for model-based pv power plant failure detection and diagnostic. *Energies*, 14(5), 1261.
- [5] Lee, S. W. , & Sim, K. B. . (2021). Design and hardware implementation of a simplified dag-based blockchain and new aes-cbc algorithm for iot security. *Electronics*, 10(9), 1127.
- [6] Li, R. , & Wan, Y. . (2021). Analysis of the negative relationship between blockchain application and corporate performance. *Mobile Information Systems*, 2021(7), 1-18.
- [7] Gaybullaev, T. , Kwon, H. Y. , Kim, T. , & Lee, M. K. . (2021). Efficient and privacy-preserving energy trading on blockchain using dual binary encoding for inner product encryption. *Sensors*, 21(6), 2024.
- [8] Wang, H. , Ma, S. , Guo, C. , Wu, Y. , & Wu, D. . (2021). Blockchain-based power energy trading management. *ACM Transactions on Internet Technology*, 21(2), 1-16.
- [9] Song, Z. , Zhang, X. , & Liang, M. . (2021). Reliable reputation review and secure energy transaction of microgrid community based on hybrid blockchain. *Wireless Communications and Mobile Computing*, 2021(8), 1-17.
- [10] Shah, C. , King, J. , & Wies, R. W. . (2021). Distributed admm using private blockchain for power flow optimization in distribution network with coupled and mixed-integer constraints. *IEEE Access*, PP(99), 1-1.
- [11] Zhang, X. , Zhao, Z. , Guo, Z. , & Zhao, W. . (2022). Research on machining parameter optimization in finishing milling with multiple constraints:. *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, 236(6-7), 968-980.
- [12] Jamil, F. , Iqbal, N. , Imran, Ahmad, S. , & Kim, D. H. . (2021). Peer-to-peer energy trading mechanism based on blockchain and machine learning for sustainable electrical power supply in smart grid. *IEEE Access*, PP(99), 1-1.
- [13] Adhvaryu, P. K. , & Adhvaryu, S. . (2021). Static optimal load flow of combined heat and power system with valve point effect and prohibited operating zones using krill herd algorithm. *Energy Systems*, 12(1), 133-156.
- [14] Bendadou, A. , Kalai, R. , Jemai, Z. , & Rekkik, Y. . (2021). Impact of merging activities in a supply chain under the guaranteed service model: centralized and decentralized cases. *Applied Mathematical Modelling*, 93(2), 509-524.
- [15] Anwar, W. S. , Abdel-Maksoud, F. M. , Sayed, A. M. , Abdel-Rahman, I. A. M. , Makboul, M. A. , & Zaher, A. M. . (2023). Potent hepatoprotective activity of common rattan (*calamus rotang l.*) leaf extract and its molecular mechanism. *BMC Complementary Medicine and Therapies*, 23(1), 1-10.
- [16] Din, H. R. , & Sun, C. H. . (2022). Centralized or decentralized bargaining in a vertically-related market with endogenous price/quantity choices. *Journal of Economics*, 138(1), 73-94.
- [17] Mirzaei, E. , & Hadian Dehkordi, M. . (2022). Simorgh, a fully decentralized blockchain-based secure communication system. *Journal of Ambient Intelligence and Humanized Computing*, 13(8), 3903-3921.
- [18] Parmentola, A. , Petrillo, A. , Tutore, I. , Felice, F. D. , & Welford, R. . (2022). Is blockchain able to enhance environmental sustainability? a systematic review and research agenda from the perspective of sustainable development goals (sdgs). *Business Strategy and the Environment*, 31(1), 194-217.
- [19] Sreenivasulu, G. , & Balakrishna, P. . (2021). Optimal dispatch of renewable and virtual power plants in smart grid environment through bilateral transactions. *Electric Power Components and Systems*, 49(4-5), 488-503.

- [20] Ren, L. , Liu, L. , Wu, Z. , Shan, D. , Pu, L. , & Gao, Y. , et al. (2022). The predictability of orthodontic tooth movements through clear aligner among first-premolar extraction patients: a multivariate analysis. *Progress in Orthodontics*, 23(1), 1-12.

Edited by: Bradha Madhavan

Special issue on: High-performance Computing Algorithms for Material Sciences

Received: Jan 16, 2024

Accepted: Mar 21, 2024



A NONLINEAR CONVOLUTIONAL NEURAL NETWORK ALGORITHM FOR AUTONOMOUS VEHICLE LANE LINE DETECTION

KANHUI LYU*

Abstract. The traditional lane line detection algorithm relies on artificial design features, which has poor robustness and cannot cope with the complex urban street background. With the rise of deep learning technology, the algorithm model with convolutional neural network as the mainstream is widely used in the field of computer vision, which provides a new idea for lane line detection. In order to improve the disadvantages of traditional lane line detection methods that are vulnerable to environmental impact and poor robustness, a nonlinear convolution neural network algorithm for driverless lane line detection is proposed. Firstly, the pretreatment method of extracting the region of interest and enhancing the contrast of lane lines is used to reduce the unnecessary image background and enhance the feature details of the image. Existing deep learning-based lane line detection algorithms still have difficulties. First, accumulated wear and tear will cause lane line to fade and fade; roadside trees and buildings can interfere with the performance of lane line detection algorithm. In addition, compared with the pixels of the whole picture, the lane line pixels are too few, and the deep convolution neural network of layer convolution is easy to lead to the loss of details. In addition, when the traffic flow is large, the lane line is easily blocked, which makes it more difficult to detect the lane line. Then the model is built based on the lane line image features extracted by CNN, and the DBSCAN clustering algorithm is used to post-process the lane line segmentation model; Finally, the least square method is used to fit the quadratic curve of the pixel peak points of the lane line, and the fitting results are regressed to the original image. The experimental results show that the accuracy and recall of the lane line detection model verification set are 91.3% and 90.6%, respectively, indicating that the model has a good segmentation effect. It is proved that the lane line detection method based on CNN combined with post-processing can effectively reduce the defects of artificial experience, and has better robustness and accuracy than the traditional lane line detection method.

Key words: Lane line detection, Convolution neural network, Deep learning, clustering algorithm

1. Introduction and examples. With the rapid development of urban traffic, traffic safety has become increasingly important. By the end of June 2017, the number of cars owned reached 205 million, and the number of car drivers reached 328 million, cars have become one of the most commonly used vehicles in our lives. Driving safety has become one of the hot issues that people are most concerned about. During the driving process, the road conditions and vehicle conditions are complex, and the driver is nervous and prone to fatigue. According to statistics, 66% of drivers are prone to drowsiness when driving alone for a long time. This shows the importance of lane line deviation warning [1].

As a key step in driverless technology, lane line detection is an important component of the sensing module. The study of lane line detection algorithm has important research value and application significance in the information exchange, traffic path planning and traffic safety accident avoidance. In the autonomous driving technology, more and more domestic and foreign researchers have carried out detailed and rich research on the lane line detection algorithm, and have achieved fruitful results. Therefore, there is a need for a lane line warning system to send out a danger warning to the driver to reduce the accident rate. This system designed to help the driver in the driving process is called the auxiliary driving system (ADAS). The goal of the system is to detect lane markings and alert the driver when he leaves the lane. In recent years, the semiconductor industry has developed and made great progress, small and powerful electronic devices are widely used in vehicles [2]. These devices can perform complex calculations and provide hardware environment protection for the auxiliary driving system. Gives the driver a safety warning at the right time. Intelligent transportation is included in the 13th Five-Year Plan, as one of the core functions of assisted driving and automatic driving, lane line detection is of self-evident importance. Lane line detection is not only related to our daily travel safety, but also an important part of development planning [3].

* Institute of Information Technology, Zhejiang Financial College, Hangzhou, Zhejiang, 310018, China (KanhuiLyu@126.com)

Lane line detection algorithm can be roughly divided into two categories: detection algorithm based on traditional digital image processing method, such as Hough transform, and lane line detection algorithm based on deep learning. Starting from the focus of the detection algorithm, the traditional lane line detection algorithm can be divided into two categories: one is the model-based detection algorithm based on the mathematical model such as curve or straight line, and the lane line detection algorithm based on the color, texture, shape and direction of the lane line.

2. Literature review. Convolution neural networks have been offered for over 20 years. The original neural network solution was also developed as a different type of neural network solution to solve problems in different research fields. It is used for image classification and image recognition in image processing. Lane line detection as one of the core functions of auxiliary driving and autopilot, the technology can in the process of driving real-time, accurate detection lane line, when the vehicle deviates from the current lane, warning, the driver operation error can be corrected, can effectively reduce the accident rate caused by the operation error and fatigue driving, so as to ensure the driver's personal and property safety. Therefore, lane line detection is crucial for the safety of drivers while driving. Neural networks are the first to be used to understand the environment of a self-driving car, such as recognizing traffic lanes captured by cameras, recognizing driving zones, and recognizing problems around the area take pictures from the face and the connection between the camera and the cap. In recent decades, many countries have invested in the direction of inquiry and the direction of research questions has become very good. Traditional line detection algorithms capture and filter the image and then use changes in line brightness to divide the lines [4].

Computer computing power and massive amounts of data have driven the development of deep learning technology, which has been applied to various fields, such as speech and images. Among them, the application of deep convolutional neural network to images is a negligible branch.

The line detection method based on neural network solution is to draw the lines of multi-dimensional scene image according to the data set, input them into the design of neural network solution, and train the neural network solution to make it clear.

The main purpose of lane line detection is to identify the lane lines from the original images. More specifically, the algorithm is used to extract the coordinates of the pixels belonging to the lane lines in the image, and then post-processing the pixels to finally obtain the actual position of the lane line in the original image.

Some assumptions about the lines on the road map. As an important part of raising environmental awareness, discovery line has been studied by researchers at home and abroad and has achieved a lot of results. The traditional line detection system can detect the line quickly and in real time, but the stability is poor, different lines of the line detection accuracy is low.

Lane lines have innate structural features, which are long and thin in shape. For a high-resolution RGB image, the proportion of lane line pixels in the whole image element tends to be small. After the deep convolutional neural network, the details of lane lines are easy to lose, resulting in the detection of lane lines.

In recent years, the rapid development of computer algorithm, gradually applied to the detection line. However, because the neural network scheme has more parameters and a large amount of calculation, the hardware cost of the detection line using the neural network scheme is high [5]. Haris, M improved line detection in complex environments, a method combined with visual data with wide distribution [6]. Ghazal, T. M proposed handwritten text recognition techniques based on neural network technology [7]. Dong, Y proposed a weighted fusion of convolutional neural network and graph attention network (WFCG) for HSI classification based on the characteristics of super-pixel-based GAT and pixel-based CNN, which proved to be successful [8].

Semantic segmentation network based on convolutional neural network for each pixel image detection, can be applied to different forms of target detection task, detection of details is more fine, so in dealing with different road scene also has advantages, such as in the car to change lanes or turn, lane line shape and structure has changed, for semantic segmentation network, can be very well processed, achieve better detection effect.

In order to improve the disadvantages of traditional lane line detection methods on environmental impact and poor robustness, a nonlinear convolutional neural network algorithm for unmanned lane line detection is proposed.

Considering that traditional line detection is easily affected by the environment and needs manual extraction,

a line detection method based on CNN is proposed. Plus the post-treatment process, it has the advantages of no manual adjustment, multiple applications and good effects.

3. Research methods.

3.1. Convolution neural network. Machine learning is an important part of the field of artificial intelligence, through the theory of probability, statistics, biology and artificial intelligence problems abstract into mathematical models, so that the model has similar to human learning ability, iteratively adjust the model parameters, to optimize the model effect, machine learning classical algorithms, including neural network, support vector machine, K mean clustering, DBSCAN and logical regression, etc. Among them, the neural network mimics the image process of the human visual system in processing pupil intake.

Compared with neural network, convolution neural network has changed the connection mode of neurons, using convolution operation and pooling operation. Convolution layer, activation layer and pooling layer are essential components of convolution neural network. The function of convolution layer is to convolution the input of convolution neural network to extract the features of the input image. The convolution process can be understood as a filtering process, a convolution kernel is a window filter, in the network training process, the convolution kernel of a user-defined size is used as a sliding window to convolution the input data [9].

The edge length of the image output by the convolution operation=(input image edge length convolution core length+1)/step length, if it cannot be divided, the result will be rounded up. Sometimes in order to ensure that the side length of the output image of a convolution operation is the same as the side length of the input image, all-zero filling will be performed around the input image, and the side length of the output image is equal to the side length of the input image divided by the step length, if it is not possible to divide and take an integer up, assume that the input image is a three-channel image of 52 * 52 * 3, the convolution kernel size is 5 * 5, and the depth is 3, that is, a convolution kernel of 5 * 5 * 3, if the convolution step is 1, the characteristic image of 48 * 48 * 1 will be output[10]. The parameters in the convolution kernel are the weights in the network, these parameters are unchanged in a forward calculation process and will not change because of the position of the convolution kernel in the input data, this is the weight sharing property of convolutional neural network. Finally, the image features are recognized by combining different features. The activation layer compensates for the linear operation of convolution, adds nonlinear elements, and makes the convolution neural network have the ability to learn nonlinear.

3.2. Image data preprocessing. The purpose of lane line image data preprocessing is to enhance the features of the target in the image, so that deep learning can better learn the feature information and obtain a model with stronger generalization ability.

3.2.1. ROI extraction. In order to remove redundant image information and improve the speed and detection effect of network training, a fixed ROI region extraction method is adopted, through OpenCV, the original image with the resolution of 720E480 is clipped to the ROI area of 720E240.

3.2.2. Brightness contrast transformation. Contrast is the ratio of the whitest and darkest brightness units. Adjusting the contrast can make the image more vivid. Take a color channel of an RGB image as an example, take the current color depth value I of the pixel as the abscissa, and output the transformed color depth value O as the ordinate to establish the coordinate system, each pixel of the traditional RGB format image can use a value of 0~255 to represent its color depth [11].

When the brightness and contrast of the image are modified at the same time, the transformation equation is as follows 3.1:

$$O = KI + J \quad (3.1)$$

where J is the increased value of image brightness, and K is the original color depth scale value. According to the actual situation, several attempts have been made to determine the added value of image contrast and brightness adjustment, and the original image has been adjusted by $K=1$ and $J=20$.

3.3. Lane line detection algorithm. The whole process of anchor-based lane line detection model can be compared to the process of region-based target detection algorithm. First, the feature extraction was

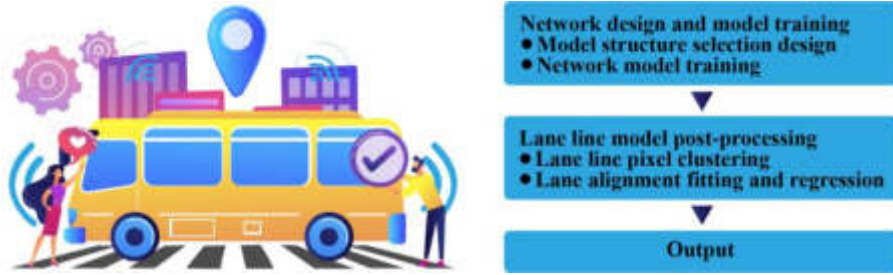


Fig. 3.1: Design steps of lane line detection algorithm

performed by placing anchor at each position of the feature graph to obtain the resulting feature vector. Finally, the feature vectors were regressed and classified separately.

The lane line detection algorithm established by the author is mainly divided into two steps, namely, network design and model training, and lane line model post-processing, as shown in Figure 3.1.

3.3.1. Model structure design. As a marker line on the road surface, it is a two-dimensional goal. Compared with the point cloud data of lidar, the model constructed by GPS and high-precision map, the images obtained by the visual sensor can express the two-dimensional features of the lane line more effectively. Moreover, compared with lidar and high-precision maps, visual sensors are low cost and cost-effective. Therefore, the lane line detection algorithm is mostly based on vision sensor, that is, the detection algorithm based on computer vision, which is also widely used in the industry to detect lane lines.

Fine-tune the VGG16 network, add a custom network on the VGG16 base network that has been trained on other classification issues, and then freeze the base network, re-train the previously added user-defined part and unfreeze some layers of the base network, and finally jointly train the unfrozen layer and the user-defined added part. The Softmax layer is added to the last layer of the network model output to modify the model framework, so as to achieve the probability distribution of multi-channel pixel points of the model output results, which is convenient for the clustering algorithm in the post-processing algorithm to cluster the lane lines [12].

3.3.2. Lane line pixel clustering. By clustering the core points and density reachable points, the area with enough high density is divided into one category. At the same time, DBSCAN can also recognize sparse noise data. The pixel distribution of the output image of the lane line segmentation model is clustered, and the image pixels belonging to the same lane line are classified into the same category. Convert the pixel probability distribution of lane line segmentation image into 3D visualization image. Sort the pixel probability distribution image output by lane line segmentation according to the axis value, and take the y item accuracy data corresponding to the image model position $x=240, 250... 480$, then extract the peak points with probability value greater than 0.5 and cluster them, the output image of the peak points with probability value greater than 0.5 is shown in Figure 3.2.

3.3.3. Lane line fitting and regression. Lane line fitting is a square theory of lane line curve according to a specific number of sampling points. Lane line fitting is to fit the pixel probability peak points of the reclassified lane line segmentation model, and then obtain the track parameter equation of the lane line[13].

Curve fitting refers to obtaining an approximate curve $y = \rho(x)$ for a given m data points $p_i(x_i, y_i), i = 1, 2, \dots, m$, so as to minimize the deviation between the curve $y = \rho(x)$ and the real curve $y = f(x)$. The deviation δ_i of $y = \rho(x)$ at point p_i is calculated as follows 3.2:

$$\delta_i = \rho(x_i) - y_i \quad (3.2)$$

where δ_i is the deviation.

The deviation minimization calculation formula is as follows 3.3:

$$\min \sum_{i=1}^m \delta_i^2 = \sum_{i=1}^m (\rho(x_i) - y_i)^2 \quad (3.3)$$

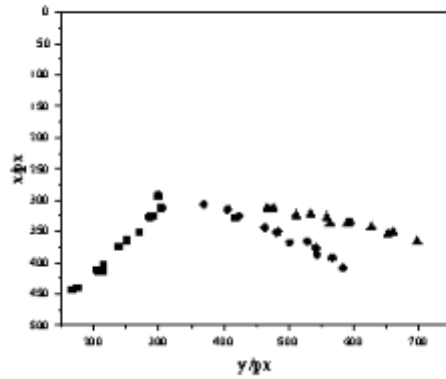


Fig. 3.2: Peak output image with probability value greater than 0.5

The least square method takes the quadratic equation as the fitting curve, and selects the fitting curve with the smallest sum of squares. The least square method is used to perform polynomial nonlinear fitting on the given m sample points, so that the approximate curve $\delta_i = \rho(x_i) - y_i$ of $y_i = f(x)$ passes through these sample points.

Assume that the polynomial of lane line to be fitted is the following equation 3.4:

$$p(x) = a_0 + a_1x + \dots + a_kx^k = \sum_{k=0}^n a_kx^k \tag{3.4}$$

The sum of squares of deviations of all pixel points reaching the approximate curve is as follows 3.5:

$$\begin{aligned} R &= \sum_{i=1}^m [y_i - (a_0 + a_1x_i + \dots + a_kx_i^k)]^2 \\ &= \sum_{i=1}^m (y_i - \sum_{k=0}^m a_kx_i^k)^2 \end{aligned} \tag{3.5}$$

Solve polynomial $a_i = (i = 1, 2, \dots, k)$ to obtain the minimum value of equation 3.6, expressed as:

$$\begin{aligned} \min R &= \sum_{i=1}^m [y_i - (a_0 + a_1x_i + \dots + a_kx_i^k)]^2 \\ &= \sum_{i=1}^m (y_i - \sum_{k=0}^m a_kx_i^k)^2 \end{aligned} \tag{3.6}$$

In order to solve the extreme value of the multivariate function of the parameter a_1, a_1, \dots, a_k , the partial derivative formula of the variable $a_i = (i = 1, 2, \dots, k)$ is obtained as follows 3.7:

$$\sum_{k=0}^n \left(\sum_{i=1}^m a_kx_i^{j+k} \right) = \sum_{i=1}^m x_i^j y_i, \quad j = 0, 1, \dots, n \tag{3.7}$$

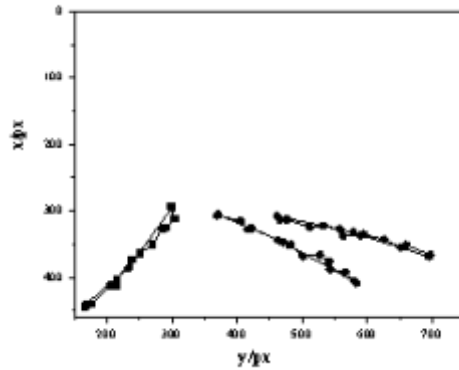


Fig. 3.3: Quadratic curve fitting image

Change the equation into matrix form as follows 3.8:

$$\begin{bmatrix} m & \sum_{i=1}^m x_i & \cdots & \sum_{i=1}^m x_i^n \\ \sum_{i=1}^m x_i & \sum_{i=1}^m x_i^2 & \cdots & \sum_{i=1}^m x_i^{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^m x_i^n & \sum_{i=1}^m x_i^{n+1} & \cdots & \sum_{i=1}^m x_i^{2n} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^m y_i \\ \sum_{i=1}^m x_i y_i \\ \vdots \\ \sum_{i=1}^m x_i^n y_i \end{bmatrix} \tag{3.8}$$

The following formula 3.9 is obtained by simplifying the Vandermonde matrix:

$$\begin{bmatrix} 1 & x_1 & \cdots & x_1^k \\ 1 & x_2 & \cdots & x_2^k \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^k \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{3.9}$$

The solved coefficient matrix Q formula is 3.10:

$$Q = X^{-1}Y \tag{3.10}$$

Equation 3.10 is the relationship of the fitting curve, the least square method is used to fit the quadratic curve of the probability peak point of the lane line pixel after clustering classification, the fitting image is shown in Figure 3.3. Regression the fitting curve to the original lane line image [14].

4. Result analysis. The author’s algorithm is based on the Keras deep learning framework, using Python language, OpenCV computer vision processing library, and tested on Ubuntu 18.04L TS system [15]. Keras is a deep learning framework based on theano / tensorflow. Is a high-level neural network API that supports fast experiments and can quickly translate your idea into results.

A random gradient is used to train a neural network model. The learning threshold was set to 0.01, the size was set to 16, and the learning period (Epoch) was set to 100. During the training process, introducing some obstacles during the training process indicates the state of the emerging model[16]. Based on the values of the output parameters of the training process, the OpenCV graphing function is used to plot the average accuracy and loss curves as a function of the number of iterations during training. The missing curve with 100 sampling iterations is shown in Figure 4.1, and the mean true curve with 100 sampling iterations is shown in Figure 4.2.

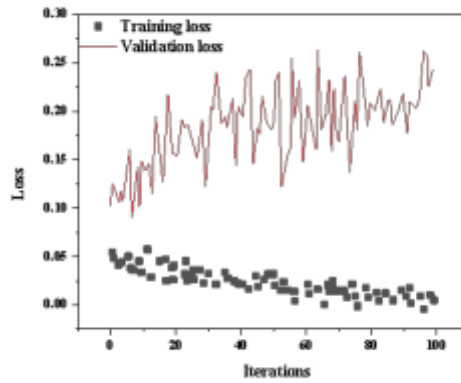


Fig. 4.1: Training loss change curve

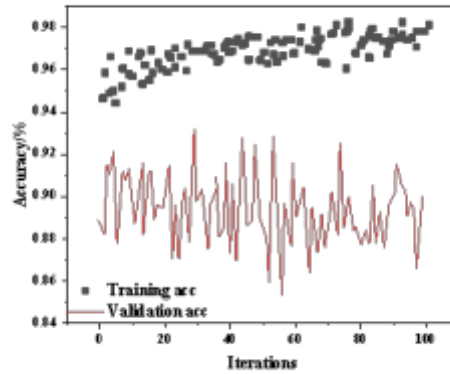


Fig. 4.2: Training accuracy change curve

The accuracy rate and recall rate of the model are defined as the basis for judging the segmentation effect of the model, that is, the greater the accuracy rate and recall value, the better the segmentation effect of the model. The calculation formula of average accuracy acc is:

$$acc = \frac{C_{img}}{T_{img}} \quad (4.1)$$

where, C_{img} is the correct number of pixels for segmentation, and T_{img} is the total number of pixels marked for the entire image[17].

All pixels labeled as lane line image areas are divided into lane line pixels and non-lane line pixels. The ratio of the two is the recall rate $recall$. The calculation formula is 4.1:

$$recall = \frac{TP}{TP + FN} \quad (4.2)$$

where, TP is the correctly predicted lane line pixel point, and FN is the incorrectly predicted lane line pixel point.

Table 4.1: Comparison between accuracy of different models and time consumption of single frame image

Model	Accuracy/%	Time consumption/ms
K-means	84.25	54
CNN+Hough	87.08	60
SegNet	90.06	50
VGG16	91.3	45

After calculation, the average accuracy rate and recall rate of the final model validation set are 91.3% and 90.6% respectively, indicating that the model has good segmentation effect. Then the lane lines in different scenes are detected and recognized[18].

For the same experimental sample, compare the detection accuracy and single frame image time of the author's model with other models, and the results are shown in Table 4.1.

The experimental results show that the accuracy of the traditional lane line detection method K-means is far less than that of other model methods combined with deep learning. Although the method of CNN combined with Hough transform improves the detection accuracy, it takes the longest average time to process a single frame image, compared with this method, the accuracy and average time of lane line detection of SegNet model have been greatly improved. Compared with the other three models, the VGG16 model has significantly improved in accuracy and single-frame image processing speed, achieving better image segmentation effect [19,20].

The features of the lines were studied using CNN in special image extraction operation and the DBSCAN clustering algorithm was used for line classification, which improves the accuracy of the model and the matching results. About this method. Experimental results show that the combination of CNN and post-processing algorithm is more accurate and reliable than conventional line detection methods.

5. Conclusion. Lane line detection is an important part of auxiliary driving and automatic driving. Lane line deviation alarm and lane line maintenance can correct the careless operation of drivers in time, reduce traffic accidents caused by wrong operation and fatigue driving, so as to effectively guarantee driving safety and reduce driving complexity. In complex road scenarios, lane line detection algorithms need to be robust and real-time.

In this paper, we take advantage of CNN in special image extraction operation to study the features of the line, and perform the post-line classification process using DBSCAN clustering algorithm, which improves the accuracy of the model and improves the matching results. about the method. The traditional lane line detection method K-means has much less precision than other model methods combining deep learning. Although the CNN method combined with Hough transformation improved the detection accuracy, the average time was the longest for processing single-frame images, and the lane line detection accuracy and average time of the SegNet model were significantly improved compared with this method. Compared with the other three models, the VGG 16 model showed a significant improvement in both accuracy and single-frame image processing speed, achieving better image segmentation.

Experimental results show that the combination of CNN and post-processing algorithm is more accurate and reliable than traditional line detection methods. The line of investigation suggested by the author is therefore of some value. However, the algorithm still has some disadvantages: on the one hand, due to the limitations of the image's own hardware during training, the training time increases, and the error resulting from training increases, which makes it impossible to learn all the network models. image features; On the other hand, when the output data of the lane-line segmentation model is postprocessed, a small amount of pixel data may be filtered out, leading to the failure to fit and missing part of the lane lines. The above questions are the key direction of the next research step. Because the input image resolution of the present network is not high enough, the localization information is not accurate enough. Moreover, when the network return is too deep, the speed of the network cannot meet the needs of real-time detection, so the results of both detection accuracy and speed cannot be realized. Next, we hope to innovate in the network input mode and network structure, so as to ensure the network detection speed and increase the resolution of the image, so as to better realize the

lane line detection.

REFERENCES

- [1] Alam, M. Z., Kelouwani, S., Boisclair, J., & Amamou, A. A. (2022). Learning Light fields for improved lane detection. *IEEE Access*, 11(3), 271-283.
- [2] Dong, Y., Patil, S., van Arem, B., & Farah, H. (2023). A hybrid spatialtemporal deep learning architecture for lane detection. *ComputerAided Civil and Infrastructure Engineering*, 38(1), 67-86.
- [3] Gopal, K. V., Rohith, C., Siddhartha, D., & Mahule, S. (2022). Lane detection on roads using computer vision. *International journal of engineering technology and management sciences*, 6(4), 8-15.
- [4] Dewangan, D. K., & Sahu, S. P. (2023). Lane detection in intelligent vehicle system using optimal 2-tier deep convolutional neural network. *Multimedia Tools and Applications*, 82(5), 7293-7317.
- [5] Shirke, S., & Udayakumar, R. (2022). Hybrid optimisation dependent deep belief network for lane detection. *Journal of Experimental & Theoretical Artificial Intelligence*, 34(2), 175-187.
- [6] Haris, M., Hou, J., & Wang, X. (2022). Lane lines detection under complex environment by fusion of detection and prediction models. *Transportation research record*, 2676(3), 342-359.
- [7] Ghazal, T. M. (2022). Convolutional neural network based intelligent handwritten document recognition. *Computers, Materials & Continua*, 70(3), 4563-4581.
- [8] Dong, Y., Liu, Q., Du, B., & Zhang, L. (2022). Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification. *IEEE Transactions on Image Processing*, 31(9), 1559-1572.
- [9] Poliak, M., Jurecki, R., & Buckner, K. (2022). Autonomous vehicle routing and navigation, mobility simulation and traffic flow prediction tools, and deep learning object detection technology in smart sustainable urban transport systems. *Contemporary Readings in Law and Social Justice*, 14(1), 25-40.
- [10] Ali, A., Zhu, Y., & Zakarya, M. (2022). Exploiting dynamic spatio-temporal graph convolutional neural networks for citywide traffic flows prediction. *Neural networks*, 145(11), 233-247.
- [11] Tian, C., Yuan, Y., Zhang, S., Lin, C. W., Zuo, W., & Zhang, D. (2022). Image super-resolution with an enhanced group convolutional neural network. *Neural Networks*, 153(15), 373-385.
- [12] Hassan, S. M., & Maji, A. K. (2022). Plant disease identification using a novel convolutional neural network. *IEEE Access*, 10(3), 5390-5401.
- [13] Sharma, A. K., Tiwari, S., Aggarwal, G., Goenka, N., Kumar, A., Chakrabarti, P., ... & Jasiski, M. (2022). Dermatologist-level classification of skin cancer using cascaded ensembling of convolutional neural network and handcrafted features based deep neural network. *IEEE Access*, 10(8), 17920-17932.
- [14] Torres, M., & Cantú, F. (2022). Learning to see: Convolutional neural networks for the analysis of social science data. *Political Analysis*, 30(1), 113-131.
- [15] Tulbure, A. A., Tulbure, A. A., & Dulf, E. H. (2022). A review on modern defect detection models using DCNNsDeep convolutional neural networks. *Journal of Advanced Research*, 35(6), 33-48.
- [16] Shlezinger, N., Eldar, Y. C., & Boyd, S. P. (2022). Model-based deep learning: On the intersection of deep learning and optimization. *IEEE Access*, 10(8), 115384-115398.
- [17] Aggarwal, K., Mijwil, M. M., Al-Mistarehi, A. H., Alomari, S., Gök, M., Alaabdin, A. M. Z., & Abdulrhman, S. H. (2022). Has the future started? The current growth of artificial intelligence, machine learning, and deep learning. *Iraqi Journal for Computer Science and Mathematics*, 3(1), 115-123.
- [18] Lauriola, I., Lavelli, A., & Aiolli, F. (2022). An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing*, 470(9), 443-456.
- [19] Zhan, Z. H., Li, J. Y., & Zhang, J. (2022). Evolutionary deep learning: A survey. *Neurocomputing*, 483(8), 42-58.
- [20] Mo, Y., Wu, Y., Yang, X., Liu, F., & Liao, Y. (2022). Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, 493(9), 626-646.

Edited by: Bradha Madhavan

Special issue on: High-performance Computing Algorithms for Material Sciences

Received: Jan 16, 2024

Accepted: Apr 5, 2024



DIMENSION EXTRACTION OF REMOTE SENSING IMAGES IN TOPOGRAPHIC SURVEYING BASED ON NONLINEAR FEATURE ALGORITHM

YANI WANG,* YINPENG ZHOU† AND BO WANG ‡

Abstract. In order to solve the problem of inaccurate image feature extraction caused by traditional extraction methods, this paper proposes a remote sensing image size extraction method based on nonlinear multi feature fusion for topographic maps. In this paper, SVM and DS evidence theory are combined to extract image features and classify pre processed remote sensing images. Based on the classification results, basic probability distributions are constructed, and a DS fusion algorithm using matrix analysis is introduced to simplify the complexity of decision level fusion algorithms; We use a multi feature fusion algorithm based on feature proximity, using the proximity vector formed by the attraction between the feature vector and the original graphics pattern as the fusion feature to complete the extraction of remote sensing image features. The simulation results show that after using this method, its soft threshold classifier outputs 0.9865, 0.9965, 0.7852, 0.9921, 0.9847, 0.6879, -0.5898, -0.5678, -0.6897, -0.4785. The algorithm in this paper can distinguish the shape features of terrain images well, and can extract the features of terrain images more accurately, which has strong feasibility.

Key words: Image feature extraction, Multi-feature fusion, Matrix analysis, Feature proximity, Feature vector

1. Introduction and examples. Remote sensing technology has become a comprehensive technology integrating multiple fields after decades of development since the 1960s [1]. Remote sensing technology refers to the observation of ground objects through some devices, obtaining relevant data, but not directly contacting the observation objects, and extracting and excavating the information from the obtained data. Different sensors have different spectral resolution. Generally, the imaging of remote sensing technology can be divided into the following categories according to the different resolution:

- ① Multispectral remote sensing image: the spectral resolution is within the range of $10-1\lambda$, and the multispectral image only contains 3-4 spectral bands, which contains less information;
- ② Hyperspectral remote sensing image: the spectral resolution is within $10-2\lambda$, often with tens to hundreds of bands, and contains rich information. This paper takes hyperspectral image as the main research object;
- ③ Ultra-spectral remote sensing image: the spectral resolution is within the range of $10-3\lambda$, and this paper does not involve ultra-hyperspectral image [2].

Hyperspectral remote sensing technology is one of the latest achievements of remote sensing technology. It scans objects and obtains relevant information through a series of narrow electromagnetic wave bands. The sensor forms a spectral image through hyperspectral remote sensing technology. The image contains continuous spectral information. Researchers can use this information to find some hidden material information, which is one of the advantages of hyperspectral remote sensing. Therefore, hyperspectral sensor imaging is a powerful assistant for researchers to identify substances in detail and accurately estimate the abundance of substances [3].

With the vigorous development of hyperspectral remote sensing technology in hardware, it is gradually recognized and accepted by more people, and the application field of remote sensing technology has been greatly expanded. But the puzzle behind the rapid development of remote sensing hardware technology is that a large number of remote sensing data are just cold numbers, which have not been fully utilized and mined, let alone used to serve mankind.

*Xi'an University, Shaanxi, Xi'an, 710065, China (YaniWang53@126.com)

†Institute of Surveying and Mapping Guizhou Geology and Mineral Exploration Bureau (Corresponding author, YinpengZhou7@163.com)

‡Shaanxi Geomatics Center, Ministry of Natural Resources Xi'an, China (BoWang28@126.com)

The data collected by hyperspectral remote sensing technology contains a lot of information, through which the spectral characteristics of the observed object can be closely linked with the spatial geographic information, and many potential characteristics of substances can be easily mined by analyzing hyperspectral data, which are important reasons why hyperspectral images are favored by more and more remote sensing experts. The hyperspectral remote sensing technology not only obtains the spectral characteristics of the ground object, but also preserves the relationship between the observed object and other surrounding ground objects. The resulting images contain rich information, which provides a strong support for us to use and analyze data.

2. Literature review. Buildings are an important component of urban areas. The technology of automatic extraction of buildings by computer is widely used in the fields of urban large-scale topographic map drawing, urban planning, urban geographic information system construction and military reconnaissance. Driven by these applications, many automatic building extraction methods have emerged. In recent years, automatic building extraction from remote sensing images has become a hot spot, attracting many scholars to discuss and study, and put forward many detection models and methods [4]. After summary, we can classify it into two kinds of methods: traditional building detection algorithm and building detection algorithm based on machine learning. Traditional building detection algorithms focus on describing the underlying features of buildings, such as building detection based on gray scale, contour, texture and other features or a simple combination of several underlying features, and building detection algorithms that add laser point clouds, DEM and other data sources [5]. Building detection algorithm based on gray level: This kind of method is mainly to analyze the gray level distribution in the image. Due to the different reflectivity of the building roof and other ground objects to the sunlight, there will be gray level differences, so the building will be extracted from the image using the region segmentation algorithm; Because most buildings have shadows, the gray contrast between the shadows around the buildings and the background is used to detect the existence of buildings.

Therefore, Li, X proposed a multivariate fusion voting network (MFFVoteNet) framework to improve the performance of 3D object detection in non-uniform and high-impact environments. Our method uses a point cloud and a synchronized RGB image as input to enable object detection in 3D space [6]. Shankar, K Review of WSI review process based on machine learning. First, the development status of WSI and CAD methods is presented. Second, we discuss WSI data reporting, segmentation, classification, and metrics for testing activities [7]. Wang, Z proposed a novel fusion model based on meta-heuristics for diagnosing COVID-19 using chest radiographs. The model includes various preplanning, extraction procedures and categories. Initially, Wiener filter (WF) technology was used for image processing [8].

This paper proposes a feature extraction method of remote sensing image based on SVM and multi-feature fusion. The simulation results show that the accuracy and speed of pixel extraction of the method proposed in this paper are much higher than those of traditional methods in the process of image extraction, which can be widely used and has strong feasibility.

3. Research methods.

3.1. Remote sensing image feature extraction principle. This is because traditional remote image feature extraction has many different parameters and is capable of extracting pixel features and recovering them in time and frequency. Therefore, we can use this model to analyze and extract the local features of remote sensing images during time and frequency changes, which is increasingly used after the remote sensing image is completed [9]. Traditional remote sensing image post-processing involves extracting the main features of multi-pixel lines from the original image by reconstructing and analyzing frequency pixels during two-dimensional frequency transfer to the area.

3.1.1. Pretreatment before feature extraction of remote sensing image.

(1) *Pixel domain subband analysis of remote sensing image.* For $P_0, (\Delta_s, s \geq 0)$ in the post-processing of captured image pixels. Restore the image pixels $(P_0f, \Delta_1f, \Delta_2f, \dots)$ of different sub-bands for the pixel range f , and the sub-band $\Delta_s f$ of multi-dimensional space contains the main features with the pixel space of 2^{-2s} .

(2) *Pixel smooth segmentation of remote sensing image.* In the process of restoring the remote sensing image pixels, set the set of $\omega_Q(x_1, x_2)$, and calculate the corresponding function ω_Q of a parameter of the restored image pixel within a two-dimensional space range of 2^{-2s} to obtain the set of Q similar results, as

follows 3.1:

$$Q = [k_1/2^s, (k_1 + 1)/2^s] \times [k_2/2^s, (k_2 + 1)/2^s] \quad (3.1)$$

Run such calculations for image pixel restoration Q of a specific algorithm. Assuming that all image post-processing $Q = Q(s, k_1, k_2)$, and the s similarity segmentation parameters k_1 and k_2 change, as shown in the following formula 3.2:

$$\Delta_s f \rightarrow (\omega_Q \Delta_s f)_{Q \in Q_2} \quad (3.2)$$

(3) *Normalization of pixels.* For a second order square, normalize $(T_Q f)(x_1, x_2) = 2^s f(2^s x_1 - k_1, 2^s x_2 - k_2)$ to calculate pixel f , and the result of Q parameter function calculated by 3.2 is $[0, 1]^2$, as shown in the following formula 3.3:

$$g_Q = (T_Q)^{-1}(\omega_Q \Delta_s f), Q \in Q_s \quad (3.3)$$

The normalization operation of pixels can collect the post-processed pixels and lay the foundation for image fusion extraction.

3.1.2. Extraction and reconstruction of remote sensing image features.

1) *Remote sensing image feature straight ridge synthesis.* After analyzing the straight ridge extracted from remote sensing image fusion, any calculation parameter is newly created from the orthogonal image pixel set, as shown in the following formula 3.4:

$$g_Q = \sum_{\lambda} \alpha(\lambda, Q) \rho_{\lambda} \quad (3.4)$$

2) *Feature normalization of remote sensing image.* The calculation in formula 3.4 can put any pixel into a suitable spatial position, as shown in formula 3.5:

$$h_Q = (T_Q)g_Q, Q \in Q_s \quad (3.5)$$

3) *Feature fusion and smooth synthesis of remote sensing image.* The result of formula 3.5 can be used as the inverse operation of image feature fusion boundary, as shown in formula 3.6:

$$\Delta_s f = \sum_{Q \in Q_s} \omega_Q - h_Q \quad (3.6)$$

4) *Time frequency subband reconstruction.* Incorporate formula 3.6, and use image pixel reconstruction formula to reconstruct and synthesize time and frequency, as shown in formula 3.7:

$$f = p_0(p_0 f) + \sum_{s>0} \Delta_s(\Delta_s f) \quad (3.7)$$

3.2. Image feature extraction method based on SVM and multi-feature fusion. Hard interval maximization support vector machine, also known as linear support vector machine, is the most primitive and simplest form of support vector machine [10]. As the basis of support vector machine theory (hereinafter referred to as SVM for short), it mainly aims at the problem of linearly separable dichotomous problems, as shown in Figure 3.1.

First, we combine SVM and DS evidence theory, extract three types of features of image shape, texture and fractal dimension after preprocessing the remote sensing image, and construct basic probability assignment with the SVM classification results of three types of nonlinear single features as independent evidence, and introduce DS fusion algorithm based on matrix analysis to simplify the complexity of decision level fusion algorithm [11]. Multi-feature fusion algorithm is introduced to calculate the spatial approximation of the original graphic pattern and feature vector, and the approximation is used as the fusion feature to complete the extraction of image features.



Fig. 3.1: Learning and application of classifier

3.2.1. Theoretical evidence of SVM . The main theoretical basis of the SVM calculation method is the certainty of the specific position of pixels in the dimension space of remote sensing image and the balance interval between each other during the post-processing of image capture. The spatial classification of any number of remote sensing image pixels in the shape, texture and fractal dimension of the image is carried out by using the formula of SVM classification and archiving. The specific method of discrimination is as follows 3.8:

$$f(x) = \text{sgn}\left(\sum_{x_i \in S_V}^n a_i y_i k(x_i, x) + b\right) \tag{3.8}$$

In the above formula, a_i is the Lagrange multiplier, S_V is the support vector, $k(x_i, x)$ is the kernel function, x_i and y_i are the basic support vector bases of the two calculation methods' functions, and b is the parameters determined according to the specific image pixel functions[12].

From the above formula, it can be seen that $k(k-1)/2$ SVM classification calculation methods are used to restore remote sensing image pixels. In the process of classifying image pixels, the computer defaults to counting once for each calculation, and sends data to all sets. Finally, the most counted image pixels are classified into a class of image pixel sets to build the theoretical evidence of SVM.

3.2.2. DS evidence theory. The basic principle of DS evidence theory is as follows: DS evidence theory fuses the trust functions corresponding to two or more evidence bodies and transforms them into a new function. The implementation process of DS evidence theory is described as follows:

Let Θ be the discriminant framework, define the function $m : 2^\Theta \rightarrow [0, 1]$ to satisfy the condition $M(\Phi) = 0$ (Φ is empty set), $\sum m(A) = 1(A \in 2^\Theta)$, and construct $m(A)$ according to the three characteristics of image shape, texture and fractal dimension, which is the fundamental probability assignment (BPA) on the framework Θ . $m(A)$ is the accurate level of trust in proposition A, and $M(\Phi)$ is the uncertainty of evidence. Assuming that m_1, m_2, \dots, m_n is BPA to distinguish different evidences on frame Θ , $m = m_1 \oplus m_2 \oplus \dots \oplus m_n$ can be converted into the following formula 3.9 and 3.10 according to $\sum m(A) = 1(A \in 2^\Theta)$:

$$m(A) = \sum_{A_1 \cap A_2 \cap \dots \cap A_n = A} \left(\prod_{1 \leq i \leq n} m_i(m_i(A)/(1-k)) \right) \tag{3.9}$$

$$k = \sum_{A_1 \cap A_2 \cap \dots \cap A_n = A} \left(\prod_{1 \leq i \leq n} m_i(m_i(A)) \right) \tag{3.10}$$

In the above formula, k is the uncertainty factor of credentials.

3.2.3. DS-synthesis algorithm with matrix analysis. This paper proposes a DS-synthesis algorithm based on matrix analysis to fuse the features of remote sensing images. In view of the situation of identifying

a target with n-type image features at the same time, the mutually independent fundamental trustable distribution values m_{ij} and uncertainty probability θ_{ij} of n-type features given to m image target categories can be described as the following formula 3.11:

$$M = \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1m} & \theta_1 \\ m_{21} & m_{22} & \cdots & m_{2m} & \theta_2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ m_{n1} & m_{n2} & \cdots & m_{nm} & \theta_n \end{bmatrix} \tag{3.11}$$

Because the sum of the mutually independent fundamental trustable distribution value m_{ij} and uncertainty probability θ_{ij} given to the target category of m images by the same image characteristics should be 1, therefore, the sum of the elements in each row of the matrix should meet the normalization condition, as shown in the following formula 3.12:

$$m_{i1} + m_{i2} + \cdots + m_{im}\theta_i = 1 \tag{3.12}$$

Under the condition of normalization, a new matrix R with $(m + 1) \times (m + 1)$ is obtained by multiplying the transposition of one row of the matrix with another row, and the DS synthesis of matrix analysis is realized[13]. The following formula 3.13:

$$R = M_i^T M_j \begin{bmatrix} m_{i1}m_{j1} & m_{i2}m_{j2} & \cdots & m_{i1}m_{jm} & m_{i1}\theta_j \\ m_{i2}m_{j1} & m_{i2}m_{j2} & \cdots & m_{i2}m_{jm} & m_{i2}\theta_j \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ m_{im}m_{j1} & m_{im}m_{j2} & \cdots & m_{im}m_{jm} & m_{im}\theta_j \\ \theta_i m_{j1} & \theta_i m_{j2} & \cdots & \theta_i m_{jm} & \theta_i m_j \end{bmatrix} \tag{3.13}$$

where the uncertainty factor k is the sum of the non-diagonal elements of the first $m \times n$ sub-matrices in the matrix, that is, the following formula 3.14:

$$k = \sum_{p \neq q} R_{pq}(p, q = 1, 2, \dots, m) \tag{3.14}$$

Fusion of matrix analysis and image features under the DS evidence theory m algorithm: this algorithm completes the matrix multiplication calculation of m+1-dimensional column vector and m+1-dimensional row vector in each implementation process. The time required to obtain the fusion result is $T((m + 1)^2)$, and the time required to obtain the fusion result is $T((m + 1)^2n)$, which is approximately linear with the number of characteristic types n[14].

3.2.4. Proximity of remote sensing image features. The feature extraction of remote sensing images is based on the fusion results of image features calculated by the -synthesis algorithm of matrix analysis. The feature space corresponding to the image to be classified is regarded as the core point of the type. The core point is attractive to all feature points in the feature space of the shape, texture and fractal dimension of the remote sensing image. The closer the feature point is to the core point, the stronger the attraction. The attraction of the core point of each image feature type is only effective in a certain spatial range.

Definition 1. In the feature space of the n-dimensional remote sensing image, the attraction between the core point x_i and the feature point x_j is defined as the following formula 3.15:

$$G_{ij} = e^{-d_{ij}^2/2\sigma^2} \tag{3.15}$$

where $d_{ij}^2 = x_i - x_j$ is the Euclidean distance of the two remote sensing image feature vectors, and σ is the parameter that controls the influence range of the core point of the image feature parameters. The core points of remote sensing image feature parameters are attractive to feature points, and the core points also attract each other[15]. The attractiveness index between the core points constitutes the proximity matrix G.

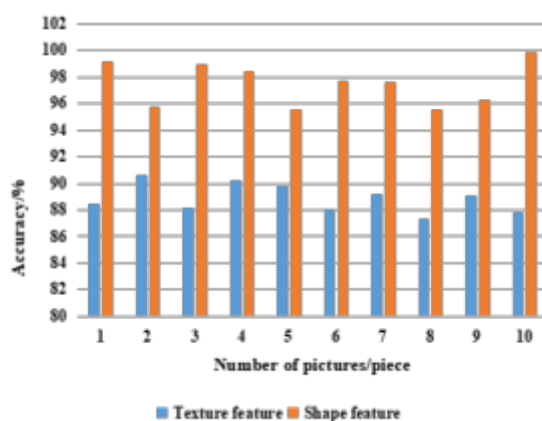


Fig. 4.1: Classification accuracy of target samples detected by SVM classification

Definition 2. Suppose there are m samples in the n -dimensional feature space, and the proximity matrix G is defined as the following formula 3.16:

$$G = \begin{bmatrix} G_{11} & G_{12} & \cdots & G_{1m} \\ G_{21} & G_{22} & \cdots & G_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ G_{m1} & G_{m2} & \cdots & G_{mm} \end{bmatrix} \quad (3.16)$$

Line i in G represents the attraction between the prototype model of image feature type and other prototype models of remote sensing image feature type, which qualitatively reflects the statistical relationship between all types of data and prototype models. G is a symmetric matrix whose diagonal elements are 1. For any remote sensing image sample feature y , it is necessary to calculate the attraction between y and the prototype model of each remote sensing image type. The feature proximity vector g of the image sample is formed by these attractive indicators, which represents the fusion feature of the remote sensing image sample y . The proximity vector is composed of the attraction between the feature vector and the primitive graphic pattern, and the Euclidean distance of the row vector in g and G is compared. If the distance between g and G is close, the feature type of the image sample can be determined to complete the feature extraction of the remote sensing image[16].

4. Result analysis. In order to prove the usefulness of the method in this paper, it is necessary to carry out experimental proof. In the experiment, 300 remote sensing images are used as target samples for testing. Among them, 150 remote sensing images are all kinds of terrain with different angles, and 150 images are other targets. 90 representative remote sensing images of grassland, road, lake and 60 other target images were selected for training after feature acquisition[17].

4.1. Image feature vector classification detection. This paper lists 10 images that are sent into the classifier to implement classification accuracy after the method changes in this paper, as shown in Figure 4.1.

After using this method, the output of its soft threshold classifier is as follows: 0.9865, 0.9965, 0.7852, 0.9921, 0.9847, 0.6879, - 0.5898, - 0.5678, - 0.6987, - 0.4785. The results show that the algorithm in this paper has a good discrimination of the shape features of the terrain image, and can extract the features of the terrain image more accurately.

The remote sensing image in the randomly collected target sample passes through the feature acquisition channel, and is brought into the optimal classification function to check whether it is a terrain image. In the case of no noise, the error of the extraction results of the three types of image features is around 0.0500[18]. Considering that the noise will affect the detection results, 0 2 Gaussian white noise is added to the original image, and its shape, texture and fractal dimension consequences are shown in Figure 4.2 to Figure 4.4.

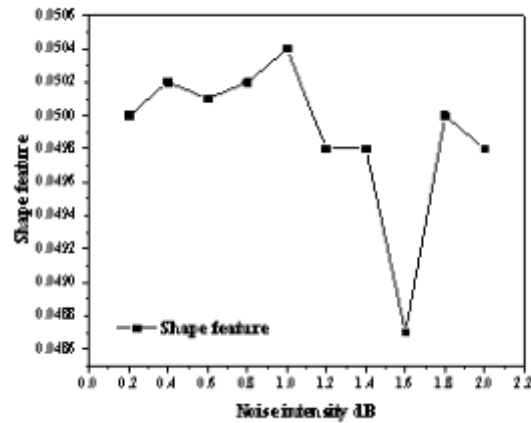


Fig. 4.2: Shape feature error after adding noise

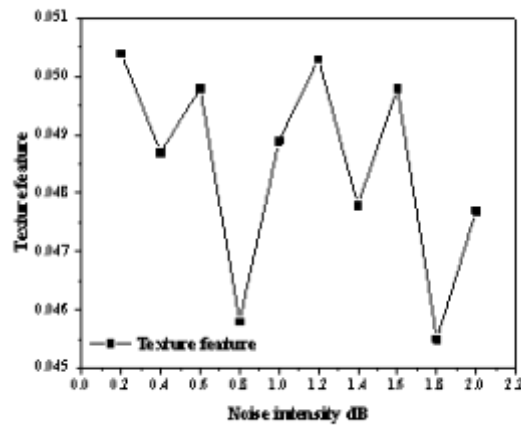


Fig. 4.3: Texture feature error after adding noise

It can be seen from Figure 4.2 to Figure 4.4 that the error range of the result after adding noise is basically the same as that of the noise-free case when the three types of feature vectors extracted by the algorithm in this paper are used for classification detection, with good noise resistance and robustness.

4.2. Comparison results with traditional methods. In order to verify the superiority of this method, we use this method and the traditional method to compare the images of the same sample set, and the results are shown in Figure 4.5 [19].

It can be seen from Figure 4.5 that the accuracy of feature extraction of the method in this paper is significantly higher than that of the traditional method when tested with the method in this paper and the traditional method respectively[20].

5. Conclusion. At present, in the post-processing of image capture, the restoration of image pixels is the key in the whole image processing process, and in the pixel restoration process, the key problem is feature extraction. In addition, the feature extraction of image pixels also has a significant impact on the subsequent

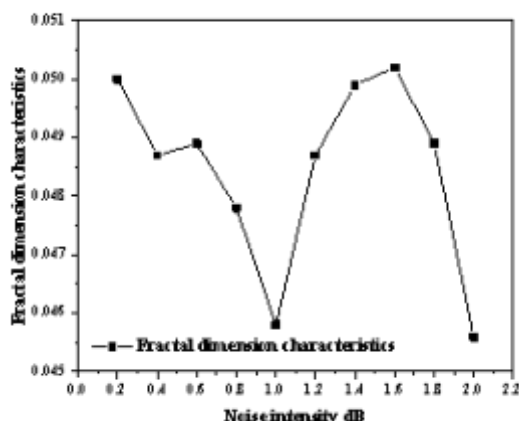


Fig. 4.4: Fractal dimension characteristic error after adding noise

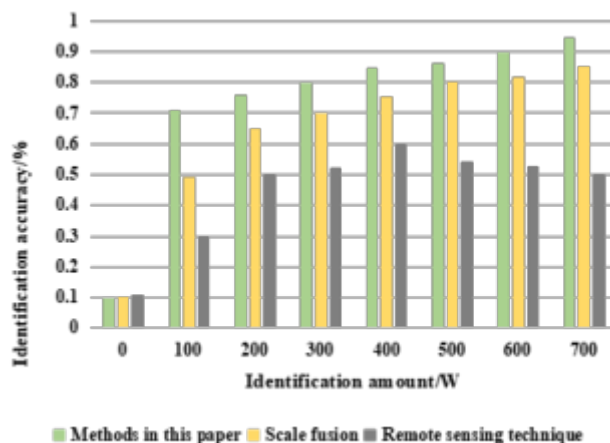


Fig. 4.5: Comparison of image feature extraction accuracy between traditional method and this method

work of image restoration and image calculation. This paper proposes a feature extraction method of remote sensing image based on SVM and multi-feature fusion. The simulation results show that the speed and accuracy of feature extraction of the proposed method are significantly improved compared with traditional methods, which verifies the feasibility of the proposed method. The method of extracting dimensions from remote sensing images of terrain surveying based on nonlinear multi feature fusion has many future research possibilities and areas for improvement. Here are some directions that can be further explored and improved:

1. Feature selection and extraction algorithm: more advanced Feature selection and extraction algorithm can be studied to better capture relevant information in remote sensing image of topographic mapping. For example, automatic feature learning techniques in deep learning models can be explored, as well as specific feature extraction methods that are more suitable for terrain surveying.
2. Data augmentation and incremental learning: Remote sensing images of terrain surveying may have different variation patterns and uncertainties. Therefore, data augmentation techniques can be studied to increase the diversity of training data through operations such as synthesis, rotation, and scaling,

thereby improving the robustness and generalization ability of the model. In addition, introducing incremental learning methods can achieve rapid adaptation to new data and model updates to cope with constantly changing terrain conditions.

REFERENCES

- [1] Hosgurmuth, S., Mallappa, V. V., Patil, N. B., & Petli, V. (2022). A face recognition system using convolutional feature extraction with linear collaborative discriminant regression classification. *International Journal of Electrical and Computer Engineering (IJECE)*, 12(2), 1468-1476.
- [2] Seçkin, A. Ç., & Seçkin, M. (2022). Detection of fabric defects with intertwined frame vector feature extraction. *Alexandria Engineering Journal*, 61(4), 2887-2898.
- [3] Vyas, R., Kanumuri, T., Sheoran, G., & Dubey, P. (2022). Accurate feature extraction for multimodal biometrics combining iris and palmprint. *Journal of Ambient Intelligence and Humanized Computing*, 13(12), 5581-5589.
- [4] Sun, L., Zhao, G., Zheng, Y., & Wu, Z. (2022). Spectralspatial feature tokenization transformer for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-14.
- [5] Ding, Y., Zhang, Z., Zhao, X., Hong, D., Cai, W., Yu, C., ... & Cai, W. (2022). Multi-feature fusion: graph neural network and CNN combining for hyperspectral image classification. *Neurocomputing*, 501, 246-257.
- [6] Li, X., Li, C., Rahaman, M. M., Sun, H., Li, X., Wu, J., ... & Grzegorzec, M. (2022). A comprehensive review of computer-aided whole-slide image analysis: from datasets to feature extraction, segmentation, classification and detection approaches. *Artificial Intelligence Review*, 55(6), 4809-4878.
- [7] Shankar, K., Perumal, E., Tiwari, P., Shorfuzzaman, M., & Gupta, D. (2022). Deep learning and evolutionary intelligence with fusion-based feature extraction for detection of COVID-19 from chest X-ray images. *Multimedia Systems*, 28(4), 1175-1187.
- [8] Wang, Z., Xie, Q., Wei, M., Long, K., & Wang, J. (2022). Multi-feature fusion votenet for 3d object detection. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(1), 1-17.
- [9] Zhao, L., & Zhu, Q. (2022). Image denoising algorithm of social network based on multifeature fusion. *Journal of Intelligent Systems*, 31(1), 310-320.
- [10] Zhang, Z., & Wang, M. (2022). Multi-feature fusion partitioned local binary pattern method for finger vein recognition. *Signal, Image and Video Processing*, 16(4), 1091-1099.
- [11] Guo, L. (2022). SAR image classification based on multifeature fusion decision convolutional neural network. *IET Image Processing*, 16(1), 1-10.
- [12] Huang, X., Liu, Y., Wang, Y., & Wang, X. (2022). Feature extraction of search product based on multi-feature fusion-oriented to Chinese online reviews. *Data Science and Management*, 5(2), 57-65.
- [13] Jin, Y. H., Oh, J., Choi, W., & Kim, M. K. (2022). Spatio-spectral decomposition of complex eigenmodes in subwavelength nanostructures through transmission matrix analysis. *Nanophotonics*, 11(9), 2149-2158.
- [14] An, S., Zhu, H., Wei, D., Tsintotas, K. A., & Gasteratos, A. (2022). Fast and incremental loop closure detection with deep features and proximity graphs. *Journal of Field Robotics*, 39(4), 473-493.
- [15] Zhou, J., Liu, L., Wei, W., & Fan, J. (2022). Network representation learning: from preprocessing, feature extraction to node embedding. *ACM Computing Surveys (CSUR)*, 55(2), 1-35.
- [16] Eltrass, A. S., Tayel, M. B., & Ammar, A. I. (2022). Automated ECG multi-class classification system based on combining deep learning features with HRV and ECG measures. *Neural Computing and Applications*, 34(11), 8755-8775.
- [17] Yang, H., Li, L. L., Li, G. H., & Guan, Q. R. (2022). A novel feature extraction method for ship-radiated noise. *Defence Technology*, 18(4), 604-617.
- [18] Zhou, Z., Dong, X., Li, Z., Yu, K., Ding, C., & Yang, Y. (2022). Spatio-temporal feature encoding for traffic accident detection in VANET environment. *IEEE Transactions on Intelligent Transportation Systems*, 23(10), 19772-19781.
- [19] Xiong, Z., Mo, F., Zhao, X., Xu, F., Zhang, X., & Wu, Y. (2022). Dynamic texture classification based on 3D ICA-learned filters and fisher vector encoding in big data environment. *Journal of Signal Processing Systems*, 94(11), 1129-1143.
- [20] Patro, K. K., Jaya Prakash, A., Jayamanmadha Rao, M., & Rajesh Kumar, P. (2022). An efficient optimized feature selection with machine learning approach for ECG biometric recognition. *IETE Journal of Research*, 68(4), 2743-2754.

Edited by: Bradha Madhavan

Special issue on: High-performance Computing Algorithms for Material Sciences

Received: Jan 17, 2024

Accepted: Mar 26, 2024



RESEARCH ON INTELLIGENT AGRICULTURE BASED ON ARTIFICIAL INTELLIGENCE AND EMBEDDED PERCEPTION ALGORITHMS

XINHUAN ZHAO*, FANG ZHANG† AND NA GAO‡

Abstract. In order to solve the problems of weak collection link, limited data coverage and poor real-time for big data in agriculture, smart agriculture by implementing artificial intelligence and embedded sensing is proposed. The front-end perceptron and wireless gateway were designed. A steady-state data collection system was constructed according to the characteristics of intelligent agricultural information data. Combining various algorithms such as data unification and data recognition, intelligent perception calculation parameters were extracted. The adaptive steady-state sensing model was designed relying on deep learning technology in the field of artificial intelligence. The experimental results show that the RMSE value of the designed system in the study is 0.028, which meets the requirements of intelligent agricultural information data perception accuracy. It is concluded that agricultural big data is a collection of data involved in the process of agricultural production, transportation and marketing, and data collection is the most important part of it.

Key words: artificial intelligence, embedded sensing, smart agriculture

1. Introduction and examples. The development of artificial intelligence in agriculture has changed people's thinking about agricultural management services. Based on the background of big data, the reasonable use of artificial intelligence can effectively monitor the production of agricultural products and build a management system that combines information monitoring and services. The system has the functions of early warning of natural disasters, prevention and control of pests and diseases, and prediction of market fluctuations, which effectively reduces the construction of agricultural platforms and creates new engines and dynamics of agricultural and rural modernization [1]. Intelligent agriculture refers to an agricultural model that utilizes modern technological means to improve agricultural production efficiency and agricultural product quality. Among them, the application of artificial intelligence and embedded perception technology can greatly improve the efficiency and intelligence level of agricultural production.

In order to accelerate the construction of digital agriculture, we should not only play the role of the government, but also mobilize the strength of all parties to form a joint effort to promote it. We should encourage and guide social capital to invest in agriculture and broaden the source of funds for agricultural and rural development. In accordance with the digital construction carried out by agricultural enterprises and new business entities, it is recommended that the government introduce relevant support policies to provide financial incentives to accelerate the process of digital platform construction. We should strengthen the construction of agricultural infrastructure, actively carry out the creation of modern agricultural parks, and fundamentally improve the conditions of agricultural practices. We should strengthen digital agriculture and rural business training, carry out digital agricultural talents to the countryside, popularize digital agriculture-related knowledge, digital technology application and management level of new business entities and high-quality farmers. We should play the role of scientific research institutions, universities, enterprises and other parties to speed up small farmers and digital agriculture interface, and accelerate the integration and application of agricultural data. At the same time, we should use digital technology to transform traditional industries in the countryside, attract outstanding urban talents to return to their hometowns for employment and entrepreneurship, inject new ideas, new thinking and new strategies into agricultural production, guarantee the scientific nature of agricultural business decisions, and further promote the high-quality development of digital agriculture [2,3].

*LangfangYanjing Vocational and Technical College, Hebei, Langfang, 065200, China (Corresponding author, XinhuanZhao7@163.com)

†LangfangYanjing Vocational and Technical College, Hebei, Langfang, 065200, China(FangZhang68@126.com)

‡LangfangYanjing Vocational and Technical College, Hebei, Langfang, 065200, China(NaGao35@163.com)

2. Literature Review. With the Internet of Things, big data, artificial intelligence and other technology applications continue to sink into the social life of production practices, digital agriculture has become the focus of the development of the agricultural industry now and in the future. Artificial intelligence can improve the accuracy and speed of agricultural production decision-making by analyzing a large amount of agricultural data. For example, using machine learning algorithms to analyze and predict agricultural data can predict crop growth periods, diseases and pests, and take timely measures to protect crops and improve crop yield and quality. The deepening of the application of digital agriculture in China's agricultural production prompts significant changes in China's traditional agricultural production methods. More and more crude, mechanical and empirical production modes are developing towards intensive, intelligent and scientific. Today, digital applications have become the mainstream of industrial production and social life, while the application of digital technology in agricultural production is still in its infancy. The agricultural sector may be the last industry where information technology and digitalization are popularized. The reasons for the slow diffusion of digital technology applications in agriculture are multi-layered, the main reason of which is that the scattered production and operation mode in rural areas of China is not conducive to the concentration of data resources in the informatization system [3]. And with the continuous improvement of rural communication infrastructure, the gradual maturity of agricultural IoT technology, and the national vigorous promotion of the construction of modern agricultural industrial parks, data integration in the whole agricultural industry chain has become possible. Liu, S. et al. constructed an agricultural Internet of Things (IoT) management system to realize the integrated management of Internet devices, environmental data, video data and agricultural expert knowledge. Then, they introduced the current status of agricultural IoT from the perspective of sensing technology, transmission technology and three intelligent information processing technologies, analyzed the economic benefits of IoT for agricultural production, and proposed future research priorities and development directions for agricultural Internet in China [4]. Vermesan, O. et al. introduced ECAS vehicles through artificial intelligence (AI) in vehicle and infrastructure-level architectures based on evolution of distributed intelligence based domain controllers, regional vehicles and federal vehicle/edge/cloud centers, and the role of AI technologies and approaches in achieving different autonomous driving and optimization functions for sustainable green transportation [5].

Agricultural big data is the overall collection of data involved in the process of agricultural production, transportation and sales, and data collection is prominent as its most important link. Embedded perception technology can achieve real-time monitoring and data collection of agricultural production environment, such as temperature, humidity, lighting and other environmental factors. At the same time, it can also achieve real-time monitoring of crop growth, such as crop growth status, pests and diseases. These data can be used to improve the accuracy of agricultural production, thereby reducing waste and improving agricultural efficiency. In the early days, the degree of agricultural informatization was low, agricultural data was small and the mining value was low. However, in recent years, the development of agricultural artificial intelligence and embedded sensing technology has made agricultural data show a spurt growth. Although the current development of agricultural big data is a big improvement over the previous, there are still some problems in the data collection link:

1. The data collection in agricultural production is uneven, only in the more developed areas of communication, and data transmission is mainly wired network;
2. The data collection is mainly based on sensor text information, with less image and video information;
3. The cost of the existing intelligent agricultural monitoring system is high and the system integration is low, it is not applicable to areas where broadband is not laid, and it is difficult and costly to deploy.

By summarizing the previous research experience, artificial intelligence technology is adopted in this study to establish an embedded sensing system with artificial intelligence adaptive sensing model as the focus to realize the sensing of agricultural information.

3. Method.

3.1. Front-end perceptron design for artificial intelligence-based embedded sensing system for agricultural information. In order to realize the collection and transmission of agricultural information, the front-end perceptron is designed. The agricultural information embedded sensing system of artificial intelligence is a system that applies modern sensing technology, communication technology, computer technology, and artificial intelligence technology to the agricultural production process. The front-end perceptron is an important component of the system, mainly responsible for collecting various parameter information in the

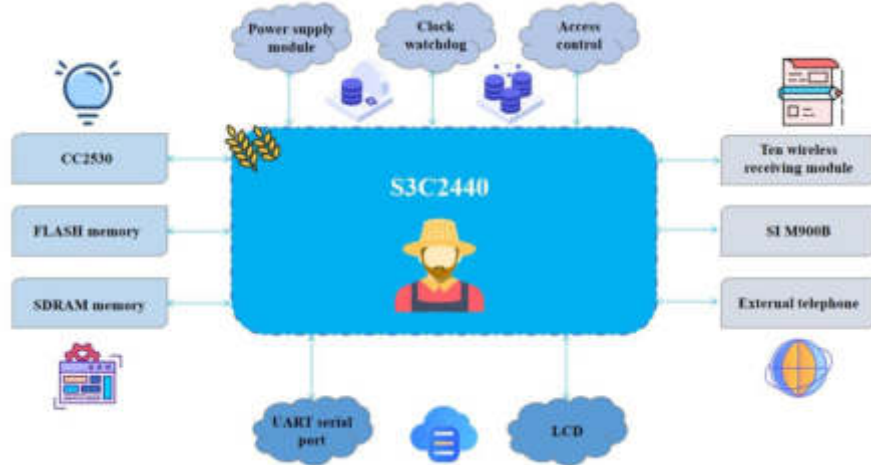


Fig. 3.1: Block diagram of the hardware structure of the gateway

agricultural production process and transmitting this information to the backend data processing center for processing and analysis. Considering the node power consumption and information reception sensitivity degree, the front-end sensing device designed in the study uses the CC2530 chip as the core, and connects the CC2530 chip to the microcontroller I/O port for the purpose of information exchange. Then it is combined with wireless transceiver module, clock module and other structures to complete the front-end sensing device design [6].

3.2. Wireless gateway design for artificial intelligence-based agricultural information embedded sensing system. The wireless gateway is designed mainly for data transmission, encapsulation and parsing. Through research, it is known that the S3C2440 microprocessor can operate at a maximum frequency of 400MHz, which can meet the working requirements of the sensing system. In this study, it is used as the design core of the wireless gateway, and then it is connected with components such as TFT-LCD display and remote control keypad. The actual hardware structure of the wireless gateway is shown in Figure 3.1.

In addition to the hardware structure of the gateway shown in Figure 3.1, the LM25965-5.0 switching voltage regulator is installed at the gateway power supply in order to enhance the stability of the gateway application.

3.3. Building intelligent agricultural information data collection system. The sensing of smart agriculture information needs to be based on data. The data collection structure of smart agriculture information shown in Figure 3.2 is designed in the study. The construction of an intelligent agricultural information data collection system requires consideration of multiple aspects, including the selection and configuration of hardware equipment, the design and implementation of data collection methods, data storage and processing, data display and analysis, etc.

In the acquisition structure shown in Figure 3.2, S denotes the data collector, C denotes the encoder, l_1 and l_2 denotes the channel length. In order to ensure the integrity of steady-state data acquisition, an in-depth analysis is conducted for the two phases of information transmission, and the acquisition information of a single data concentrator is clarified, and the acquisition information is calculated by the coding function to generate the input signal. A moment in the data acquisition structure shown in Figure 3.2 is selected, and the data acquisition channel is described by Equation 3.1.

$$Y_a = X_{ja} + Z_a, Z_a \in N \quad (3.1)$$

In Equation 3.1, X denotes the input signal, Y denotes the output signal, a denotes the acquisition moment,

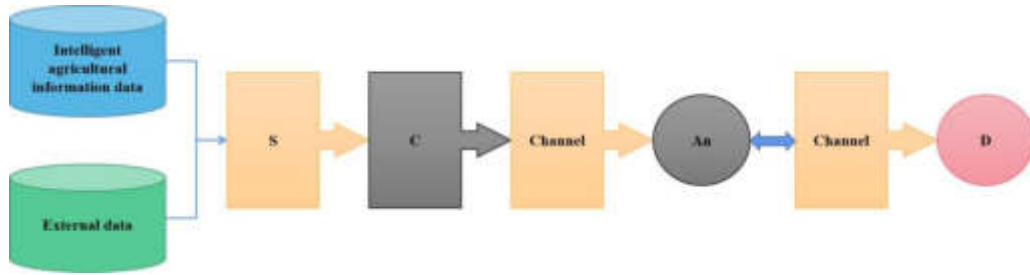


Fig. 3.2: Structure of intelligent agricultural information data collection

m denotes the number of data acquisition points of intelligent agricultural information, and denotes a data collection point, Z denotes interference noise, and N denotes variance [7,8].

Setting the critical capacity of the channel to transmit information in line with the minimum coded transmission requirements, the channel upper limit is calculated as Equation 3.2.

$$Q = L_1 \sum_j^m \log\left(1 + \frac{P_j}{N_j}\right) + L_2 \left[\log\left(1 + \frac{P_f}{N_f}\right) + \log\left(1 + \frac{P_z}{N_z}\right) \right] \tag{3.2}$$

In Equation 3.2, Q denotes the upper limit of the channel, $P_j P_f P_z$ denotes the average noise power of the information transmission process, $N_j N_f N_s$ denotes the noise variance of the information transmission process. For each average noise power analysis, the power constraint of each information transmission stage can be derived.

$$\begin{cases} P_j \geq \frac{1}{n} \sum_{a=1}^n [X_{j1}(w_j, a)]^2 \\ P_f \geq \frac{1}{n} \sum_i^n (X_{12}, \dots, X_{m2}, i)^2 \\ P_z \geq \frac{1}{n} \sum_i^n Z_i^2 \end{cases} \tag{3.3}$$

In Equation 3.3, ω indicates the agricultural information state quantity. Combining with the constraints shown in Equation 3.3, the spacing of the front-end data collectors in the steady-state data collection structure is set to realize the overall collection of steady-state data.

3.4. Extraction of intelligent perception calculation parameters. Based on the results of agricultural information data collection, techniques such as data unification and data recognition are applied to extract the mainstream features of steady-state data. Considering that the collected data are associated with both time and space, a data unification multi-layer model is designed in the study to identify the data states. Each feature quantity for the collected steady-state data is recorded to form the following matrix.

$$D = \begin{bmatrix} d_{11} & \dots & d_{1\alpha} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ d_{\beta 1} & \dots & d_{\beta z} \end{bmatrix} \tag{3.4}$$

In Equation 3.4, D denotes the acquisition matrix, d denotes the number of individual features of the acquisition data, $\alpha\beta$ denotes the number of columns and rows of the acquisition matrix.

$$\tilde{\omega}_\alpha = (v_1, v_2, \dots, v_\alpha) \tag{3.5}$$

In Equation 3.5, $\tilde{\omega}$ denotes a sequence of steady-state data vectors, v denotes the matrix column vector.

Considering that the frequency of steady-state data collection varies, some of the collected data have the problem of loss. In the study, a dynamic time programming method is used to calculate the similarity of discrete sequences, and complete the sequence expansion and compression to ensure the uniformity of the sequence scale. A column vector is randomly selected as the reference vector within Equation 3.5, and the Euclidean distances of other column vectors are calculated to generate multiple distance matrices.

$$O_k = \begin{bmatrix} B_{11} & \dots & B_{1\beta} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ B_{\beta 1} & \dots & B_{\beta\beta} \end{bmatrix} \tag{3.6}$$

In Equation 3.6, k denotes the column vector, O_k denotes the distance matrix, B denotes the Euclidean distance. The distance matrix is extrapolated to form several distance loss matrices to complete the calculation of the column vector similarity.

$$\theta = \begin{bmatrix} \varepsilon_{11} & \dots & \varepsilon_{1\beta} \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \varepsilon_{\beta 1} & \dots & \varepsilon_{\beta\beta} \end{bmatrix} \tag{3.7}$$

$$Q = \{Q_1, Q_2, \dots, Q_\beta\} \tag{3.8}$$

In Equation 3.7 and 3.8, θ denotes the distance loss matrix, ε denotes the degree of loss, and Q denotes the optimally adjusted sequence and also the set of shortest paths within the matrix. The distance between the steady-state data vectors is adjusted by dynamic regularization techniques to ensure the distance minimization.

Then, using principal component analysis, the validity of the steady-state data is evaluated, duplicate redundant information is removed, and the complexity is calculated. First, the normalization process is performed for the adjusted vector distances to obtain the normalization matrix shown below.

$$U = \xi - \frac{\xi}{\beta} \tag{3.9}$$

In Equation 3.9, U denotes the normalized matrix and ξ denotes the interval distance adjusted covariate data, and based on the calculation of Equation 3.9, the covariance matrix and singular value decomposition formulas are obtained as follows.

$$E = \frac{1}{\beta} U \tag{3.10}$$

$$svd(E)[H, R, F] \tag{3.11}$$

In Equation 3.10 and 3.11, E denotes the covariance matrix, svd denotes the singular value decomposition, H, R, F denotes the matrix formed after decomposition, H denotes the dimensionality reduction matrix, and the main data vectors are dimensionally reduced by using the dimensionality reduction matrix.

Relying on the above dimensionality reduction data, the intelligent perception parameters of a single sample are calculated in combination with support vector machines, and then the likelihood functions of the unknown parameters are calculated with reference to the independence of each observed object within the collection sample. In summary, the extraction of intelligent perception computational parameters is completed [9].

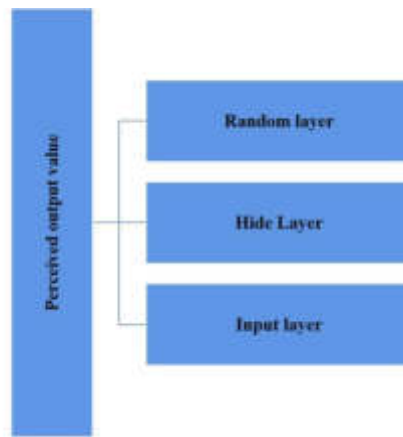


Fig. 3.3: Artificial intelligence-based adaptive perception model

3.5. Building an artificial intelligence adaptive perception model. The embedded sensing system designed in the study is centered on artificial intelligence technology, i.e., an artificial intelligence adaptive perception model is used as the focus of the research, and deep learning techniques within the field of artificial intelligence are applied to construct an adaptive perception model. The adaptive perception of the steady-state action behavior is divided into two parts, on the one hand, the input to output calculation of the AI neural network, and on the other hand, the parameter weights are modified based on the output perceived posture values. Deep learning technology is a machine learning method based on artificial neural networks. Its main feature is that it can automatically learn and extract features from data, and thus construct more accurate and efficient models. In embedded sensing systems, deep learning technology can be applied to the construction and optimization of perception models. By training and learning the data collected by sensors, more accurate and adaptive perception models can be constructed, improving the perception accuracy and stability of the system.

The adaptive perception model relying on artificial intelligence technology consists of four main layers of structure, as shown in Figure 3.3.

According to the schematic diagram of the perception model shown in Figure 3, it can be seen that the input layer includes the current steady-state condition of intelligent agricultural information, and according to the two parameters mentioned above, the input vector is described as:

$$\lambda(t) = (\lambda_1(t), \lambda_2(t), \dots, \lambda_2(t)) = \{\rho(t), \rho(t-1), \dots, \rho[t - (\partial - 1)\tau]\} \quad (3.12)$$

In Equation 3.12, $\lambda(t)$ denotes the input vector of the perceptual model in time input vector, ∂ denotes the attack time interval, ρ denotes the steady state condition of agricultural information data, τ denotes the time delay.

The information of the input layer of the adaptive perception model is passed to the hidden layer, which is computed via multiple hidden nodes to obtain.

$$\mu(t) = \frac{1}{1 + \psi^r} \quad (3.13)$$

In Equation 3.13, μ denotes the hidden layer output result, ψ denotes the constant, and r denotes the parameter weights.

The output results of the hidden layer are applied to the random layer to calculate the Gaussian distribution characteristics of the steady-state data as a way to describe the distribution of the output data. Considering the results of Gaussian distribution calculation for each hidden node, which is directly influenced by the intelligent perception parameters, the random layer output is expressed as:

$$\eta[\mu(t), r_0] = \frac{1}{1 + \psi^{\mu(t)r_0}} \quad (3.14)$$

In Equation 3.14, η denotes the random layer output result, r_0 denotes the hidden node parameter weights.

Finally, an adaptive reinforcement learning mechanism is added to the output layer to further analyze the output results of the stochastic layer, which is expressed as a one-dimensional Gaussian function. The steady-state perception results are obtained using intelligent perception calculation parameters, and then adaptive learning is performed for the deviations in the stochastic layer to update the parameter weights and obtain more accurate adaptive perception results for the steady-state operational behavior.

3.6. Developing the embedded sensing system. After the software design is completed, an embedded real-time operating system is used for software development. Its main feature is that it can respond to external events in real-time, while ensuring that the system can complete the processing of events within a specific time range. The application of embedded real-time operating system divides the software development into several subtasks and ensures that each subtask is responsible for the corresponding responsibilities and gives each subtask the corresponding operation order.

Considering that the perception system designed in the study is an embedded operating system, a comparative analysis of commonly used embedded real-time operating systems shows that the UCOS- system has free real-time characteristics and can support more than 250 tasks to be developed simultaneously. Therefore, UCOS- is chosen as the system software development platform in the study [10-11].

Usually, embedded real-time operating systems let the tasks in the front of the operation order run first during software development and can interrupt other tasks in the operation order for CPU preemption at any time. This development model optimizes the response time of software subtask development. This development model is applied to the development of the perception system, so as to make the functional software development into task-oriented software development and realize the simplification of the logical structure of the intelligent agricultural information data perception system. Finally, the software structure is set to three layers by using the embedded real-time operating system to avoid presenting the underlying hardware directly in the visualization interface, which facilitates the expansion of software and hardware respectively. So far, the overall design of the intelligent agricultural information data embedded sensing system is completed.

3.7. System test. In the study, the artificial intelligence technology is relied on to design an intelligent agricultural information data embedded sensing system. In order to verify the practical application effect of the system, a system test is conducted. During the test, the IEEE39 node system is used as an example to apply the system designed in the study to obtain steady-state sensing results and clarify the feasibility of the system designed in the study.

3.8. Building the test environment. Considering the embedded architecture of the design system in the study, the system testing process is based on Linux system, and multiple virtual machines are used to build the system testing environment to display the perception results in a visualized form in front of the user while the intelligent agricultural information data, and the system testing environment is shown as follows. Through the establishment and testing of the system test environment, the correctness and stability of the system can be verified, and the potential problems can be found and solved in time to ensure that the system can achieve the desired effect in practical application.

Combined with seven virtual machines, the test environment is completed by using Linux Ubuntu version of the operating system and JDK version programming components. Among them, four virtual machines act as Data Node slave nodes, two act as master nodes, and the remaining one is a management node. The actual configuration information is shown in Table 3.1.

According to the configuration information shown in Table 3.1, the IP address division of the virtual machine is realized, the programming components are installed on each virtual machine separately, and the environment variables are configured after the programming software is installed.

The configuration of environment variables starts from the settings of SSH protocol and Hadoop users. First, the SSH protocol is installed in each virtual machine, and a directory with the .SSH suffix is created to facilitate subsequent system startup and command execution. Then, the SSH protocol is used to generate keyless password pairs for Hadoop users and save them in the SSH directory. Finally, after the installation of Hadoop components is completed, the core component core-site.XML and MapReduce framework files are configured to complete the address configuration of slave and master nodes [12,13,14,15]. During this test, the

Table 3.1: Virtual machine address assignment

Nodes	IP Address
CDH Management Node	192.168.155.1
CDH Primary NameNode	192.168.155.2
CDH Secondary NameNode	192.168.155.3
CDH DataNode 1	192.168.155.4
CDH DataNode 2	192.168.155.5
CDH DataNode 3	192.168.155.6
CDH DataNode 4	192.168.155.7

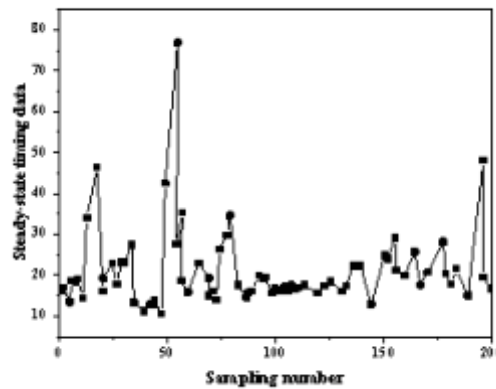


Fig. 3.4: Time-series data of agricultural information data

node system is the main result, which contains 10 generators, 46 lines and 19 load nodes in the system. In the above test environment, the sensing system proposed in the text is run to obtain intelligent agricultural information data.

3.9. Setting perceptual model parameters. In order to improve the accuracy of the test results, reasonable parameters are set for the artificial intelligence adaptive sensing model before the system is run. Running the IEEE39 node system is shown in Figure 3.4. The Nessus software is applied to scan the system acquisition characteristics, and the professional software is used to simulate network attacks during the scanning process to collect the 200 posture timing data shown in Figure 3.4.

As shown in Figure 3.4, agricultural information data can be regarded as nonlinear sequences, and agricultural information data situational awareness is accomplished by nonlinear mapping from different dimensional output spaces. Using the above 200 steady-state time-series data, 197 and 195 sets of test samples can be obtained when the dimensionality of the input vector is set to 3 and 5, respectively. The above data samples are applied to train the artificial intelligence adaptive perception model and compare the errors of the system output results under different parameters, so as to determine the final parameters of the model. It is known from the study that when the input vector dimension is set to 5, the number of nodes in the implicit layer of the model is 20, and the prediction results for the next time period of agricultural information data are more accurate [16,17,18].

4. Results and Discussion. After the parameters of the artificial intelligence adaptive sensing model are set, the embedded sensing system proposed in the study is applied to sense the steady-state operational behavior changes of the IEEE39 node system in one day, and the sensing results are combined with the actual detected state values to generate the line graph of the sensing results shown in Figure 4.1. The differences

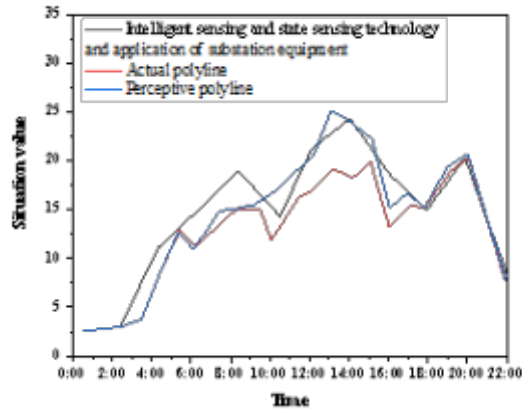


Fig. 4.1: Line graph of sensing results

between the sensed and actual data are analyzed to clarify the application performance of the designed system in the study.

According to the sensing results shown in Figure 4.1, it can be seen that, compared with the intelligent sensing and condition sensing technology and application of substation equipment, the stable posture values obtained by the sensing system designed in the study match the actual posture values in most cases, and the sensed posture is opposite to the actual posture only at ten and fifteen points. In order to describe the application effect of the sensing system more intuitively, the accuracy of the sensed posture value is calculated by using the RMSE value index in the study.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n |x_i - x_i|^2} = 0.028 \tag{4.1}$$

In Equation 4.1, RMSE denotes the mean square root error, n denotes the amount of steady-state action behavior data, i denotes a sample of steady-state data, x_i denotes the actual posture value, x_i denotes the perceived potential value. According to Equation 4.1, the RMSE value of the designed system in the study is, which satisfies the accuracy requirement of intelligent agricultural information data sensing [19,20].

5. Conclusion. In this study, it is proposed to realize intelligent agriculture by implementing artificial intelligence and embedded sensing. With the advancement of artificial intelligence technology, big data analysis can be more perfect. Secondly, artificial intelligence services provide a flexible infrastructure for agricultural big data analysis, which greatly simplifies the system scale of big data analysis and can be expanded according to demand, making it easy to manage workload. Finally, artificial intelligence services make users to perform big data processing without large-scale big data resources, which greatly reduces the big data system operation costs of agriculture-related enterprises and organizations and brings great value to agricultural development. In the era of big data, artificial intelligence is the strategic direction of future agricultural development, which can effectively improve agricultural production and quality, promote agriculture in the direction of green and ecological development, and then achieve the goal of smart agriculture.

The application of artificial intelligence and embedded sensing technology in intelligent agriculture can greatly improve the efficiency and quality of agricultural production, laying the foundation for the future development of intelligent agriculture. Firstly, through the application of artificial intelligence technology, intelligent agricultural management and decision-making can be achieved. For example, using artificial intelligence technology to analyze data such as farmland soil and crop growth status, predict crop growth trends and yields,

and provide more accurate and scientific decision-making basis for agricultural management. Secondly, the application of embedded sensing technology can achieve real-time monitoring and remote control of agricultural production. For example, using embedded sensors and controllers to monitor environmental factors such as weather, soil, and water quality, providing real-time feedback on data and controlling irrigation, fertilization, and other operations to improve crop production efficiency and quality.

REFERENCES

- [1] Breitzkreuz, C. , Herzig, L. , Buscot, F. , Reitz, T. , & Tarkka, M. . (2021). Interactions between soil properties, agricultural management and cultivar type drive structural and functional adaptations of the wheat Environmental microbiology, 23(10), 5866-5882.
- [2] Zhoulin, W. U. , Wang, W. , Lili, J. I. , Hou, B. , Bai, T. , & Zhang, J. . (2022). Teaching reform and practice of animal products processing under the background of intelligent agriculture. Asian Agricultural Research, 14(11), 3.
- [3] Ze-mengFENG, Yun-huaZHANG, Yu-minHE, QuanWANG, Jiao-genZHOU, & LunYE, et al. (2021). Intelligent breeding: the research and application of pig behavior. Research of Agricultural Modernization, 42(01), 1-9.
- [4] Liu, S. , & Wu, Y. . (2021). Economic benefit evaluation and analysis based on intelligent agriculture internet of things. Journal of Mathematics, 2021(5), 1-12.
- [5] Vermesan, O. , John, R. , Pype, P. , Daalderop, G. , & Waldhr, S. . (2021). Automotive intelligence embedded in electric connected autonomous and shared vehicles technology for sustainable green mobility, 17(6), 4318-4321.
- [6] Martini, B. G. , Helfer, G. A. , Barbosa, J. , Modolo, R. , & Leithardt, V. . (2021). Indoorplant: a model for intelligent services in indoor agriculture based on context histories. sensors, 21(5), 1631.
- [7] Kashyap, P. K. , Kumar, S. , Jaiswal, A. , Prasad, M. , & Gandomi, A. H. . (2021). Towards precision agriculture: iot-enabled intelligent irrigation systems using deep learning neural network. IEEE Sensors Journal, PP(99), 1-1.
- [8] Mateusz Szczepanski, Pawlicki, M. , Kozik, R. , & Chora, M. . (2023). The application of deep learning imputation and other advanced methods for handling missing values in network intrusion detection. Vietnam Journal of Computer Science, 10(01), 1-23.
- [9] Secinaro, S. , Mas, F. D. , Massaro, M. , & Calandra, D. . (2022). Exploring agricultural entrepreneurship and new technologies: academic and practitioners' views. British Food Journal, 124(7), 2096-2113.
- [10] Martin, R. . (2022). Integrative and exclusionary roles of trust in timber value chain in the southern highlands of tanzania. forum for development studies, 49(1), 27-52.
- [11] Ziyu, H. U. . (2022). Integration and development of china's characteristic agricultural industry against the backdrop of rural revitalization strategy. Agricultural Research, 14(2), 23-25.
- [12] Coti-Zelati, P. E. , Teixeira, M. , Machado, M. M. , DLAD Araújo, & RMD Pereira.(2021). Perception of the sociology of absences in the agricultural machinery industry supply chain. revista de Economia e Sociologia Rural, 60(4), 1-19.
- [13] Lillo-Saavedra, M. , V Gavilán, A García-Pedrero, C Gonzalo-Martín, & Rivera, D. . (2021). Ex post analysis of water supply demand in an agricultural basin by multi-source data integration. remote Sensing, 13(11), 2022.
- [14] Lopez, I. D. , Figueroa, A. , & Corrales, J. C. . (2021). Multi-label data fusion to support agricultural vulnerability assessments. IEEE Access, PP(99), 1-1.
- [15] Alemu, B. A. , Habteyesus, D. G. , & Abate, K. A. . (2022). The implication of intra-rural migration on crop output commercialization in ethiopia. Migration and Development, 11(1), 126-141.
- [16] Song, D. . (2022). Current situations, evolution trend, and improvement measures of soil organic matter in cultivated land of northeast china. Research, 14(3), 4.
- [17] Y Commandré, Macombe, C. , & Mignon, S. . (2021). Implications for agricultural producers of using blockchain for food transparency, study of 4 food chains by cumulative approach. sustainability, 13(17), 9843.
- [18] Wang, Z. , Kong, Y. , & Li, W. . (2022). Review on the development of china's natural gas industry in the background of "carbon neutrality". Natural Gas Industry B, 9(2), 132-140.
- [19] Amarasinghe, I. A. , & Hadiwattege, C. . (2022). Enablers for facilitating life cycle assessment: key stakeholder perspectives of sri lankan construction industry. built Environment Project and Asset Management, 12(4), 590-612.
- [20] Morano, L. , Bagasbas, J. C. , Remiter, M. , Ragas, M. J. , Bron, R. M. , & Lima, M. , et al. (2021). Lowering the minimum age of criminal responsibility (macr) in the philippines: a study on its implications and public perception in naga city, camarines Journal of City and Development, 3(2), 69-78.

Edited by: Bradha Madhavan

Special issue on: High-performance Computing Algorithms for Material Sciences

Received: Jan 17, 2024

Accepted: Mar 26, 2024



THE APPLICATION OF INTELLIGENT WELDING ROBOTS AND VISUAL DETECTION ALGORITHMS IN BUILDING STEEL STRUCTURES

WEI ZHANG* AND JUNHUA LI†

Abstract. In order to understand the application of welding robots in building steel structures, the author proposes a research on the application of intelligent welding robots in building steel structures. The author first analyzed the characteristics of complex and diverse structural forms of building steel structure components, small batches, no repetitive components, diverse welding joint forms, low precision of components and assembly control, and proposed specific requirements for the application of intelligent welding robots, such as fast programming to meet diverse structural forms, rich and powerful welding process databases, and high adaptability to parts and assembly deviations. Secondly, it is described that existing welding robots have mature contact sensing and arc tracking functions, have offline programming software, and can achieve thick plate groove welding. They already have the technical foundation to achieve automatic welding in building steel structural components. Finally, combined with practical cases, taking a commercial building project as an example, the construction land area is 31689m², using Q460GJC steel, with a maximum plate thickness of 100mm, the steel structure includes extended arm honing frame, radial separation frame, ring separation frame, and other contents, and is connected to the core tube as a whole. Describe the application content and methods of robot technology in the manufacturing of building steel structures, including parameter selection, programming methods, welding processes, and more. In order to achieve the goal of automatic welding of building steel structure robots. According to the application results, compared with traditional manual welding, intelligent Robot welding welding has higher efficiency and better quality, which is worth popularizing and applying comprehensively. Comprehensive comparative analysis, compared with manual welding, intelligent Robot welding has the advantages of higher efficiency and more stable quality, and has the technical conditions for comprehensive promotion.

Key words: Intelligent welding, Robots, Building steel structure

1. Introduction. With the rapid development of the construction steel structure industry, there are more and more steel structures for large-span, venue, and super high-rise buildings, and the types of components are becoming increasingly complex. Their design and production accuracy requirements are high.

At present, the low efficiency and unstable quality of manual welding operations often become the biggest obstacles to improving production efficiency and product quality stability. The improvement of welding level, especially automatic welding level, in steel structure manufacturing enterprises is the key to achieving rapid development of steel structure technology, the actual production components are shown in Figure 1. Although the welding workload in the construction steel structure manufacturing industry is large, there are significant difficulties in achieving fully automated welding due to the current non-standard design of construction steel structures, multiple types of components, small batch production of single pieces, complex processes, and low cutting and assembly accuracy in the previous process; However, the continuous innovation of advanced welding technology in the steel structure manufacturing industry and the application of efficient and intelligent welding equipment are gradually improving the quality of steel structures. At present, the vast majority of steel structure enterprises are still in the wait-and-see stage in the application of welding robots, considering the high cost and immature application of welding robots. With the development of the welding robot industry, the application of robots in the welding of U ribs and plate units in bridge projects is relatively mature; And in non bridge projects, enterprises have already put welding robots and supporting tooling systems into actual component production lines; Under the booming trend of robotics, the emergence of small welding robots and emerging technologies is also driving the application of intelligent steel structure welding. Welding robots have been

*Department of Construction Engineering, Hebei Vocational University of Industry and Technology, Shijiazhuang, Hebei, 050091, China (WeiZhang7631@163.com)

†Basic course teaching Department, Hebei Vocational University of Industry and Technology, Shijiazhuang, Hebei, 050091, China (Corresponding author, JunhuaLi28@126.com)

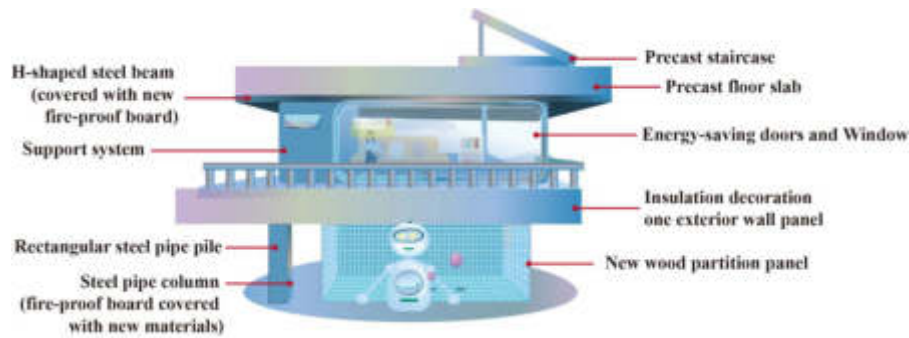


Fig. 1.1: Application of Intelligent Welding Robot in Building Steel Structures

widely used in industries such as automotive manufacturing and medical devices due to their high efficiency, performance, and quality. These products have the characteristics of standardization and large batches.

At present, the level of welding automation in industrialized countries around the world has reached as high as 80%, resulting in significant advantages in terms of efficiency and quality. According to the estimated consumption of welding materials for manual and automatic welding, the nominal degree of welding automation is 30%, which is a significant difference compared to this. With the development of building welding structures towards large-scale, heavy-duty, and high-precision parameters, the low efficiency and unstable quality of manual welding operations often become the biggest obstacles to improving production efficiency and product quality stability. In order to meet the special requirements of high-strength, thick plates, and long welds, the improvement of welding level, especially automatic welding level, is the key to achieving the rapid development of steel structure technology. Therefore, rapidly improving the level of welding automation has become an urgent and important task [1].

2. Literature Review. Intelligence and interconnection have become the mainstream direction for the future development of welding robots. The so-called intelligence mainly refers to the precise tracking and sensing of welding seams. In order to replace manual welding operations, robots need to accurately track welding based on the actual situation of the groove. Therefore, the mainstream development trend is to shift from a single teaching type to a multi-sensor and intelligent flexible processing system centered on intelligence. At present, most of the intelligent welding robots retained in the market are still teaching type, and the precise tracking and sensing technology for welds is not yet mature, and there is still great room for improvement.

Compared to the slow progress of intelligence, there has been significant progress in the interconnection of welding robots. The welding robot regional interconnection technology introduced by China Construction Steel Structure Co., Ltd. has been successfully applied to the on-site installation of steel components. Interconnection technology connects welding robots and terminal devices through regional networks, enabling remote information operation of welding. This not only improves welding efficiency, but also ensures the safety of operators.

At the same time, the real-time operation of the welding site can also be synchronized to the company's monitoring system through the network, further strengthening the control of welding quality. The intelligent and interconnected development of welding robots is the overall trend of their future development, and they will also make significant progress in other areas, such as automatic cleaning of welding slag, welding of complex components, and long installation time before welding, which require step-by-step technological breakthroughs, these are not bottlenecks that constrain the development of welding robots. Driven by market demand, the future development potential will be enormous. The position of the steel structure industry in the national economic development system is irreplaceable. With the increasing domestic steel production year by year, the application of welding technology in construction is also becoming increasingly popular, which has a huge impact on engineering safety and functional applications. Diaz-Cano, et al., propose an online robot programming approach that eliminates the traditional unnecessary steps in robot welding, allowing the operator to complete

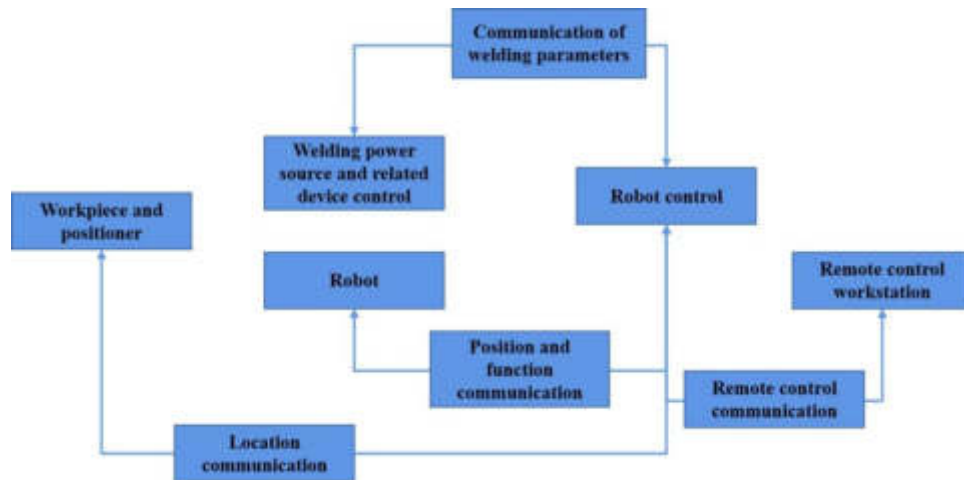


Fig. 3.1: Typical Welding Robot System Structure

the welding task by performing only three steps[2]. Canfield, S. L. et al., propose a collaborative robot model and model calibration strategy to aid teaching and monitoring of welding tasks. This method uses a torque estimation model based on robot momentum to create an observer to evaluate external forces [3,4].

At present, the structural form of steel structure buildings is complex. Introducing welding robots into the technology can lead the industry to gradually develop towards digitization and industrialization, fully meeting the requirements of technological innovation and environmental protection, and comprehensively improving the quality and efficiency of steel structure welding.

3. Research methods.

3.1. Technical methods for the application of intelligent welding robots . In the welding of structural components in industries such as bridges and construction machinery, there are problems such as large workpiece sizes and plate thicknesses, poor welding groove processing, and poor accuracy in workpiece assembly, in order to achieve good welding results, robots need to have sensing and tracking functions equivalent to human vision, touch, and other senses - that is, tracking and correction functions. The welding robot system can achieve functions such as finding the starting point of welding and tracking the weld seam through devices such as contact sensors, arc sensors, and laser tracking sensors (Figure 3.1).

3.2. Contact sensing function. The contact sensing function is a collection of starting point sensing, 3-directional sensing, welding length sensing, arc sensing, root gap sensing, multi-point sensing, etc. The robot senses voltage through the end of the welding wire (or welding gun nozzle), detects deviation and groove size of the welding workpiece, and remembers the position of the workpiece or weld seam.

Through the combined application of these functions, the welding process can be unaffected by errors caused by workpiece processing, assembly welding, and welding clamp positioning. It can automatically find the starting position of the weld seam and identify the weld seam situation, compensate for weld seam offset, deformation, length, and groove width changes, and ensure that the robot can weld smoothly. The principle is shown in Figure 3.2, and the contact sensing function is mainly used for locating the starting point of the weld and locating the groove [5,6].

(1) *Finding the starting point of the weld seam.* The starting point of the weld seam is located by sensing the surface of the workpiece in three directions, in order to perceive the actual position of the component weld seam to be welded [7,8]. The deviation between the actual position and the position of the component weld surface during teaching is calculated through a program, and then the deviation is added to the welding position during programming to find the correct welding position, correct the welding position deviation caused by assembly, assembly, and welding, and achieve the goal of ensuring welding quality.

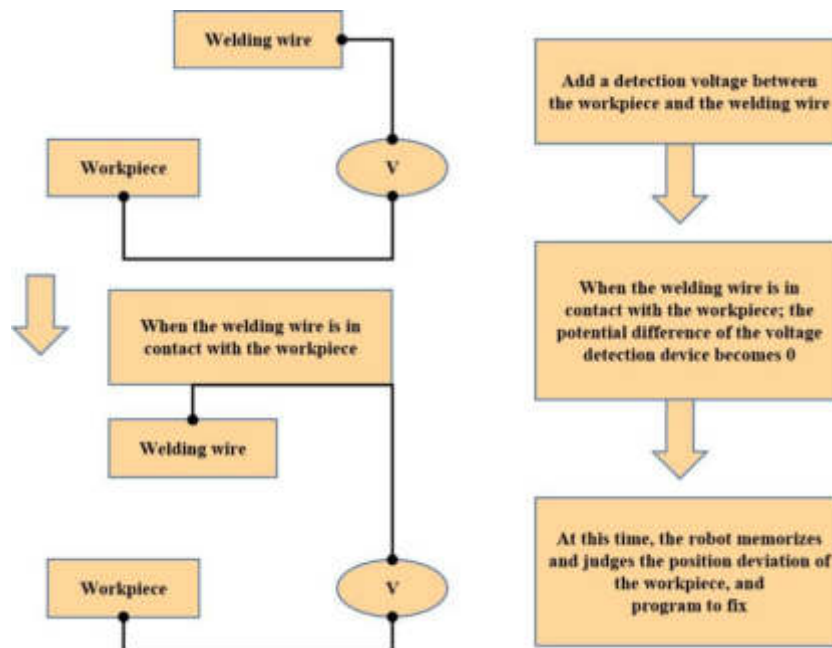


Fig. 3.2: Schematic diagram of contact sensing function

(2) *Groove sensing and groove positioning sensing.* Through the contact sensing of welding wire (or welding gun nozzle), the specific position of the weld groove can be quickly and conveniently found, and the width and depth of the groove can be automatically detected. At the same time, the groove angle can be calculated, which can be provided for the welding program to judge and adjust.

3.3. Arc tracking function . The arc tracking function is a function that searches for the center of the welding line, corrects the deviation of the welding workpiece in real-time, automatically detects the position of the welding line, and tracks the position deviation based on the feedback value of the welding current during swing welding. Especially in the multi-layer and multi-pass welding process, the workpiece change information obtained during the first layer of welding is utilized, and after the control system organizes and calculates, the results are directly applied to the welding after the second layer. Arc tracking is divided into welding line tracking (left and right direction tracking and up and down direction tracking) and groove width tracking [9].

(1) *Welding line tracking.* After the position of the starting point is determined, the correctness of the welding direction also needs to be ensured using the arc tracking function. Arc tracking refers to the real-time monitoring of welding voltage and current changes by the welding robot system through software during the welding process, analyzing and calculating the changes in arc length, and correcting the deviation of the weld seam through software adjustment of the robot's posture, thereby achieving seam position tracking.

(2) *Groove width tracking.* The arc welding robot system detects multiple points on the entire weld seam before welding, calculates the width of the weld seam groove through software, and then obtains the change in the width of the entire weld seam. During the welding process, by automatically adjusting the welding swing amplitude and welding speed, a weld seam with consistent height and shape is obtained, achieving the goal of improving welding quality [10,11].

3.4. Teaching Programming and Offline Programming. All welding robots have the function of teaching programming, which guides the welding gun to the starting point through the teaching box, and then determines the position, motion mode (linear or arc interpolation), swing mode, welding gun posture, and various welding parameters, at the same time, the movement speed of peripheral devices can also be determined through the teaching box. The welding process operations include arc striking, arc extinguishing, and filling

arc pits, which are also given through the teaching box. After the teaching is completed, the welding program can be produced. For the welding seams of structural components with complex structures, different shapes, and larger volumes, especially for the production of multi variety, small batch, and single variety, non batch welded structural components, online teaching will inevitably spend a lot of time, reduce equipment usage, and increase the labor intensity of operators. The method of online teaching directly restricts the application of welding robots, as is the case in the steel structure industry [12,13,14].

Offline programming technology utilizes 3D modeling software to copy the real work robot system into a computer, and then imports the component models generated by SolidWorks or ProE software into the robot scene in the computer. The workpiece can be programmed in an offline simulation environment, allowing the robot to complete welding programming independently on the computer without the need for the robot equipment itself. The program for on-site teaching and editing can be copied, translated, and other basic programs that can be read by offline programming software, greatly reducing the preparation time for welding programming and improving the utilization rate of intelligent robots [15].

4. Experimental analysis.

4.1. Project Introduction. Taking a certain commercial building project as an example, the construction land area is 31689m², using Q460GJC steel with a maximum plate thickness of 100mm. The steel structure consists of extended arm honing frame, radial separating frame, ring belt separating frame, etc., and is connected to the core tube as a whole. The steel used in this project is about 120000 tons, and the cotton frame layer structure is complex. In terms of design, high-strength bolts are used to connect with welding. The weight of a single component is 90 tons, and the welding quality and installation accuracy requirements are strict. The steel is mainly Q345GJC, which is a low alloy high-strength structural steel with a thickness range of 20-130mm. In actual construction, it is controlled according to foreign welding standards [16].

4.2. Application method.

4.2.1. Parameter determination. The robot in this building adopts a modular development route, with trajectories, actuators, multi degree of freedom welding guns, control platforms, and intelligent control modules, which can fully meet the on-site installation and welding needs of steel structures. In order to meet the various welding operation requirements on site, the main parameter selection is: in terms of technical parameters, the robot adapts to the welding position and supports horizontal, vertical, upward and 360° all position welding; Supports straight seams, circumferential seams, and irregular welds, supports circular workpiece sizes with a diameter exceeding 168mm, and the robot's walking speed is between 0-160cm per minute; The angle swing of the welding gun adopts a strip conveying method, with a swing speed of 0-255cm per minute and an amplitude of ±25mm; The horizontal tracking stroke is 200mm, the vertical tracking stroke is 150mm, and the programmable parameter adjustment amplitude is ±20%. In terms of melting efficiency, the efficiency of thick plate long weld welding is 1.5 times higher than that of arc welding; The magnetic adsorption type track of the robot is driven by friction, with a fine and compact body structure and convenient installation [17,18].

The comparative analysis of the technical performance of different mechanisms is shown in Table 4.1. Finally, a cage type positioner continuous flipping device was selected, and two workstations were arranged simultaneously. During welding operations, one workstation could be used for loading and unloading of vertical components, improving efficiency and the utilization rate of welding robots.

4.3. Programming method. This welding robot system supports three programming methods:

(1) *Online teaching programming.* Online teaching programming cannot be widely applied in building steel structures due to long equipment usage time and low efficiency, and can only be used as a supplementary application for on-site adjustment.

(2) *Offline teaching programming.* Through traditional offline teaching programming software, offline programming of all steel structure components can be achieved, meeting the requirements of welding usage. However, due to the numerous types of building steel structures, the standardization of welding structural components is not high, and the 3D models generated by the CAD software currently used by steel structure enterprises cannot be directly imported into offline programming software, and the 3D models of components need to be

Table 4.1: Comparative Analysis of Technical Performance of Different Mechanisms

Types	Work situation	advantage	disadvantage
L-type 90° flipping device	Build and purchase on the jig frame, and after welding on each side, flip the device with $N \times 90^\circ$ flip, experimental full position welding	Moderate price Can achieve continuous automatic welding	Unable to synchronize with robot control system
Continuous flipping of cage positioner	Built and loaded into a cage positioner, capable of achieving 360° continuous rotation	Can be linked with the robot control system to meet the requirements of continuous automatic welding	Relatively high price The system is relatively complex

rebuilt, it will consume a lot of time in modeling and editing robot action trajectories in offline programming software, so traditional offline programming software is difficult to meet the needs of actual production.

(3) *Intelligent rapid programming software.* In order to address the above issues, an intelligent and fast offline programming software has been developed. The rapid programming system adopts parameter driven, similar to building blocks, to construct a model of the welded part. The H-beam, box column, and cross column are defined as the main body of the workpiece, and the rib plate, bracket, and combined bracket are defined as modules, by inputting the size parameters of the main body of the workpiece, the size parameters and quantity of the module, and the parameters of the installation position on the main body of the workpiece, the two-dimensional model of the workpiece, the three-dimensional model that can be recognized by the offline teaching software, and the Robot welding implementation program are automatically generated. At the same time, according to the input size information, the welding database of the corresponding weld is automatically selected (Figure 4.1). The automatically generated robot program can be verified in an offline system.

The application of intelligent rapid programming software has greatly improved the welding programming time of building steel structure components, shortened the ratio of programming time to welding time, and laid an important technical foundation for the promotion and application of building steel structure welding robots. Online teaching programming requires a long investment time and low work efficiency, making it difficult to fully utilize in building steel structures and can only assist in on-site adjustments. In offline teaching, programming of all steel structure components can be achieved, fully meeting the requirements of welding applications. However, due to the variety of steel structure types, the standardization level of structural components is low, and the 3D model generated by the CAD software used by the enterprise cannot be directly imported into offline software. Therefore, it is necessary to rebuild the model and spend a lot of time editing the robot's motion trajectory, which is not in line with actual needs. In this regard, the building adopts intelligent offline programming software, which is parameter driven and similar to the principle of building blocks to construct a welded workpiece model. It uses box shaped columns and cross column workpiece bodies to support and strengthen modules such as plates, and automatically generates a two-dimensional model. By inputting parameters such as workpiece body parameters, size, quantity, etc., the offline teaching software of the three-dimensional model is recognized. Based on the input size information, a matching database is automatically selected, Enable the robot program to be verified in the offline system, saving more programming time, and effectively ensuring the welding efficiency of the steel structure of this project.

4.3.1. Welding process. This type of robot can walk along a predetermined route, repeatedly operate a certain process for a long time, and is easy to track and control. The system operation is stable and reliable, with high work efficiency. It is suitable for prefabrication and welding at various locations on site, and can effectively solve the problem of automatic welding for long welds and multi-position steel structure installation in construction projects. In this project, the application of welding robots includes the following content. In the welding of the extended arm truss, the truss layer of this project is a key and difficult welding point, mainly

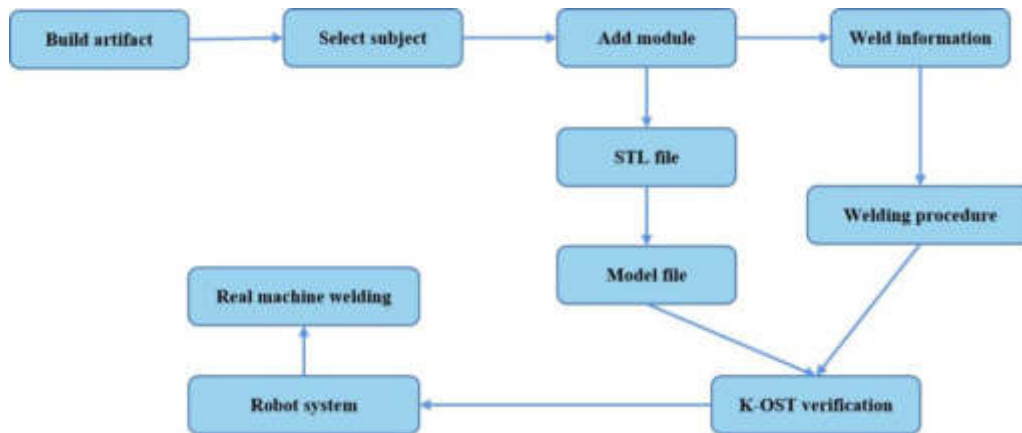


Fig. 4.1: Intelligent Fast Programming Process

Table 4.2: Welding flux parameters of damper quality box

project	numerical value
Number of steel cables	13 pieces
steel type	Q355B
Quality box quality	1100t
Steel plate thickness	85mm
Weld length	10m

made of Q390GJC material. The maximum length of the vertical weld seam of the extended arm truss is 4m, and the plate thickness is 140mm, one weld seam requires two welders to work continuously for 40 hours to complete. According to the relevant regulations for steel structure welding, the difficulty level reaches level D. This project includes a total of eight truss layers, each with a degree of 2-4m, a plate thickness of 80-140mm, and more than 50 welds. Due to the large amount of welding, low manual operation efficiency, and unstable welding quality, robot technology should be introduced to achieve high-altitude welding goals. The welding operator determines the length of the track based on the actual situation on site, configures the power control box, wire feeder, etc. required for automatic welding, and connects the cable to the welding trolley. The cable length is about 25 meters. This equipment can be placed at a high altitude around the car for welding. The welding protection gas cylinder can be connected to the control box through a gas pipe, ensuring that the center of the weld pool is the same as the weld seam during welding, optimizing and improving the welding parameters, and achieving the goal of continuous welding. In the welding of the damper quality box, the equipment is located on the 125th floor, with a height of 27m. There are 4 groups of 13 pieces suspended on the 125-131 floor, and the specific parameters are shown in Table 4.2 [19].

In order to ensure the reliability of root welding, a 5mm gap should be reserved during assembly. Generally, welding wires with a diameter of 5mm should be placed on both sides and in the middle during assembly, and then positioned for welding, the length of the weld seam is between 30-40mm, with a spacing of 400-500mm, it should be positioned at the small groove position, and the end should be spot welded and fixed before polishing smoothly, treat it as the arc starting point for robot operation. After the point welding is completed, use a flame gun to heat and remove the welding wire. Arrange the two robots symmetrically to minimize welding stress and deformation while ensuring welding progress and quality.

4.4. Application effect. In this project, the welding robot was used to achieve the welding objectives of the boom honing frame and the damper mass box. Under the same position and conditions, it has significant advantages compared to traditional welding modes, and the application effect is mainly reflected in: Firstly, the

welding quality is high, the appearance is beautiful, the weld seam is smooth with the base material, and the non-destructive testing results meet the standards[20]; Secondly, the welding efficiency is high, and automatic slag cleaning can be achieved during the welding process, achieving continuous operation, which doubles the efficiency compared to previous manual welding; The third is to greatly reduce the intensity of manual work. Technicians only need to adjust the welding parameters. After completing the weld seam teaching activity, the robot can automatically repeat welding, which is very convenient.

5. Conclusion. Currently, there are various forms of construction, and the types of steel structural components are becoming increasingly complex, with the characteristics of small batches without repetition, which puts forward higher requirements for the function of welding robots. In practical projects, appropriate parameters and programming methods should be selected based on the actual situation, advanced and reliable welding processes should be applied to achieve the goal of automatic welding of steel structure components, compensate for the quality defects of previous manual welding, and better meet the requirements of digitalization and technology in steel structure manufacturing, making component manufacturing more efficient and reliable. The application of intelligent and rapid programming software greatly improves the welding programming time of building steel structure components, reduces the ratio of programming time and welding time, and lays an important technical foundation for the promotion and application of building steel structure welding robots.

REFERENCES

- [1] Wang, G., & Arora, H. (2021). Research on continuous trajectory planning of industrial welding robot based on cad technology. *Computer-Aided Design and Applications*, 19(S2), 74-87.
- [2] Diaz-Cano, I., Quintana, F. M., Lopez-Fuster, M., Badesa, F. J., Galindo, P. L., & Morgado-Estevéz, A. (2022). Online programming system for robotic fillet welding in industry 4.0. *Industrial Robot*14(3), 49.
- [3] Canfield, S. L., Owens, J. S., & Zuccaro, S. G. (2021). Zero moment control for lead-through teach programming and process monitoring of a collaborative welding robot. *Journal of Mechanisms and Robotics: Transactions of the ASME*24(3), 13.
- [4] Chen, Y., & Hu, Q. (2022). Dual-robot stud welding system for membrane wall. *Industrial Robot*36(1), 49.
- [5] Chen, H., Ma, H., Jiang, H., Lv, S., Li, Y., & Liu, H. (2022). The influence of control parameters on precision of welding seam tracking in manually control master-slave robot remote welding system. *Journal of Physics: Conference Series*, 2218(1), 012045-.
- [6] Liu, C., Wang, H., Huang, Y., Rong, Y., Meng, J., & Li, G., et al. (2022). Welding seam recognition and tracking for a novel mobile welding robot based on multi-layer sensing strategy. *Measurement Science and Technology*, 33(5), 055109 (13pp).
- [7] Rippegather, D. (2021). Valk welding takes production of robot torches, cable assemblies and shock sensors into their own hands. *Welding and Cutting*36(2), 20.
- [8] Ebel, L. C., Maa, J., Zuther, P., & Sheikhi, S. (2021). Trajectory extrapolation for manual robot remote welding. *Robotics*, 10(2), 77.
- [9] Chen, S., Liu, J., Chen, B., & Suo, X. (2022). Universal fillet weld joint recognition and positioning for robot welding using structured light. *Robotics and Computer-Integrated Manufacturing*, 74(2), 102279.
- [10] Fengqun, J. Z. (2021). Analysis of typical working conditions and experimental research of friction stir welding robot for aerospace applications. *Proceedings of the Institution of Mechanical Engineers, Part C. Journal of mechanical engineering science*, 235(6)2.
- [11] Haitao, L., Tingke, W., Jia, F., & Fengqun, Z. (2021). Analysis of typical working conditions and experimental research of friction stir welding robot for aerospace applications. *Proceedings of the Institution of Mechanical Engineers, Part C. Journal of mechanical engineering science*42(6), 235.
- [12] Luo, H., Zhao, F., Guo, S., Yu, C., & Wu, T. (2021). Mechanical performance research of friction stir welding robot for aerospace applications. *International Journal of Advanced Robotic Systems*, 18(1), 172988142199654.
- [13] Sharma, A., Singh, P. K., & Sharma, R. (2021). Numerical simulation of temperature distribution in robotic arc welding by aristo tm robot. *IOP Conference Series Materials Science and Engineering*, 1116(1), 012117.
- [14] Liu, M., Zhu, S., Huang, Y., Lin, Z., & Ge, D. (2021). A self-healing composite actuator for multifunctional soft robot via photo-welding. *Composites Part B Engineering*, 214(1), 108748.
- [15] Zhou, B., Zhou, R., Gan, Y., Fang, F., & Mao, Y. (2022). Multi-robot multi-station cooperative spot welding task allocation based on stepwise optimization: an industrial case study. *Robotics and Computer-Integrated Manufacturing*, 73, 1(0)2197.
- [16] Wang, Z. (2021). Design of laser welding workstation control system based on industrial robot. *IOP Conference Series Earth and Environmental Science*, 714(3), 032082.
- [17] Zhao, X., Wu, C., & Liu, D. (2021). Comparative analysis of the life-cycle cost of robot substitution: a case of automobile welding production in china. *Symmetry*, 13(2), 226.
- [18] Yaseen, S., & Prakash, J. (2021). Analysis of numerical method on inverse kinematics of robotic arm welding with artificial intelligence. *Journal of Physics: Conference Series*, 1964(6), 062104 (16pp).
- [19] Loukas, C., Williams, V., Jones, R., Vasilev, M., & Gachagan, A. (2021). A cost-function driven adaptive welding framework for multi-pass robotic welding. *Journal of Manufacturing Processes*, 67, 5(4)5-561.

- [20] Parameshwaran, R., Maheswari, C., Nithyavathy, N., Govind, R. R., & Vasanth, M. (2021). Labview based simulation on welding seam tracking using edge detection technique. IOP Conference Series Materials Science and Engineering, 1055(1), 012026.

Edited by: Bradha Madhavan

Special issue on: High-performance Computing Algorithms for Material Sciences

Received: Jan 17, 2024

Accepted: Mar 5, 2024



APPLICATION OF PHYSICAL MODELING AND VIRTUAL SIMULATION TECHNOLOGY IN MEASURING THE PERFORMANCE OF SUBWAY TRAIN TRACKING AND OPERATION

XIUXUAN WANG* AND HONGWEI LIANG†

Abstract. The purpose of this paper is to study the application of virtual simulation technology in the measurement of subway train tracking performance. Firstly, the safety braking model required by virtual reconnection technology is discussed around the operation simulation and performance measurement of virtual reconnection. The two trains are combined into "one virtual reconnection train set", the tracking train and the head train run with relative moving block, and the two run with the front train set with moving block. Then, according to the characteristics of Metro train tracking operation, a virtual reconnection model and an improved station tracking model are proposed for the bottleneck area of the station. Finally, the proposed model is verified by numerical calculation and computer simulation modeling. The simulation results reproduce the dynamic characteristics of train flow during metro train operation. The results show that the departure interval and minimum tracking interval of the train are all 58 s. The improved station tracking model has the moderate passing capacity as the relative moving block, and the virtual reconnection model has the largest passing capacity. After the initial delay of the system, the delay recovery ability of the virtual reconnection model is the strongest.

Key words: virtual reconnection, relative moving block, train tracking interval, performance measurement, simulation system, Subway operation system

1. Introduction and examples. Virtual simulation technology is Simulation technology and virtual reality technology combined product. It builds the whole system A complete virtual environment is typically characterized and integrated and controlled through the virtual environment Make a large number of entities. Entities interact in a virtual environment, or with a virtual Environmental effects are the real characteristics of the objective world.Virtual simulation means "true Real people manipulate virtual systems in a virtual environment while conducting simulations, in the most In recent years, the development has been very rapid, and a series of successful in a wider range of fields Apply.

Driven by information technology, simulation technology has developed into a universal and strategic technology for human beings to understand and transform the objective world, which requires it to further absorb and integrate other related technologies on the original basis. Virtual simulation technology is the combination of simulation technology and virtual reality technology. It is typically characterized by a unified and complete virtual environment of the whole system, and integrates and controls a large number of entities through the virtual environment [1]. The interaction between entities in the virtual environment or with the virtual environment is the real characteristics of the objective world. Virtual simulation refers to the simulation conducted by "real people manipulating virtual systems in virtual environment". It has developed very rapidly in recent years and has been successfully applied in a wide range of fields. Metro train operation system can ensure safe and smooth train operation. But it is a systematic process from design, construction to system commissioning, involving large quantities, high investment and complex system. So it is difficult to quickly find and handle problems only by virtue of experience. We combine virtual simulation technology with Metro train operation system to simulate a real subway operation environment on computer, which can be used for train operation control strategy, system integration scheme analysis, key subsystem testing and driving training. It has the characteristics of controllability, safety, repeatability and economy, which is of great significance for urban traffic development [2].

*Zhengzhou Railway Vocational & Technical College, Zhengzhou, Henan, 451460, China (Corresponding author, XiuxuanWang@163.com)

†Zhengzhou Railway Vocational & Technical College, Zhengzhou, Henan, 451460, China(HongweiLiang7@163.com)

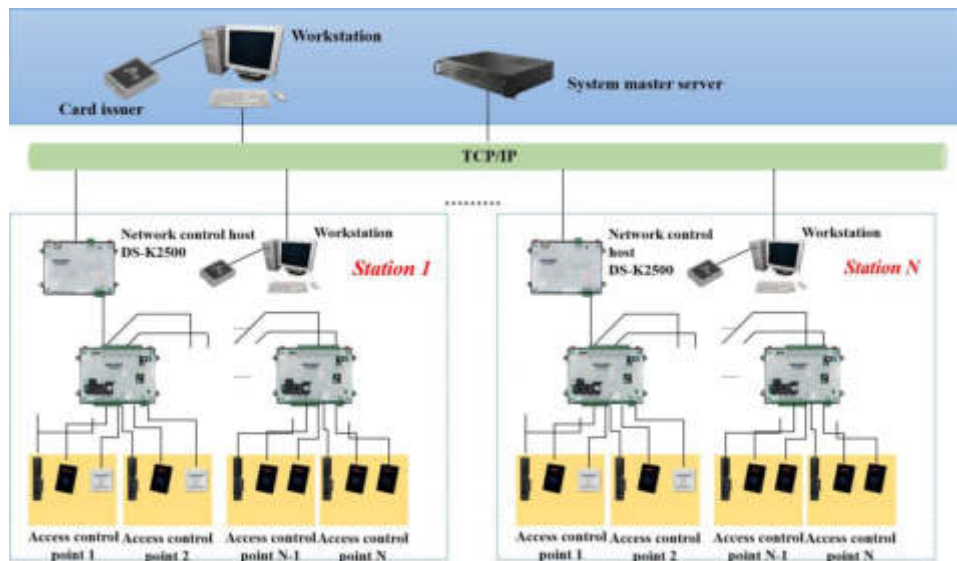


Fig. 1.1: Metro train tracking operation

The subway train operation system can ensure the safe and smooth operation of the train, but it is a systematic process from design, construction to system trial operation, which involves a large amount of engineering, high investment and complex system, and it is difficult to quickly discover and master the problem just by relying on experience. We combine virtual simulation technology with subway train operation system. To carry out research, simulate a real subway operating environment on the computer, can it be used for train operation control strategy, system integration scheme analysis, key subsystem test and driving training. It is controllable, safe, repeatable and economical. Sex and other characteristics are of great significance to the development of urban traffic.

In terms of the current development of the railway, on the main trunk lines of the railway line, with the increase of speed and traffic density, there will be mixed trains of different speed grades on the line at the same time, and their body weight and speed are different to some extent. Due to the different speeds of trains, it is impossible for trains to operate in an undisturbed environment. In most cases, multi-trains tracking operation with mutual influence (Fig. 1.1).

In this process, the operation of the two trains will affect each other. In case of failure or speed restriction of the current train during operation, the recommended speed curve and train schedule of ATO system are no longer applicable to the following train [3]. At this time, if the following train still operates according to the speed curve generated offline before leaving the station, it may produce unnecessary braking or parking, or even more serious accidents when receiving the fault information of the preceding train. After braking or parking, increasing speed or starting again will produce traction energy consumption, which increases the total energy consumption of the train.

According to the characteristics of section tracking operation, the following train obtains the position, speed and other information of the front train in real time through communication. Therefore, the impact of the front on the following should be taken into account in the energy-saving and optimized driving of the following train. So that it can update its target speed curve in time according to the actual situation, avoiding unnecessary deceleration or waiting. So as to reduce the total energy consumption of the train [4]. When the current train is disturbed during operation and has an impact on the following train, the ATO system of the following train will adjust or re-plan its operation speed curve in real time according to its current information and the information of the front train. It will adopt corresponding intelligent control methods to make the train take the updated speed as the operation basis, reducing delay and unnecessary energy consumption.

2. Literature review. The VR virtual simulation operation system of rail transit trains can help train operators to deal with various emergency situations, such as train accidents, equipment failures and so on. By simulating the actual situation, the operator can be trained in the virtual environment, familiar with the emergency handling procedures and operating procedures, improve the emergency response capacity and processing capacity, and greatly improve the operation safety. The VR virtual simulation operation system of rail transit trains can monitor and analyze the operation of trains in real time, identify potential faults and problems in advance by predicting and analyzing the operation status of trains, help dispatchers make reasonable operation plans and maintenance plans, and improve the reliability and operation efficiency of trains. The traditional training method needs to spend a lot of manpower, material resources and financial resources, and the VR virtual simulation operation system of rail transit trains can be trained in the virtual environment, which greatly reduces the training cost, and can also avoid accidents and losses caused by improper operation.

For this study, considering the constraints such as the actual line condition and speed limit of train operation, Z L ü et al. constructed the train operation strategy using genetic algorithm (GA), and gave the form of the optimal solution in ATO system and the minimized energy consumption of train driving control [5]. Liu, Y. et al. use the Lagrange multiplier and maximum principle method to obtain a set of optimal control and control conversion points of the train, and optimize the running time [6]. Song, H. et al. proposed a global optimal driving strategy for the train running on the track with different slopes, and calculated the local optimal switching point on each slope using the principle of local energy minimization. The algorithm can continuously update the optimal speed profile [7]. Considering the utilization of train regenerative braking energy, Zhang, M. et al. proposed a semi-analytical solution to discretize optimization problem of the train energy consumption, and applied the Lagrange multiplier algorithm to solve the speed optimization [8]. Dong, J. et al. proposed a distance based train speed trajectory search method, which uses the speed level of each preset position obtained during operation, and uses the search method combined with ant colony algorithm, GA and dynamic programming to search the train speed [9].

Based on the analysis of train energy-saving operation strategy, Zhanjun, W. proposed a fusion algorithm to optimize the global optimal operation strategy of train. The fusion algorithm is based on energy saving, meets the requirements of passengers, and can obtain the optimal timetable and optimal driving strategy of the train at the same time [10]. For multi train joint control, Li, Z. et al. proposed the corresponding controller and stability criteria. And the stability conditions of the proposed algorithm are given using Lyapunov stability criterion. The controller can use the information of the nearest train to ensure the operation performance of multi-train sequence [11]. Gao, R. et al. mainly studied the absorption of energy consumption generated by train regenerative braking and proposed the method of multi-train combination. That is, the energy consumption generated by the braking of the front train can be absorbed by other trains on the line [12]. Zheng, P. and others mainly used GA and particle swarm optimization algorithm to determine the train speed and minimize the error between it and the actual running speed. At the same time, Kalman filter is used to determine the position of transponder according to the changed parameters. When the transponder position reaches the optimum, the train speed error is also reduced [13]. Aiming at the train optimal control, Zhou, H. et al. firstly analyzed the stress of the train and established the train analysis model. Then, according to the selection principle of optimal coasting point and the energy-saving of regenerative braking, the single train interval operation optimization model and multi-train energy-saving operation model are established respectively. The particle swarm optimization algorithm based on Gaussian white noise disturbance variation is used to solve the above model, and the optimal control strategy of the train in each case is obtained. Finally, the energy-saving adjustment scheme under the condition of train delay is discussed [14].

3. Virtual simulation of train operation system. Scene model is the basis of simulation. Modeling is an essential part in the construction of virtual simulation system which needs to model the physical attributes and motion laws of virtual environment and virtual objects like visual appearance and surface friction. Its purpose is to give drivers and customers a sense of immersion in the actual scene around the current train.

(1) *Scene modeling content.* The modeling quality of scene model has a direct impact on the analysis and evaluation of the whole simulation system. After a high-quality model runs, users will have a strong sense of immersion [15]. Therefore, the three-dimensional model of subway track and trackside ancillary facilities must

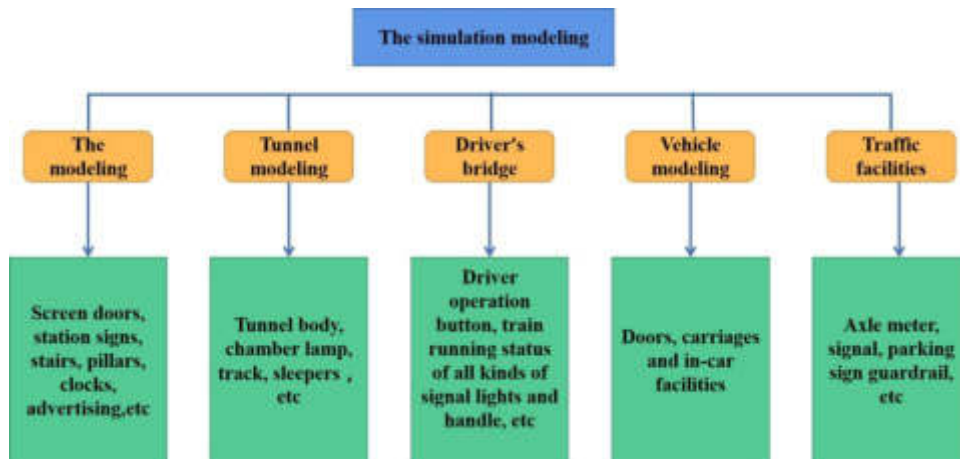


Fig. 3.1: Scene Modeling content

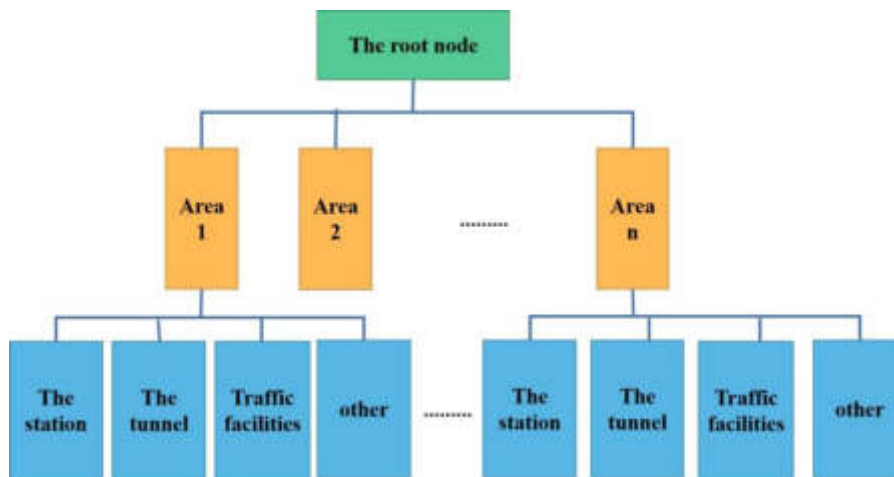


Fig. 3.2: organizational structure of scene model

be established according to the reconstructed line drawings of Beijing Metro Line 2. The design results can be observed and displayed from any angle, so as to realize the roaming function by interactively changing the position of viewpoint or preset motion route. We use the modeling software creator launched by American MultiGen paradise company for modeling. The virtual simulation modeling of the system mainly includes the following six modules, as shown in Figure 3.1.

(2) *Organizational structure of the model.* The organizational structure of scene model directly affects the operation efficiency of the system. The scene model is generally expressed in a multi-level tree structure. According to the characteristics of banded distribution of subway lines [16], the tree structure based on spatial location relationship is adopted to realize the establishment of scene model. The basic idea is to divide the whole scene into several regions along the line direction, which correspond to the first-level nodes in the scene model tree (Figure 3.2).

The idea of hierarchical modeling method is generally adopted for the modeling of the underlying facilities. It uses the tree structure to represent each component of the object. It not only provides a convenient natural segmentation method, which can arrange various levels of the database from large to small, but also is very

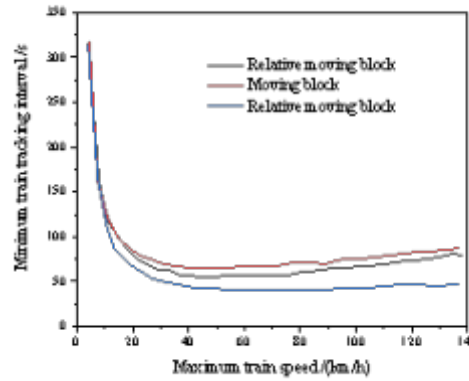


Fig. 4.1: Influence of station tracking interval on maximum operation speed under different tracking modes

beneficial to the modification of complex models [17]. Take the subway station model as an example: the independent modules such as platform, wall, roof, column and annunciator can be placed at the same level, while the stairs can be placed at a lower level than the platform because they are on the platform. The stair nodes include handrails, steps, etc., which are the next node of the stairs. This tree hierarchical scene modeling based on spatial location relationship can quickly realize various convenient operations on the scene model of the specified area from top to bottom, reduce mutual restraint and interference, have clear structure and clear hierarchy, and greatly improve the operation efficiency of the system.

The whole scene is divided into static scene and dynamic scene. The static scene is the fixed part of the scene, and the dynamic scene is the part that changes in the scene. In the simulation process, the dynamic scene should be controlled. The screen door, train door, signal and train operating status indicator in the station are all dynamic scenes, and we need to control them in real time according to the operation of the train, such as controlling the opening of the screen door and the train door, the color change of the signal and the indicator, and the movement of the virtual handle. The control of dynamic scenes can be achieved through the Switch node and DOF provided by Creator.

The system software mainly includes three parts: the model system initialization, the 3D model loading and management and the main program of vision generation. The initial configuration of the model was completed through the Vega graphical interface LynX, and the.adF file was formed. After the initialization work is completed, the virtual simulator loads the required 3D model from the file into the memory, and at the same time reads the signal device description file and determines the state of the device and the position of the moving object to operate the model accordingly.

4. Performance measurement and result analysis.

4.1. Steady state performance measurement. The simulation parameters are set as follows: train length $L_T = 140m$, protection section length, $L_S = 15m$, train starting acceleration $a = 1m/s$, train braking deceleration $b = 1m/s$, braking reaction time $T_R = 3s$ and stop time $T_R = 3s$. Based on formulas 4.1 to 4.4, the relationship between station tracking interval and train speed of moving block, relative moving block, virtual reconnection model and improved station tracking model can be obtained [18]. The simulation results are shown in Figure 4.1.

Under the condition of train tracking operation, the general formula for calculating the passing capacity of the line is:

$$T_1 = \frac{2a(L_T + L_S) + v_{max}^2}{2av_{max}} + T_R + T_D \quad (4.1)$$

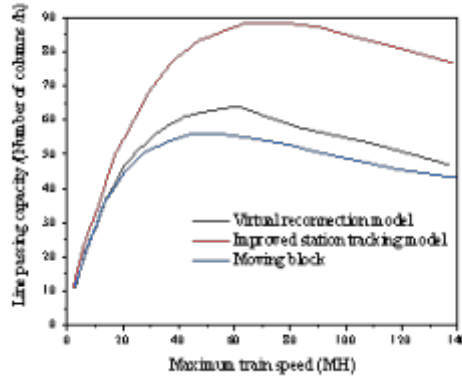


Fig. 4.2: Influence of passing capacity on maximum speed under different tracking modes

Table 4.1: Comparison of station tracking interval and passing capacity under different tracking modes

Block system	The station tracking interval/s	Passing capacity/ (cars/h)
Moving block	60.21	54
Relative moving block	54.35	58
Virtual reconnection model	33.26	86
Improved station tracking model	54.46	88

$$T_2 = \sqrt{\frac{2b(L_T + L_S)}{a^2 + ab}} + T_R + T_D + \frac{v_{max}}{b} \quad (4.2)$$

$$T_1 = \frac{4a(L_T + L_S) + v_{max}^2}{2av_{max}} + T_R + T_D + \frac{v_{max}}{b} / 2 \quad (4.3)$$

$$T_4 = T_2 \quad (4.4)$$

$$N = \frac{3600}{t} \quad (4.5)$$

where: N refers to the maximum number of trains that can pass through the line within 1h, t is the minimum tracking interval of the train. The control value of the minimum tracking interval of the train generally occurs during the parking operation of the front train. When multiple trains travel in sequence along the same track and the same direction [19], there must be sufficient tracking interval between the follow-up train and the front train to ensure a certain safe distance between adjacent trains, so as to avoid abnormal braking, parking or collision of the follow-up train. Therefore, the tracking interval of the station is used to calculate the line passing capacity. The line passing capacity ignores many factors affecting the capacity in the actual line operation, so it is only discussed here as the upper bound of the theoretical capacity. The relationship between the line capacity and the maximum train speed under different tracking modes is shown in Figure 4.2.

See Table 4.1 for station tracking interval and passing capacity under different tracking modes when $v_{max} = 72km/h$. It can be seen from Figure 4.1 and Figure 4.2 that the station train tracking interval is not the greater the speed, the smaller the interval, but the minimum tracking interval under a special speed [20]. After exceeding this speed limit, the greater the speed is, the greater the tracking interval will be, and the corresponding

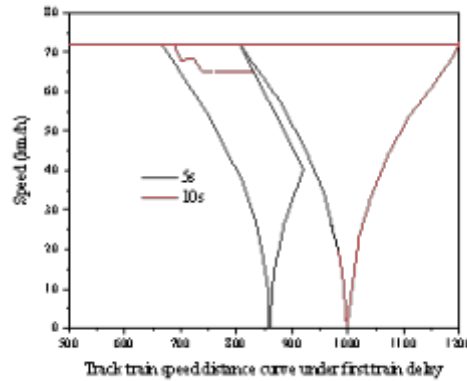


Fig. 4.3: Speed distance curve of tracking train under head train delay

theoretical passing capacity will be smaller and smaller. According to the simulation data in table 4.1, under the same maximum driving speed limit, the tracking interval of the moving block station is the largest, and the tracking interval of the improved station tracking model is the same as that of the relative moving block, while the station tracking interval of virtual reconnection model is the smallest. For the corresponding passing capacity, virtual reconnection model is the largest, the relative moving block is equal to the improved station tracking model, and the moving block is the smallest.

It can be seen that under the moving block system, adding the concepts of virtual reconnection model and improved station tracking model, will theoretically improve the line passing capacity (in the data in Table 4.1, the passing capacity of virtual reconnection model is 64.1% higher than that of moving block, and the passing capacity of improved station tracking model is 11.3% higher than that of moving block). This is only two train formation. If more trains are formed, the line passing capacity will be further improved [21,22,23]. But correspondingly, the train control system will be more complicated.

4.2. Dynamic performance measurement. The simulation scene and simulation parameters of the system are set as follows: $L_T = 140m$, $L = 2000m$. The total evolution time of the system is 2000 s, the acceleration and deceleration of the train are $1m/s^2$, maximum operating speed $v_{max} = 72km/h$, safety interval $L_S = 15m$. The system sets up a station at 1000m in the center, and the parking position of the locomotive in the station is 1000m. In particular, the simulation scene in this paper is that only the storage line and turn back line (or arrival departure site) are set at the end station, one station line is reserved in the upper and lower directions of the section transfer station, and there is no turnout connection with the section main line. Therefore, there is no "station arrival departure interval" at the station.

Figure 4.3 is the speed distance curve of follow-up tracking trains when the initial delay of the first train occurs at the station under moving block (the departure interval and minimum tracking interval of trains are all 58s). As can be seen in Figure 4.3, when the first train has no initial delay, the follow-up tracking train will slow down and stop evenly, and accelerate to leave the station after arriving at the station without interference by the first train [24]. However, when the head train is delayed, the tracking train will deviate from the planned speed distance curve. The longer the head train is delayed, the farther the tracking train deviates from the planned speed distance curve, and even stops outside the station waiting for the outgoing train, which is consistent with the actual situation.

In order to evaluate the delay propagation of initial delay in different systems under different tracking modes, the number of delayed trains caused by initial delay is calculated during simulation. Figure 4.4 is the relationship between different departure intervals and the number of train delays when the initial delay of the first train is 120s.

It can be seen from Figure 4.4 that appropriately increasing the departure interval can effectively reduce

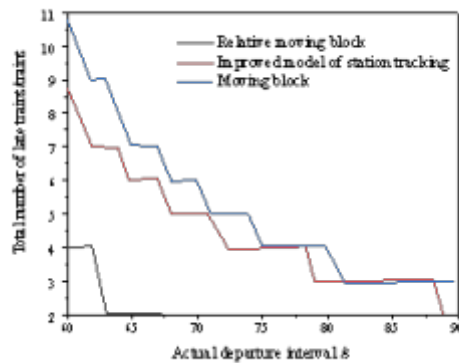


Fig. 4.4: Relationship between departure interval and train delay number under different tracking modes

the number of follow-up train delays caused by the initial delay of the first train [25]. On the premise of a certain actual departure interval, the number of delayed trains is reduced to a certain extent compared with moving block under the conditions of improved station tracking model in most cases. The number of delays in the improved station tracking model is basically the same as that in the relative moving block. The virtual reconnection model has the least number of delayed disturbed trains.

5. Conclusion. Virtual reconnection technology is a technology to control the tracking operation of trains on the track in formation. This technology uses vehicle-to-vehicle communication to coordinate the operation of each train, so as to realize the operation of "safe space dimension and closer time dimension". It improves the flexibility of urban rail transit operation, so as to adapt to the changing traffic demand and improve the passing capacity of the line. Aiming at the bottleneck area of the station, a tracking interval model based on virtual reconnection is proposed: that is, the combination of two trains is "one virtual reconnection train set". The relative moving block is used between the tracking train and the head train, and between the train set and the front train set. Then an improved station tracking model is proposed: relative moving block tracking is adopted only in the station area and moving block tracking is adopted in the section.

Therefore, on the premise of large space in the station platform, the design of virtual reconnection technology for the station area is an effective and feasible method to improve the overall line passing capacity. Under the condition of limited space in the station platform, the improved station tracking model can further improve the line passing capacity. In the aspect of running time optimization, the train running time in the section is only required within the allowable error range. In actual operation, when the train runs in multiple stations (multiple sections), it cannot only ensure that a certain section meets the requirements of operation time, but also need to adjust the overall operation time. For example, when the train runs ahead, it needs to redistribute the excess time. Or when lagging behind, it needs to speed up to reduce the delay, Therefore, optimizing the timetable can better achieve the purpose of optimizing the speed of the train. This system can not only meet the needs of the subway train running simulation system, but also provide various reference information for the subway line designers to make decisions about its design. It has high fidelity and credibility to provide drivers with simulation training for driving under various signal systems on different lines.

REFERENCES

- [1] Yubao, Q., Weifeng, Q., Jiayi, L., Feng, G., & Qiang, Z. (2017). Application of virtual simulation and computer technology in experiment and practical teaching. *Revista de la Facultad de Ingenieria*, 32(2), 450-459.
- [2] Zhao, C. (2021). Application of virtual reality and artificial intelligence technology in fitness clubs. *Mathematical Problems in Engineering*, 2021(20), 1-11.

- [3] Y Ding, Y Li, & Cheng, L. (2020). Application of internet of things and virtual reality technology in college physical education. *IEEE Access*, PP(99), 1-1.
- [4] Wu, G., Yang, R., Li, L., X Bi, Liu, B., & Li, S., et al. (2019). Factors influencing the application of prefabricated construction in china: from perspectives of technology promotion and cleaner production. *Journal of Cleaner Production*, 219(MAY 10), 753-762.
- [5] Z Lü, Sheng, W., Liu, H., Sun, L., & Li, R. (2017). Application and challenge of virtual synchronous machine technology in power system. *Zhongguo Dianji Gongcheng Xuebao/Proceedings of the Chinese Society of Electrical Engineering*, 37(2), 349-359.
- [6] Liu, Y., Hamid, Q., Snyder, J., Wang, C., & Wei, S. (2016). Evaluating fabrication feasibility and biomedical application potential of in situ 3d printing technology. *Rapid Prototyping Journal*, 22(6), 947-955.
- [7] Song, H., & Schnieder, E. (2019). Development and validation of a distance measurement system in metro lines. *IEEE Transactions on Intelligent Transportation Systems*, 20(2), 441-456.
- [8] Zhang, M., Zhu, Z., & Tian, Y. (2020). Application research of virtual reality technology in film and television technology. *IEEE Access*, PP(99), 1-1.
- [9] Dong, J. (2017). Research and application of virtual reality technology in the restoration of ancient buildings in huizhou. *Acta Technica CSAV (Ceskoslovensk Akademie Ved)*, 62(1), 289-299.
- [10] Zhanjun, W. (2017). Application research of virtual reality technology in environmental art design. *Acta Technica CSAV (Ceskoslovensk Akademie Ved)*, 62(1), 215-224.
- [11] Li, Z., Huo, G., Feng, Y., & Ma, Z. (2021). Application of virtual reality based on 3d-cta in intracranial aneurysm surgery. *Journal of Healthcare Engineering*, 2021(1), 1-11.
- [12] Gao, R. (2017). Application of modern measurement technology in energy demand forecasting. *Revista de la Facultad de Ingenieria*, 32(14), 576-581.
- [13] Zheng, P., & Huang, L. (2017). Research on the application of 3d virtual reality technology in the protection and development of ancient villages. *Revista de la Facultad de Ingenieria*, 32(5), 766-774.
- [14] Zhou, H. (2017). Application and research of virtual reality technology based on udk in interior design. *Boletin Tecnico/Technical Bulletin*, 55(19), 456-461.
- [15] Zhao, B., Yang, H., D . Wang, & Xu, N. (2017). Application of computer aided technology and visual system design in engineering cost management. *Boletin Tecnico/Technical Bulletin*, 55(18), 359-365.
- [16] Yang, J., & Jin, H. (2020). Application of big data analysis and visualization technology in news communication. *Computer-Aided Design and Applications*, 17(S2), 134-144.
- [17] Yongwei, Li, Yuman, Li, Hongfei, & Wang, et al. (2018). Application of control quality evaluation technology in complex industrial process. *International Journal of Computer Applications in Technology*, 57(2), 149-156.
- [18] Wu, L., & Peng, H. (2017). Application of isp technology in the design of intelligent instruments. *Acta Technica CSAV (Ceskoslovensk Akademie Ved)*, 62(1), 483-493.
- [19] Di Nardo, M.; Murino, T.; Osteria, G.; Santillo, L.C. A New Hybrid Dynamic FMECA with Decision-Making Methodology: A Case Study in an Agri-Food Company. *Appl. Syst. Innov.* 2022, 5, 45.
- [20] Zhang, Y., & Meng, F. (2017). Research on optimization technology and application of bim in building optimization. *Revista de la Facultad de Ingenieria*, 32(6), 233-240.
- [21] Li, W., Luo, Q., Ca I, Q., & Zhang, X. (2018). Using smart card data trimmed by train schedule to analyze metro passenger route choice with synchronous clustering. *Journal of Advanced Transportation*, 2018(PT.4), 2710608.1-2710608.13.
- [22] Bai, Y., Hu, Q., Ho, T. K., Guo, H., & Mao, B. (2019). Timetable optimization for metro lines connecting to intercity railway stations to minimize passenger waiting time. *IEEE Transactions on Intelligent Transportation Systems*, PP(99), 1-12.
- [23] Li, S., Yang, L., & Gao, Z. (2019). Efficient real-time control design for automatic train regulation of metro loop lines. *IEEE Transactions on Intelligent Transportation Systems*, 20(2), 485-496.
- [24] Feng, L., & Zhao, J. (2017). The study on the application of the intelligent technology in the sightseeing agricultural parks. *Revista de la Facultad de Ingenieria*, 32(15), 12-17.
- [25] Pang, C., Huang, S. C., Liu, J. C., & Zhao, W. (2017). Multi sensor cross cueing technology and its application in target tracking. *Yuhang Xuebao/journal of Astronautics*, 38(4), 401-409.

Edited by: Bradha Madhavan

Special issue on: High-performance Computing Algorithms for Material Sciences

Received: Jan 17, 2024

Accepted: Mar 26, 2024



THE EMPLOYMENT OF CARBON NANOTUBES IN BIOMEDICAL APPLICATIONS

JAJAAR FAHAD A. RIDA*

Abstract. Carbon nanotubes (CNTs), a prominent application of nanotechnology, find extensive use across various fields. Their electrical and optical characteristics, which are affected by the manufacturing process and any impurities introduced during production, are crucial in establishing their suitability for use. This research focuses on the utilization of carbon nanotubes in medical applications, exploring their properties both as electrical conductors and semiconductors, comparable to silicon used in precision medical equipment and devices. When functioning as electrical conductors, CNTs exhibit characteristics similar to traditional conductive materials. This property is harnessed in medical applications, particularly in targeted cancer treatments that minimize impact on healthy cells. CNTs' efficient conduction of electrical current makes them valuable components in medical devices and equipment. Furthermore, CNTs showcase semiconductor properties akin to silicon. This characteristic is crucial for developing advanced medical equipment, enabling accurate diagnostics and medical imaging. The semiconductor behavior allows the creation of intricate medical devices with enhanced precision. The research underscores the significance of CNTs in shaping the future of medical technology, especially when integrated with artificial intelligence applications. The ability of CNTs to function both as conductors and semiconductors highlights their versatility in the medical field, promising advancements in healthcare technologies. Their use holds potential for targeted cancer treatments, accurate diagnostics, medical imaging, and enhanced performance through integration with artificial intelligence.

Key words: Carbon Nanotubes, Biomedical Engineering, Nanotechnology Engineering, Bio- Nanotechnology, and Carbon Nanotube Applications

1. Introduction. Nanotechnology, emerging as an alternative to microtechnology, introduces the possibility of manufacturing nanoelectronic and electromechanical devices, drastically reducing their size compared to micro devices. This transformative innovation is anticipated to bring about significant advancements across various scientific and engineering disciplines. Enthusiasts foresee its broad influence on contemporary medicine, the global economy, international relations, and the daily lives of individuals. The potential to arrange matter particles in unprecedented ways at a lower cost sparks imaginations of supercomputers integrated into pen tips, and fleets of medical nanorobots administered to treat blood clots, tumors, and currently incurable diseases[[1], [2]]. The study highlights the contrast between micro and nano technologies, emphasizing nanotechnology's potential to revolutionize electronic and electromechanical devices. In contrast to micrometers, nanoelectronic and electromechanical devices can be reduced in size by a factor of a thousand, resulting in enhanced performance. This paradigm shift is not merely a theoretical concept; it has tangible applications across various industries. For instance, polymers in micro-devices lead to increased device longevity, while metals like gold, nickel, and aluminum contribute to the reliability of early devices [3], [4]. The discussion extends to the unique properties of nanomaterials, showcasing their exceptional hardness, transparency, and transformative effects on material behavior. Spherical nanoparticles made of silicon, ranging from 40 to 100 nanometers, exhibit hardness surpassing even that of sapphire and approaching that of diamond. Transparency, a characteristic of nanoparticles due to their dimensions being smaller than light wavelengths, opens up possibilities for applications like transparent packaging and cosmetic products [5], [6], [7].

Nanotechnology is acknowledged as the fifth generation of electronic technologies, following the progression from electronic valves to transistors, integrated circuits, and microprocessors. The development of nanoelectronic and electromechanical devices, facilitated by advancements in synthetic chemistry, holds promise for applications in health, medicine, information technology, and beyond. Notable examples include IBM's creation of a microscope for imaging and recording atoms at the Nano level and the aspiration to replace electricity with light ([8], [9], [10]) potentially leading to the advent of optical computers such as shown in figure 1.1.

*College of Engineering, University of Sumer, Dhiqar, Iraq. (j.fahad@uos.edu.iq & jafaarfahad@gmail.com & jafaarfahad@uos.edu.iq)

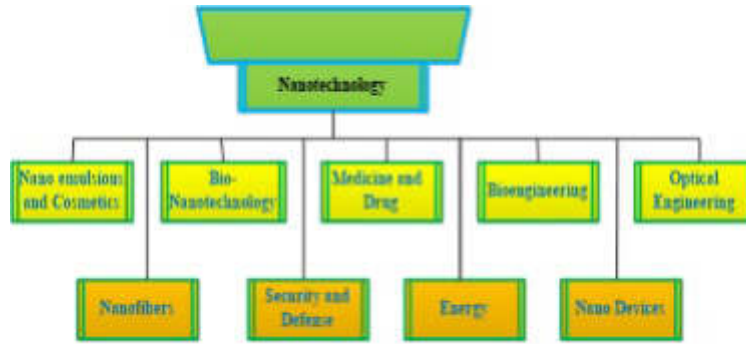


Fig. 1.1: Nanotechnology holds promising prospects and applications in many fields.

1.1. Overview of Carbon Nanotubes (CNTs). During the early 1990s, Sumio Iijima made a groundbreaking discovery of carbon nanotubes (CNTs), which brought about a significant transformation in numerous scientific and technical domains. Carbon, which is vital for human existence, serves as the fundamental building block for the bonding arrangement of diamond, graphite, nanotubes, and fullerenes. With a wide range of desirable properties, CNTs have become integral in structural science, material science, chemistry, biology, and electronics [11], [12]. CNTs, resembling stretched graphite strips, exhibit exceptional properties: tensile strength surpassing steel, superior thermal conductivity, and electrical conductivity equivalent to copper. Divided into single-wall (SWNTs) and multi-wall (MWNTs) categories, CNTs' key characteristics include diameter, chirality angle, and number of walls, each influencing unique physical and chemical properties. Fabrication techniques involve chemical methods and physical techniques such as chemical vapor deposition (CVD), arc discharge, and laser ablation. CNTs, incorporated into nanosystems like polymer electrical nano materials, possess a nanoscale diameter, contributing to their versatility. Electronic properties vary based on diameter and chirality, with approximately one-third exhibiting metallic structure and the rest being semiconducting. The optical properties of CNTs have garnered attention in photonics, showcasing a short recovery time and high third-order optical nonlinearity. Chiral nanotubes, categorized based on chiral vectors, contribute to the diverse landscape of carbon nanotubes. The interplay between valence electrons and the lattice in rigid covalent-bond materials impacts the electronic structure. Calculations show that around one-third of CNTs have a metallic structure, dependent on nanotube diameter and chiral angle [13], [11], [14].

The structure of single-walled carbon nanotubes (SWNTs) is intricately linked to their electrical properties, particularly influenced by the chiral vector (Ch), which determines the orientation of the honeycomb lattice. This unique parameter serves as the key characteristic affecting whether SWNTs behave like metals or semiconductors. The chiral vector, represented by (n, m) indices, defines the tube's structure, with n and m being integers that influence the nanotube's diameter and chirality angle (θ) . It is essential to comprehend the structure of the end of a single-walled carbon nanotube (SWNT), which is commonly described using Hamada indices (n, m) , as this knowledge is critical for accurately predicting and controlling the electrical characteristics of SWNTs. The chiral angle and the specific arrangement of carbon atoms significantly impact the electronic structure. The correlation between the electrical structure and geometry, as demonstrated in fullerenes, is proposed to be a universal trait in nanostructured carbon materials such as carbon nanotubes. This reliance is attributed to the enhanced interaction between valence electrons and the lattice in rigid C-C covalent-bond materials. In materials with stiff covalent bonds, valence electrons interact more strongly with the lattice, influencing the electronic structure based on geometrical details. The electrical energy band structure of a nanotube is intricately connected to the energy band structure generated by the 2D graphite honeycomb sheet used in the production of the nanotube. This connection highlights the significance of the underlying hexagonal lattice

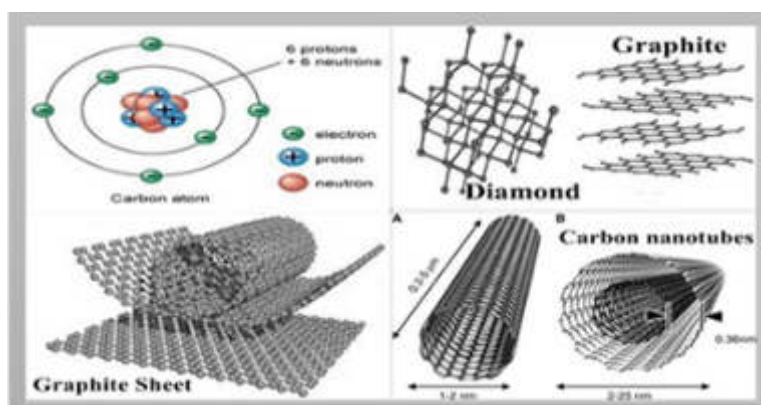


Fig. 1.2: Full carbon nanotube (CNT) manufacturing process, from atomic carbon to graphite sheets to rolled-up form tubes.

structure in determining electronic properties. Calculations reveal that approximately one-third of carbon nanotubes exhibit a metallic structure, while the remaining two-thirds have a semiconducting structure [14][2][15]. This observation is contingent on factors such as nanotube diameter (dt) and the chiral angle (θ). Chiral nanotubes, categorized as zigzag and armchair nanotubes, represent another classification based on chiral vectors (Ch), adding to the diversity of carbon nanotube structures as observed in the figure 1.2.

Carbon nanotubes (CNTs) have potential biomedical applications in various fields. They have applications in antimicrobial materials, dentistry, drug delivery, biosensing, cancer therapy, tissue engineering, diagnostic imaging, and regenerative medicine. Carbon nanotubes (CNTs) have distinctive characteristics, including a large surface area, exceptional mechanical robustness, electrical conductivity, and thermal properties, which render them well-suited for these specific applications. Functionalization of CNTs enhances their biocompatibility and enables biomolecule loading for targeted drug delivery and immobilization support. Carbon nanotubes (CNTs) can undergo modifications with diverse functional groups to enable the concurrent transportation of many molecules, facilitating targeted delivery, therapeutic interventions, and imaging purposes. They have been employed for the delivery of tiny medicinal molecules, peptides, proteins, and genes and have demonstrated therapeutic effectiveness in both *in vivo* and *in vitro* experiments. In addition, carbon nanotubes (CNTs) have been utilised in the advancement of biosensors for the detection of biological and biomedical chemicals. However, there are challenges related to cytotoxicity and biodegradation that need to be addressed for their safe implementation in clinical trials [16], [17].

Properties of carbon nanotubes have High Strength that mains Carbon nanotubes are known for their exceptional strength, making them ideal for use in biomedical implants and scaffolds [3], [18], [19]. Electrical Conductivity is they exhibit excellent electrical conductivity, allowing for applications in bioelectronics and biosensors. Thermal Stability is with high thermal stability, carbon nanotubes are suitable for various biomedical applications, including heating-based therapies. Carbon nanotubes (CNTs) exhibit unique optical properties that make them highly valuable in various applications. Some key aspects of their optical properties include: Optical Absorption and Emission are CNTs demonstrate strong optical absorption in the near-infrared region. Their emission properties can be tuned based on the nanotube structure, offering possibilities for applications in sensors and imaging. Photoluminescence is Carbon nanotubes can emit light upon absorbing photons, a phenomenon known as photoluminescence. This property is influenced by the nanotube's diameter and chirality, providing opportunities for designing nanoscale light sources and devices [20]. Nonlinear Optical Behavior is CNTs exhibit nonlinear optical behavior, making them suitable for applications in nonlinear optics. Their response to intense light can be harnessed for developing optical switches and modulators. Optical Transparency is Depending on their structure, some carbon nanotubes are optically transparent. This transparency, combined with their excellent electrical conductivity, is advantageous for applications in transparent conductive films and coatings. Light Polarization can be Carbon nanotubes exhibit polarization-dependent optical prop-

erties. This polarization sensitivity is beneficial in designing devices for polarized light applications, such as in optoelectronics and photodetectors. Light Scattering and Reflection have CNTs can scatter and reflect light, and their interaction with light is influenced by factors like diameter and length. These properties are relevant in applications such as anti-reflective coatings and light-absorbing materials. Photoconductive Response have Carbon nanotubes can show a photoconductive response, meaning their electrical conductivity changes upon exposure to light. This property is utilized in developing light-sensitive devices like photodetectors and photovoltaic cells. Broadband Absorption: CNTs have broadband absorption capabilities, covering a wide range of the electromagnetic spectrum [13], [21], [22]. This feature is advantageous for applications in solar cells and broadband photodetectors.

Utilizing carbon nanotubes (CNTs) in biomedicine presents promising opportunities, but it also comes with several challenges and limitations that need to be addressed. Some key considerations include: Biocompatibility Concerns is the biocompatibility of carbon nanotubes is a significant challenge. Some studies have raised concerns about potential toxicity and inflammatory responses when CNTs interact with biological systems. Addressing these issues is crucial for safe biomedical applications. Functionalization and Surface Modifications are Pure carbon nanotubes may lack specific functional groups required for targeted drug delivery or interactions with biological molecules. Surface modifications are often necessary to enhance biocompatibility, solubility, and the attachment of biomolecules [3], [4], [23], [24]. Biodistribution and Clearance have Understanding the biodistribution and clearance of carbon nanotubes from the body is essential for their safe use. The long-term fate of CNTs, especially in terms of potential accumulation in organs or tissues, needs careful investigation. Regulatory Approval can be the regulatory approval process for medical applications involving carbon nanotubes is challenging. Establishing standardized protocols for testing and ensuring the safety and efficacy of CNT-based biomedical products is crucial for regulatory acceptance. Large-Scale Production is Scaling up the production of high-quality, well-characterized carbon nanotubes for biomedical applications remains a challenge. Ensuring consistency in size, structure, and purity is essential for reproducibility in research and clinical settings. Cost is the production and functionalization of carbon nanotubes can be expensive, limiting their widespread adoption in healthcare. Cost-effective manufacturing methods need to be developed to make CNT-based technologies more accessible. Limited Understanding of Long-Term Effects can be the long-term effects of exposure to carbon nanotubes, especially in the context of chronic diseases or repeated treatments, are not fully understood. Further research is needed to assess any potential cumulative impact on health over extended periods. Interaction with Immune System has the interaction between carbon nanotubes and the immune system is complex. Depending on their properties, CNTs can trigger immune responses, impacting their effectiveness and safety in biomedical applications. Intracellular Fate is understanding how carbon nanotubes behave inside cells and their intracellular fate is crucial. This includes investigating whether they remain intact, undergo degradation, or lead to the formation of toxic byproducts. Multifunctionality Challenges are while the multifunctionality of carbon nanotubes is advantageous, integrating multiple functionalities (e.g., imaging, drug delivery, and sensing) in a single platform without compromising performance remains a technical challenge. Carbon nanotubes (CNTs) possess unique properties such as large surface area, mechanical strength, electrical conductivity, and biocompatibility, making them ideal for biomedical applications. CNTs have been used as antibacterial agents, dental materials (scaffolds, bone-grafting, tissue engineering), and drug delivery systems for cancer therapy. The modifications of CNTs with metal and metal oxide nanoparticles, such as zinc oxide (ZnO), have enhanced their antibacterial properties [13], [22], [25], [26], [27].

The graphite's hexagonal mesh structure, which serves as the basis for carbon nanotubes, can be conceptualised as a cylindrical formation resembling rolled-up chicken wire, owing to the organisation of carbon atoms in stacked layers. Within the electrical density of states (DOS) of carbon nanotubes, there are singularities called Van Hove Singularities. Each nanotube possesses four distinct energy levels two for conduction and two for valence (CNTs). In contrast to metal carbon nanotubes, semiconducting carbon nanotubes exhibit a direct band gap that increases with the nanotube diameter, enabling them to efficiently conduct electrical current. Single-walled nanotubes generally exhibit sizes within the range of 1-2 nm, but multi-walled nanotubes can exhibit diameters ranging from 2-25 nm. Furthermore, nanotubes can vary in length, ranging from 0.2 to 5 micrometers, which provides a diverse array of structural options. Carbon nanotubes can exhibit either metallic or semiconducting properties, depending on their structure and orientation. This dual nature makes

them valuable for a wide range of electronic applications, from high-conductivity components to semiconducting devices in advanced technology and nanoelectronics [7], [21], [28], [29]. The authors Mohd et al. (2023) provided information. The user's text is empty. Carbon nanotubes possess significant potential for utilisation in several biomedical applications, including the fabrication of antibacterial materials, enhancement of dental operations, drug delivery, and the advancement of biosensors. The study conducted by K. Victor et al. (2022) highlighted the several potential biomedical uses of carbon nanotubes, which encompass therapeutic, tissue engineering, diagnostic, and imaging applications. Additionally, carbon nanotubes can be utilised for drug transport and exhibit antibacterial properties. [Sarika Verma et al, (2023)] described The potential biomedical applications of carbon nanotubes include their usage in diagnoses, tissue regeneration, selective drug delivery, and as tissue engineering scaffolds. [Mahdiah Darroudi et al, (2023)] explained Carbon nanotubes have the potential to be used in several biomedical applications, such as diagnosing, treating, and preventing infectious and neoplastic disorders. They can also be used for gene transfer and anti-inflammatory therapy. [Lopamudra Giri et al. (2023)] discussed the biomedical applications of carbon nanotubes, highlighting their potential in various biological applications. [Duygu Harmanaci et al. (2023)] discussed Carbon nanotubes (CNTs) show great potential for theranostic applications in various fields, including cancer diagnosis and therapy, infectious diseases, central nervous system problems, and tissue engineering. Functionalized carbon nanotubes (CNTs) have been successfully used in pharmaceuticals and medicine due to their unique properties . Study of the properties of carbon nanotube when it acts as a conductor of electric current or acts as a semiconductor, based on the materials added to the graphite material, as well as on the valence band and conduction in each nucleus and the opening between them. It is considered a carbon nanotube that is symmetrical, unlike semiconductors made of silicon, which are asymmetrical and the bias voltage is high. Compared to the effort required for carbon nanotubes. Single walled carbon nanotubes (SWNTs) have a single cylindrical wall. The typical diameter of the nanotubes falls within the range of 1 to 2 nm, although their length can vary from 0.2 to 5 μm , and in some cases, even extend to a few centimeters, as depicted in figure 3. It has been processed to have optical qualities that are compatible with optical systems. The photonic and biological research communities have been paying an increasing amount of attention to the possible uses of carbon nanotubes. The latter property has given rise to dreams of using nanotubes to make extremely dense electronic circuitry and the last year has seen major advances in creating basic electronic structures from nanotubes in the lab, from transistors up to simple logic elements. The volumes of SWNTs produced are currently small and the quality and purity are variable. Multi-walled carbon nanotubes have two or more cylinders within cylinders. Nanotubes typically have diameters in the range of two to twenty-five nanometers, and their lengths may vary anywhere from two to five micrometers or a few centimeters. The gap between the walls is 0.36 nm, and the spacing between walls is 0.36 nm. In these progressively more sophisticated systems, the several SWNTs that combine to form the MWNT could have quite different architectural compositions (length and chirality). MWNTs have an average length that is one hundred times greater than their width, and their outside diameters are almost always measured in the tens of nanometers. [5], [8], [10], [13], [22]. Despite the fact that it is simpler to produce significant quantities of multi-wall nanotubes than single-wall nanotubes, the structures of multi-wall nanotubes are not as well understood as those of single-wall nanotubes due to the greater complexity and variety of multi-wall nanotubes, as shown in figure 1.3.

2. Research Methodology. When a single sheet of graphite crystal is rolled up into a cylinder, the result is a single wall carbon nanotube (SWTN). This cylinder has a thickness of one atom and a relatively low density of atoms (2040 per micron in diameter and length) along its axis. The nanotube represented by the chiral vector C_h .

$$C_h = n * a_1 + m * a_2 \quad (2.1)$$

where n and m are two integers denoting the number of unit vectors $n * a_1$ and $m * a_2$ in the hexagonal honeycomb lattice contained in this vector, where $|a| = |a_1| = |a_2|$ and where $|b| = |b_1| = |b_2|$ and where $|c| = |c_1| = |c_2|$ Graphite lattice vectors a_1 and a_2 are a pair of real space vectors with the following values: $a_1 = (a, 0)$, $a_2 = (0, a)$, where $a = 0.246\text{nm}$, where $a_3 = (a, a)$, where $a_4 = (0, a)$ is the c - c bond length. As can be seen in Figure 1.3, the chiral vector may be used in conjunction with the zigzag or the direction to create an angle (the chiral angle. On a graphene sheet with a honeycomb structure, the vector connects two O and A sites that are crystallographically

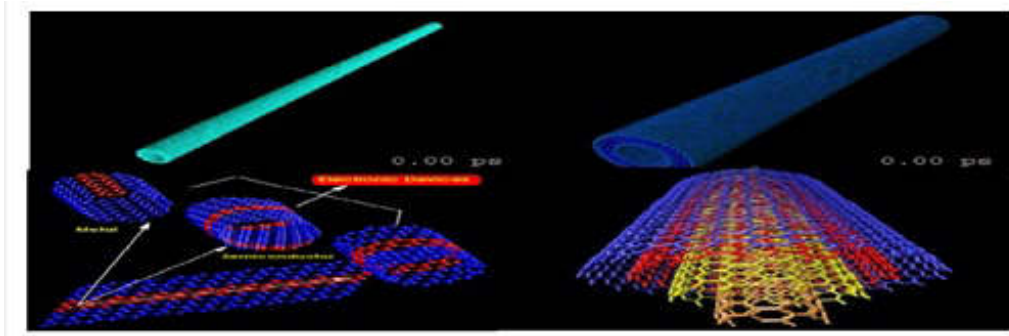


Fig. 1.3: Schematic representation of rolling graphite to create single- and multi-walled carbon nanotubes.

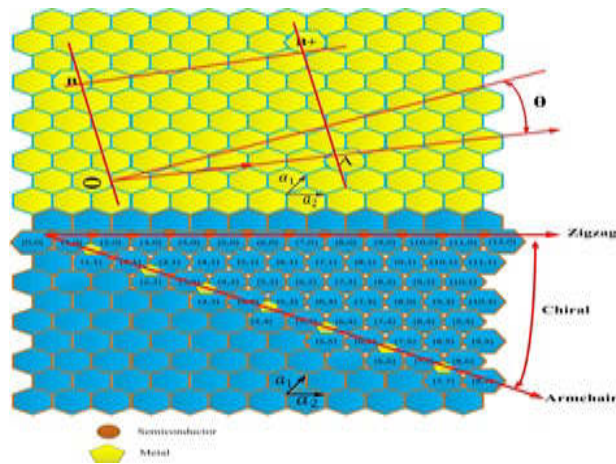


Fig. 2.1: Clarification of the process through which graphite sheets are converted into carbon nanotubes

similar to one another. Each vertex of the honeycomb structure contains a carbon atom. The axis of the zigzag nanotube is located at the value $\phi = 0$ for the parameter, the axis of the armchair nanotube is located at the value $\phi = 30$ for the parameter, and the axis of the chiral nanotube is located at a value that is $0 < \phi < 30$. Attaching the line AB to the parallel line OB in figure 2.1 creates the smooth cylinder connection that the nanotube needs in order to function properly [30], [31], [32], [33]. The following is an equation that expresses the diameter of a nanotube, denoted by the notation d_t , in terms of the numbers n and m: equation 2.2.

$$d_t = \frac{|a| * \sqrt{(n^2) + n * m + m^2}}{\pi} \tag{2.2}$$

The distance between the two carbon atoms of the closest neighbor is 1.421 or 0.142 in graphite, C_h is the length of the chiral vector, and the chiral angle (ϕ) may be obtained by solving the following equation 2.3.

$$\phi = \tan^{-1} \frac{\sqrt{3} * m}{2n + m} \tag{2.3}$$

Thus, the (n, m) indices or their equivalent, d_t , may be used to describe a nanotube.

Figure 2.1 shows a) The unit cell of one-dimensional nanotubes is described using the nomenclature of the unit cell of the honeycomb lattice, which is often encountered in two dimensions; b) the Brillouin zone of two-dimensional graphite represented by rhombuses and a shaded hexagon; and c) The unit cell of one-

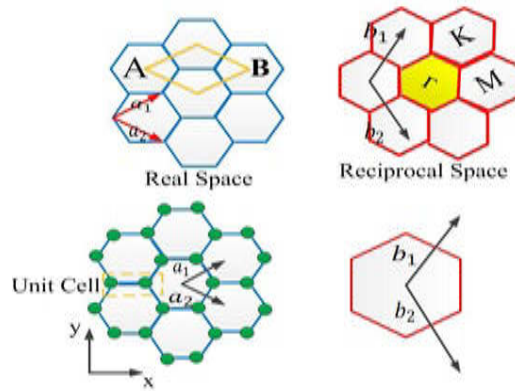


Fig. 2.2: Actual and reciprocal features of the structure's fundamental building block, the unit cell

dimensional nanotubes is described using the nomenclature of the unit cell of the honeycomb lattice, which is often encountered in two dimensions.

Figure 2.2's coordinates (x, y) are used to express the real space basis vectors a_1 and a_2 , which are written as equations 2.4 and 2.5 respectively.

$$a_1 = \left(\frac{\sqrt{3}}{2}a, \frac{a}{2}\right), a_2 = \left(\frac{\sqrt{3}}{2}a, -\frac{a}{2}\right) \tag{2.4}$$

and

$$b_1 = \left(\frac{2\pi}{\sqrt{3}a}, \frac{2\pi}{a}\right), b_2 = \left(\frac{2\pi}{\sqrt{3}a}, -\frac{2\pi}{a}\right) \tag{2.5}$$

where $a = |a_1| = |a_2| = 0.246\text{nm}$, a is the lattice constant of two-dimensional graphite, which is consequently the basis vectors b_1 and b_2 of the reciprocal lattice. b_1 and b_2 are the reciprocal lattice's basis vectors. The reciprocal ($b=2.949\text{nm A}$) hexagonal lattice has basis vectors b_1 and b_2 that are rotated by 30 degrees with respect to the real space hexagonal lattice's basis vectors a_1 and a_2 . To do this, we choose the shaded region of Figure 2.2 to represent the first brilluoin zone, the centers of Points K and M, and the corners of Figures 2.1 and 5 to represent the highest symmetry points. Calculations have been made to determine the energy dispersion relations for the MK triangle depicted in Figure 2.1 with dotted lines. As shown in figure 1.3, the unit cell is defined by the smallest repetition distance along the axis of the 1D nanotube, which is denoted by the letter OB. This allowed us to determine the translation vector (T).

$$T = t_1 * a_1 + t_2 * a_2 \tag{2.6}$$

The equation 2.7 the variables n and m are connected to the coefficients t_1 and t_2 , respectively.

$$t_1 = \frac{(2n + m)}{(dR)}t_2 = \frac{-(2n + m)}{(dR)} \tag{2.7}$$

where dR is the greatest common divisor of $(2n+m, 2m+n)$, and the equation 2.8 for dR may be obtained by clicking here.

$$dR = \begin{cases} d, & \text{if } n - m \text{ is not a multiple of } 3d \\ d3, & \text{if } n - m \text{ is a multiple of } 3d \end{cases} \tag{2.8}$$

where d is the greatest common divisor of all the other (n, m) . The magnitude of the translation vector T , equal to the vertical bar symbol.

$$|T| = \frac{\sqrt{3} * L}{dR} \quad (2.9)$$

The length of the chiral vector C_h is denoted by L , while the diameter of the nanotube is denoted by d_t . The region bounded by the vectors T and C_h is the area that is referred to as the nanotube's unit cell. The values (n,m) and the formula are used to determine the number of hexagons, N , that are contained within the one-dimensional unit cell of a nanotube. This number may be thought of as the number of individual nanotubes 2.10.

$$N = \frac{(2(n^2 + n * m + m^2))}{dR} \quad (2.10)$$

Two carbon atoms are added, which is represented by a single hexagon in the honeycomb structure shown in Figure 1.2. We chose carbon nanotubes with the following chirality ratios $(17, 0)$, $(12, 5)$, $(10, 10)$, $(16, 2)$, $(15, 0)$, and $(11, 11)$ based on the assumption that a $(c-c)=0.142$ nm. The nanotube has a larger real space unit cell compared to the 2D graphene sheet, resulting in a substantially smaller 1D Brillouin zone (BZ) for the nanotube compared to a single 2D graphene unit cell. The reason for this is that the actual spatial unit cell for the nanotube is far greater than that for the 2D graphene sheet. Brillouin Zone-folding methods have been extensively utilised to establish approximate connections between the dispersion of electrons and phonons in carbon nanotubes (n, m) possessing specified symmetry. The user's text is empty. The nanotube's crystal structure closely mimics that of a graphene sheet, which is why the Brillouin Zone is relatively modest. The vectors k_1 and k_2 may be calculated using the relationship. $R_i * K_j = 2 * i * j$, where the lattice vectors in real space are denoted by $R(i)$, and the vectors in reciprocal space are denoted by K_j . It is possible to write form k_1 and k_2 as

$$k_1 = 1/N(-t_2 * b_1 + t_1 * b_2) \text{ and } k_2 = 1/N(m * b_1 - n * b_2) \quad (2.11)$$

The reciprocal lattice vectors of a graphene sheet in two dimensions are represented by the symbols b_1 and b_2 , which are correspondingly indicated by the equation that describes them 2.5. Given a set of N wave vectors $k(1) (=0, \dots, N-1)$, it is feasible to compute N discrete k vectors in the circumferential direction. A one-dimensional band of electronic energy appears for each of the discrete values of the circumferential wave vectors, and phonon dispersion relations extend in six distinct directions depending on the value of $.$ Creating a nanotube from a graphite sheet by rolling it along the chiral vector C_h . The resulting nanotube is rolled (n,m) . Nanotubes can be characterised by their diameter (d_t) and chiral angle (π) in relation to the zigzag axis. Single-wall nanotubes (SWNTs) may be thought of as hollow cylinders that are formed by rolling a graphite sheet. It is possible to describe it in an unambiguous manner by a vector denoted by the letter C_h in terms of a set of two numbers denoted by the letters n and m , which correspond to graphite vectors a_1 and a_2 . It is possible to create two standard nanotubes from a single graphite sheet by rolling it in opposite directions. The nanotubes can exist in three different configurations: zigzag $(n, 0)$, armchair (n, m) when $n = m$, and chiral (n, m) . If n is a value greater than m and less than zero, the coordinates of the chiral nanotubes are as follows: $(17, 0)$, $(15, 0)$, $(12, 5)$, $(16, 2)$, $(10, 10)$, and $(11, 11)$. Both the lattice constant and the intertube spacing are essential requirements for the proper creation of a bundle of single-walled carbon nanotubes (SWNTs). Experiments and theoretical studies have reached the conclusion that the average length of C-C bonds in MWNT should be 0.34 nanometers, and that the spacing between tubes should also be 0.34 nanometers. In light of this, equations 2.1 and 2.2 may be used to simulate a variety of tube architectures and interpret experimental data. They are now taking into account the energetics or the stability of nanotubes. During the fabrication of a SWNT from a graphite sheet, the strain energy is equal to $1/d_t$ per tube, or $1/d_t^2$ per atom.

The diameter of SWNTs that are often seen in experiments ranges from 0.6 to 2.0 nanometers, however it may be as tiny as 0.4nm or as big as 10 nanometers (3.0nm). The electrical properties of a nanotube can

be determined by analysing the dispersion relation of a graphite sheet with wave vectors (k_x, k_y) , in the most basic scenario.

$$E(k_x, k_y) = \gamma \sqrt{4\cos((3 * k_x * a)/2)\cos((k_y * a)/2) + 4\cos^2((k_y * a)/2)} \quad (2.12)$$

Considering the stated values of $a = 0.246nm$, $\gamma = 2.5eV - 3.0eV$ (the lattice constant), and $a = 0.246nm$ (the closest neighbour hopping parameter), it is plausible that γ originates from many sources. In order to construct a nanotube from graphite, it is necessary to maintain a periodic boundary condition either around the circumference of the tube or in the C directions. This may happen either along the circumference of the tube or along the C directions [30], [31], [32], [34], [35], [36], [37], [38]. The two-dimensional wave vector $k = (k_x, k_y)$ is quantized in this direction due to the constraint $k.c = 2q$, where k that satisfies this condition is allowed when q is an integer. The following need, therefore, must be satisfied in order for metallic conductance to occur: equation 2.13

$$(n - m) = q_{metallic} (2n + m) = 3q \quad (2.13)$$

2.1. Semiconducting of Carbon Nanotubes . The valence energy band is located at the lower section of the energy curve, while the conduction energy band is located at the upper section of the energy curve. When s equals zero, the valence band and the conduction band become symmetric, adopting the shape of a ball, which is described by equation 2.12, and $\gamma = 2.9eV$. This occurs when the electron spin is equal to zero. 2D graphite is classified as a zero-gap semiconductor due to the conduction and valence bands intersecting at the six corners of the Brillouin zone, which are known as high symmetry sites $E = E_{sp}$ [30], [38], [39], [40]. The positive sign is used to represent the valence band, whereas the negative sign is used to indicate the conduction band. Graphene exhibits symmetrical conduction and valence bands in terms of both structure and distribution. On the other hand, silicon, which is an indirect band gap semiconductor, exhibits dissimilar band structures for both electrons and holes. The Fermi points are the places on the edges of the Brillouin zones where the energy troughs may be found. These sites are also known as the Brillouin zone corners. The vectors serve as the foundation for the reciprocal lattices b_j . The wave vectors were limited to the specified range as a direct consequence of the periodic boundary restriction that was applied in the circumferential direction .

$$k.c = 2 * \pi * q \quad (2.14)$$

where k is a wave vector that can be allowed, and q is an integer that denotes the quantum number. This may be allowed. The equation that describes the conductance of SWNT, MWNT rope, or SWNT cable is as follows 2.15:

$$G = G_O * M = ((2e^2)/h) * M \quad (2.15)$$

Since nanotubes' conductance is quantized, the resistance of nanotubes is equal to $6.5 * 10^3$ ohms; M is an apparent number of conducting channels that takes into account electron-electron coupling and intertube coupling effects in addition to the intrinsic channel $G = (6.5k_{ohm})^(-1)$ which is the value of the intrinsic channel. The results of the combined STM and STS studies are consistent with the following hypotheses: 1) approximately two-thirds of the nanotubes are semiconducting, and one-third of them are metallic; 2) the density of states exhibits van hove singularities, which is characteristic of the expectations for a one-dimensional system; and 3). The energy gaps of semiconducting nanotubes are proportional to the square root of the distance traveled over time $(1/d_t)$. An inductor's E_k value is determined by the extra kinetic energy associated with the current. This represents the equilibrium Fermi energy for electrons moving between the source and the drain of a field-effect transistor. This demonstrates that kinetic inductance has a ballistic origin. Therefore, making the right option while selecting the diameter is helpful. At a temperature of 300 kelvin, the typical values for the band gap are as follows: - (1.12 eV) for germanium; (0.67 eV) for silicon; and (1.43 eV) for GaAs. Electronic density of states (DOS) is measured in terms of the equilibrium Fermi energy per unit of nanotube length, whereas availability is expressed in terms of the number of electron-phonon pairs per unit of nanotube length. Both of these measures are independent of one another. Both of these measurements are expressed in nanotube length units. Due to the fact that the bias window only contains right-moving carriers, the density of states must be divided by two in the calculation for the mean number of electron-phonon couplings. This is because the bias window contains only the right-moving carriers [26], [27], [41], [42], [43].

2.2. Nanotechnology Applications in the Medical Sciences. Because of its close relationship to human life and health, nanomedicine is often regarded as one of the most significant uses of nanotechnology. Some even argue that it is the most significant application of all. This is the age of nanomedical technology, when the principles of illness prevention, diagnosis, and treatment have been rewritten in light of recent advances in nanotechnology. Whereas nanotechnology, by way of illustration, opens up novel pathways for drug carriers within the human body, allowing them to specifically target different cells, and to take on some of humanity's deadliest diseases, like cancer. This has spawned a great deal of nano research and experimental applications at labs all over the world. The nanosensors, on the other hand, may be surgically inserted into the brain of the paraplegic patient to give them the ability to move and walk again. Research indicates that it will appear on the farthest extent of techniques for repairing living cells, as well as nano-neural electronic links, and if this happens, a real revolution will occur in the world of treatment and therapy. There are a lot of applications in the field of health care and the manufacture of nanomedical devices. Using this method, it is possible to capture images of the body's cells with ease, akin to taking a conventional snapshot. Furthermore, these cells can be manipulated and moulded into various configurations [13], [14], [22]. An institute in California has set a general framework for what nanotechnology can offer us in the field of medicine, for those people who suffer from certain diseases, and for older people who suffer because of the incorrect sequence of atoms.

2.2.1. Delivery of the Drug to the Tissues. Delivery of drugs to tissues is one of the priorities of research in the field of nanomedicine, as it depends on the manufacture of micro-nanomaterials that improve the bioavailability of the drug. This means that the drug molecules are located in the targeted place in the body, where they work with maximum effectiveness, and thus the rate of drug consumption decreases and its side effects are reduced, as well as the total cost of treatment. Pharmacology is one of the sciences that needs high accuracy due to its connection to human health. A drug's efficacy is diminished and it causes undesired side effects if it is absorbed by healthy tissues along with the sick ones. For instance, we see that conventional approaches such as radiation and chemotherapy have serious adverse effects and are not very successful in treating cancer [11], [21], [44], [45]. As a result, anti-cancer medications need highly targeted administration in order to have any effect at all. Therefore, scientists are studying one of the future nano applications, which is represented in the drug delivery technology using one of the nano devices called dendrimer. Methods of drug delivery are based on nanotechnology, some of which depend on very small-scale tubes that have the ability to move and can be directed to the area to be treated. Others rely on smart systems of very small size that can be implanted inside the body and have the ability to control drug doses and the appropriate time for delivery. It appeared that carbon nanotubes could be used by linking them with peptide compounds to introduce them to the immune system in the body and thus use them in the delivery of traditional vaccines [43].

2.2.2. Pharmaceuticals and Therapeutic Drugs. A new term has now been introduced into the science of medicine, nanobiotics, which is the new alternative to antibiotics. Hangbang University researchers in Seoul successfully incorporated nano-silver into antibiotics. Silver possesses the capability to eradicate 650 pathogenic pathogens while maintaining the safety of the human body. This technology will solve a lot [5], [31], [46], [47], [48]. One of the problems of antibiotic-resistant bacteria that have caused mutations that prevent the effect of the antibiotic on these bacteria is that nanobiotics puncture the cell wall of bacteria or cells infected with the virus, allowing water to enter the cells and they are exterminated. This technique will eliminate the strains of bacteria that are resistant to antibiotics that have caused mutations to prevent them from affecting them. Where the nanobiotic punctures the cell wall of the bacteria or virus, and when millions of them enter the gel membrane of the bacterium, they are chemically attracted to each other and gather in the form of long tubes or many pins that puncture the cell membrane and other groups work to expand the hole in the bacterial cell wall so that it dies within minutes as a result. They dissipate the external electrical potential of her membrane and then destroy it within minutes, and she cannot adapt her immune system with it [4,44].

2.2.3. A Nano-Sized Robot is an Assistant in Surgical Operations. For use in delicate and high-risk procedures, Corv is has developed nanoport optical transducers with nanoscale scales. By using a specialized gadget, the surgeon may direct the robot to do the procedure with more precision and less human error than is possible with conventional approaches. The surgeon manipulates the robot arm, which holds the instruments and camera, via a joystick. Through this, huge motions may be reduced to micromanipulations, enhancing the

accuracy of surgical procedures. Because of its extreme hardness, scientists used carbon to create a nanopore only 1 millimeter wide, allowing it to pass through human blood arteries. Magnetic resonance imaging (MRI) and computed tomography (CT) scans may be used to track the robot while it works within the body to verify that it has reached the correct organ or sick tissue.

2.2.4. Diabetes Treatment. Niue University in the United States has successfully created a nanotechnology-based device that can regulate blood sugar levels in the body. This device offers an alternative to insulin injections for diabetic patients. Additionally, nanotechnology is being used in the treatment of kidney diseases. Specifically, researchers are studying the atomic-level formation of kidney proteins and using nanotechnology for imaging purposes. The aim was to study the biological processes that occur in kidney cells and the use of nanoparticles in the treatment of kidney diseases, where solutions to many kidney diseases can be reached by understanding the physical and chemical properties of kidney proteins, and many doctors dream of an artificial kidney using nanotechnology. The cell has the potential to improve the lives of many kidney patients significantly.

2.2.5. Medical Imaging of Medical Applications. Researchers and medical professionals now have the ability, thanks to nanoimaging, to monitor every movement that takes place inside the live tissue of the human body. This allows medical professionals to properly detect the movement of the medicine within the sick tissue. Studying some cells of the body is difficult, and from here scientists resort to coloring them. There is another problem, which is that the cells that emit light waves of different lengths do not always work in the same way or in the same way, which makes medical imaging processes face problems in terms of correct diagnosis. Scientists were able to solve this problem by using some nanoparticles that show reactions different due to the different wave frequencies arising naturally from the difference in the length of the face. Nanotechnology will contribute to advancing its development in terms of its efficiency, performance, speed of work, and increased safety, due to the entry of nanotechnology into the manufacture of electronic chips, electrical conduction circuits, and data processors used in these devices.

2.2.6. Diagnostic Applications of Nanotechnology Medical. The main goal is to discover the disease in the early stages so that it can be eliminated before it causes symptoms or complications. By using nanotechnology, biological tests to measure the presence or activity of the tested materials become faster, more accurate and flexible. The presence of specific molecules or microbes can be combined, and similarly, gold nanoparticles can be combined with short sections of DNA to identify a sequence of genes in a sample. There is also the technology of nano-holes to analyze DNA, which converts its sequence of units directly into electrical signals, and by using nanoparticles as contrast agents (as an alternative to dye), we obtain MRI and ultrasound images with better contrast and distribution, and even luminous nanoparticles can help the surgeon during the operation Surgical procedures to identify the location of the tumor and thus make the process of eradicating it more easily [4], [6], [17], [20], [25], [29], [30], [38], [45], [48], [49], [50].

3. Result and Discussion. The electrical characteristics of carbon nanotubes (CNTs) are determined by the chiral vector (C_h), which is calculated using Equation (1). Various arrangements of carbon nanotubes, such as (17, 0), (12, 5), (10, 10), (16, 2), (15, 0), and (11, 11), exhibit different chiral angles and integer pairs. For instance, the (17, 0) and (15, 0) nanotubes, known as zigzag nanotubes, can function as either metals or semiconductors, depending on the fabrication method. The focus of this paper is on chiral semiconducting nanotubes, specifically (12, 5) and (16, 2), with $n = 12, m = 5$, and chiral angles analyzed in the range of 0 to 30 degrees. Additionally, armchair nanotubes (10, 10) and (11, 11), with $n=m=10$ and a chiral angle ($=30$), operate as metals. Simulation results reveal nanotube diameters ranging from approximately 1.33 nm to 1.356 nm. The band gap in semiconducting nanotubes is determined by solving for the energy difference between the Fermi energy and electronic density of states (DOS). Equation 3.1 is employed to find the band gap (E_g) in semiconducting nanotubes by used MATLAB program. This is found by comparing the Fermi energy per carbon nanotube atomic unit cell to the electronic density of states (DOS) (unit measurement, arbitrary, unit) given by equation 3.1

$$E_g = \frac{2 * a(c - c) * \gamma_o}{d_t} = \frac{(0.8eV)}{d_t} \quad (3.1)$$

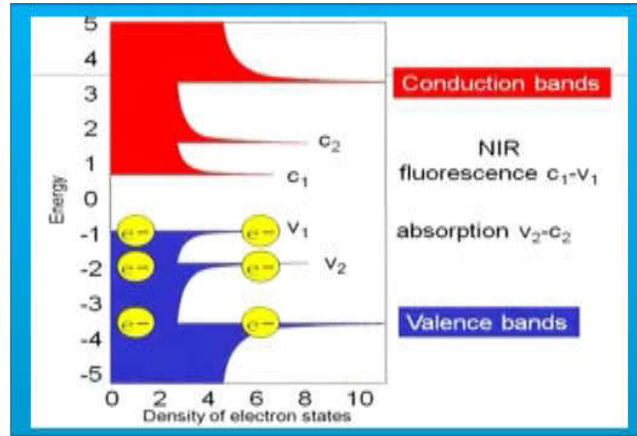


Fig. 3.1: An example of two van hove singularities in semiconductor carbon nanotubes

Figure 3.1 shows the energy band gap, denoted as E_g , for a semiconductor with two conduction bands and two valence bands, where the Fermi energy is $E_f = 2.9 eV$, and where two van Hove singularities are handled. The semiconducting properties of carbon nanotubes are described by the equation (2/3) if the condition that $2n+m = 3q$ is equation (3.13). The minimal value, which is provided by the equation, is used to compute the kinetic energy, denoted by E_k , of the lowest subband 3.2.

$$K(c,q) = \frac{2}{(3d_t)} \tag{3.2}$$

The equations that describe the energy output of carbon nanotubes, where d_t represents the diameter of the nanotube, are as follows: 3.3

$$(K_t) = \left(\frac{3 * a(c - c) * \gamma_o}{2} \right) * \sqrt{(k_t^2 + (2(3d_t)^2)} \tag{3.3}$$

and

$$E_o = \frac{(3 * a(c - c) * \gamma_o)}{2} \tag{3.4}$$

where E_o is the energy gap for graphite when it is at its standard temperature. After solving equations 3.3 and 3.4, we are left with equation 3.5, which describes E_{out} , also known as the output energy band gap.

$$E_{out} = \frac{(E(k_t))}{E_o} \tag{3.5}$$

Doping the electron density of states during manufacture has a direct impact on the output energy band gap of carbon nanotubes, which may be described as an equation 3.6.

$$D(E_k) = \frac{8}{(3 * \pi * a(c - c) * \gamma_o)} * \frac{(E(k_t))}{\sqrt{(E(k_t) - (E_g/2)} \tag{3.6}$$

The electrons in the valence band (v_2) are excited by the incident light (photon) and move into the conduction band (c_2), leaving a hole in the valence band (v_2), which allows an electron to move from the valence band (v_1) to the valence band (v_2), which in turn allows an electron to move from c_2 to the conduction band (c_1) and back to v_1 such as shown in figure 3.1.

Certainly, let's delve deeper into specific aspects of the electronic construction of carbon nanotubes (CNTs) and their implications:

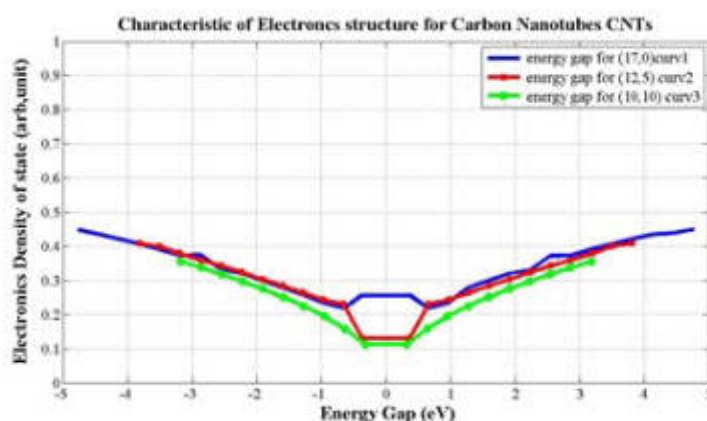


Fig. 3.2: The energy gap to the DOS in this system in order to present the results for three distinct carbon nanotubes (17, 0), (12, 5), and (10, 10)

- *Chirality and Electronic Properties:* The chirality angle (θ) determined by the chiral vector plays a crucial role. Zigzag nanotubes ($=0$) exhibit unique properties, potentially serving as either metals or semiconductors. Semiconducting nanotubes with specific chiral angles (e.g., (12, 5), (16, 2)) offer controlled electronic behavior.
- *Simulation and Diameter Impact:* The simulation results showing nanotube diameters are vital. Diameters influence the electronic structure, affecting band gaps and conductivity. Understanding this parameter aids in tailoring nanotubes for specific applications.
- *Energy Band Gap in Semiconducting Nanotubes:* Equation 3.1 provides a direct link between the energy band gap and nanotube diameter. This relationship is essential for predicting and controlling the semiconducting properties crucial for electronic applications.
- *Visualization of Energy Band Gap (E_g):* Figure 3.1's visualization of the energy band gap is a key element. It illustrates the distinct bands and van Hove singularities, giving a comprehensive view of the electronic structure. This understanding is fundamental for designing nanotubes for specific electronic functionalities.
- *Output Energy Band Gap (E_{out}):* Equation (20) and the concept of E_{out} are critical. This parameter encapsulates the nanotube's output energy band gap, offering insights into its behavior and suitability for various applications. It aids in predicting the nanotube's response to external stimuli.
- *Impact of Doping on Electronic States:* Equation 3.6 underscores the impact of doping on the electron density of states. Doping introduces additional electronic states, influencing the output energy band gap. This insight is valuable for engineering nanotubes with tailored electronic properties.
- *Applications in Electronics:* The discussed parameters collectively contribute to the understanding of how carbon nanotubes can be harnessed in electronic devices. Their unique electronic properties, such as high conductivity and tunable band gaps, make them promising candidates for future electronic applications.
- *Challenges and Future Directions:* Despite their potential, challenges and limitations in utilizing carbon nanotubes in electronics should be acknowledged. Issues like uniformity, scalability, and reproducibility are areas of ongoing research. Future directions may involve overcoming these challenges for widespread implementation.

The electronic construction of carbon nanotubes, shaped by factors like chirality and diameter, offers a versatile platform for tailoring their behavior in electronic applications. The ability to control their electronic properties makes them valuable in the development of advanced electronic devices.

Figure 3.2 illustrates that carbon nanotubes can exhibit both positive and negative values for the energy band gap. This high level of symmetry is a unique characteristic, indicating the versatile electronic behavior

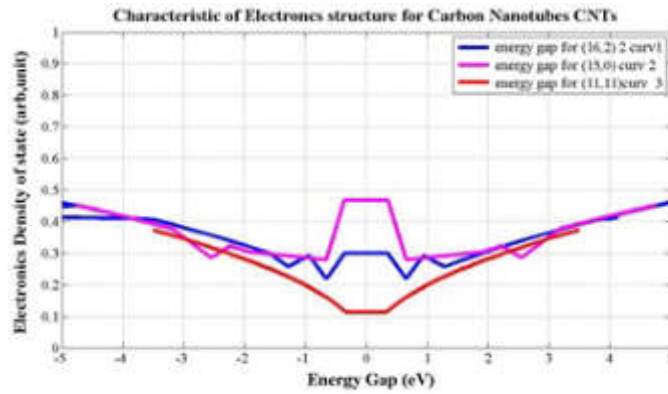


Fig. 3.3: Connection between energy gap and DOS for three distinct carbon nanotubes (16, 2), (15, 0), and (11, 11)

of nanotubes. Coordinates (17, 0) and (12, 5) in Figure 3.2 demonstrate the energy gap with two valence bands (v_1 and v_2) and two conduction bands (c_1 and c_2) according to van Hove notation. The presence of both positive and negative energy gap values highlights the tunable electronic states in these nanotubes. Unlike semiconductor nanotubes, (10, 10) nanotubes exhibit metallic behavior with only one conduction band and a valence band. The absence of a band gap in these nanotubes classifies them as metals. This behavior is critical for applications requiring high electrical conductivity.

Employing equations 3.2-3.5, another set of findings is presented in Figure 3.3, reinforcing the observation that carbon nanotubes can have both positive and negative energy gap values. This emphasizes the exceptionally high symmetry of carbon nanotubes' electronic structure. In Figure 3.3, coordinates (15, 0) and (16, 2) exhibit two conduction van Hove bands (c_1 and c_2) and two valence bands (v_1 and v_2). The symmetry of the energy gap in semiconductor carbon nanotubes is influenced by doping in the electronic density of states (DOS), resulting in either positive or negative energy gaps. For (11, 11) nanotubes, the band gap is zero, indicating metallic behavior. These nanotubes act as metallic carbon nanotubes with a continuous range of energy levels. Equations 3.7 and 3.8 provide insights into the behavior of carbon nanotubes as semiconductors (S) and metals (M) during a symmetric transition at $p=q$. The ability of nanotubes to operate in positive and negative bias with low bias is a crucial characteristic for certain applications. The equations suggest that the symmetric transition in the energy level occurs at $p=q$, allowing carbon nanotubes to operate under both positive and negative biases with low bias conditions (160 mV - 200 mV). This bias-dependent behavior is significant for practical electronic applications.

$$E_p p^S = \frac{(2 * \pi * a(c - c) * \gamma_o)}{d_t} \tag{3.7}$$

For semiconductance carbon nanotube,

$$E_p p^M = \frac{(6 * \pi * a(c - c) * \gamma_o)}{d_t} (6 * p * a(c - c) * o) / d_t \tag{3.8}$$

for metallic carbon nanotube.

Figure 3.2 and 3.3 and associated equations underscores the intricate electronic properties of carbon nanotubes, including their symmetry, tunability, and diverse behavior as semiconductors and metals under different conditions. These characteristics are fundamental for leveraging carbon nanotubes in a wide range of electronic applications. The synergy between nanotechnology and artificial intelligence (AI) represents a transformative leap in technology. Nanotechnology, with its ability to manipulate materials at the nanoscale, coupled with AI's data processing and analytical capabilities, promises enhanced precision and efficiency across various domains.

In the realm of medical devices and healthcare, the convergence of nanotechnology and AI holds great promise. The precision offered by nanoscale materials, such as carbon nanotubes, can be complemented by AI algorithms for more accurate diagnostics, treatment delivery, and monitoring of health conditions. Carbon nanotubes, with their remarkable electrical, thermal, and mechanical properties, stand out as a key component of nanotechnology. Their high conductivity, strength, and unique structure make them versatile for applications ranging from electronics to biomedical devices. The discussion emphasizes the significant role carbon nanotubes will play in the future of nanotechnology. In the biomedical field, these nanotubes have shown promise for drug delivery, imaging, and diagnostics due to their ability to penetrate cell membranes and interact at the molecular level. Carbon nanotubes, integrated with AI, can revolutionize healthcare by enabling advanced diagnostics, personalized medicine, and real-time monitoring. The combination of nanoscale materials and intelligent algorithms enhances the capabilities of medical devices for more accurate and timely interventions. The collaborative use of nanotechnology and AI in medical diagnostics can lead to highly accurate and efficient diagnosis. Nanoscale sensors, possibly incorporating carbon nanotubes, can detect biomarkers at ultra-low concentrations, while AI algorithms analyze complex datasets for disease identification. The statement underscores the belief that carbon nanotubes, as part of nanotechnology advancements, will significantly shape the future of healthcare. This suggests a transformative era where nanoscale materials, guided by AI, contribute to breakthroughs in diagnostics, treatment, and overall healthcare management. Alongside the potential benefits, it's crucial to acknowledge and address challenges and ethical considerations associated with the integration of nanotechnology and AI in healthcare. Ensuring the safety, privacy, and ethical use of these technologies is paramount for their widespread acceptance and positive impact. the convergence of nanotechnology, artificial intelligence, and the potential of carbon nanotubes holds immense promise for advancing healthcare. The precision, efficiency, and transformative capabilities offered by these technologies signify a future where medical devices are more accurate, personalized, and effective in improving patient outcomes. The difference is clear in Table No. 3.1 of the specifications and properties of carbon nanotube and its use, like semiconductors, silicon, and germanium, in terms of work, heat tolerance, and working with ultra-high frequencies up to THz. as well as form figure 3.2 and 3.3.

4. Conclusion. The development of nanoparticles, their size control, and the study of their physical properties have revolutionized medical diagnostics. The advancement has opened up possibilities for incorporating nanoparticles, specifically carbon nanotubes (CNTs), into well-established diagnostic techniques such as magnetic resonance imaging (MRI), ultrasound, CT scans, and nuclear medicine equipment. The ability to control nanoparticle properties enhances diagnostic efficiency, enabling early disease detection and providing detailed information about disease location, size, and progression. Around the world, ongoing studies focus on incorporating nanotechnology advancements into medical fields. These efforts include safety assessments for human use, aiming to turn these applications into a daily reality in hospitals. Nanotechnology applications span various sectors, including technology, electronics in medicine, biology, pharmaceutical industries, and disease detection. The enormous potential of CNTs in biological applications is evident. CNTs are highly adaptable molecules that show promise in diverse contexts, acting as sensors, drug transporters, imaging aids, bioelectrodes, and reinforcement for composites. CNT-based sensors are envisioned as simple, rapid, sensitive, and cost-effective tools for monitoring various analyses. Their design flexibility allows for tailoring to specific needs, surpassing the limitations of prior analytical methods. The construction of CNT-based sensors from scratch enhances simplicity and performance. These sensors are considered more straightforward to work with, exhibiting improved detection limits, sensitivities, specificities, and repeatabilities. The potential of CNT-based sensors as effective tools for monitoring targets is recognized, offering a feasible option to meet the urgent demand for numerous analyses. Future research on CNTs-based biosensing is expected to emphasize *in vivo* detection methods. These methods aim for minimal cytotoxicity, high sensitivity, and long-term stability to meet the requirements for reliable point-of-care diagnostics under physiological conditions. The continuous development of CNT-based technologies underscores their significance as a transformative tool in healthcare, contributing to disease treatment and human health preservation. While silicon remains the workhorse of the semiconductor industry, carbon nanotubes offer unique properties that make them attractive for certain applications, particularly in emerging fields such as flexible electronics and advanced sensors. Ongoing research aims to harness the strengths of both materials for future semiconductor technologies. Table No. 1 in the specs outlines the

Table 3.1: Comparison Carbon nanotubes (CNTs) and silicon are both materials with distinct semiconductor properties.

Serial No.	Criteria	Carbon Nanotubes (CNTs) Semiconductors	Silicon Semiconductor:
1.	Structure	CNTs have a tubular structure composed of carbon atoms arranged in a hexagonal lattice.	Silicon is a traditional semiconductor with a crystalline structure.
		They can be single-walled (SWNT) or multi-walled (MWNT), and their electrical properties depend on their structure and chirality.	It is widely used in the electronics industry, forming the basis of most semiconductors
2.	Electrical Properties	CNTs can exhibit either metallic or semiconducting behavior depending on their structure.	Silicon is typically a semiconductor with an indirect bandgap.
		Semiconducting CNTs have a bandgap that varies with their diameter and chirality.	Its electrical properties can be manipulated through doping.
3.	Advantages	High electrical conductivity, comparable to or even better than silicon.	Well-established technology with mature fabrication processes
		Exceptional mechanical strength, flexibility, and thermal conductivity.	Integrated into a variety of electronic devices and circuits.
4.	Applications	Widely explored for Nano electronics, including transistors and interconnects.	Dominant material in the semiconductor industry for integrated circuits and microelectronics.
		Promising in applications like flexible electronics and high-performance sensors.	Commonly used in transistors, diodes, and solar cells.
5	Bandgap Control	CNTs offer a tunable bandgap based on their structure.	Silicon bandgap is controlled through doping
6	Mechanical Properties	CNTs have superior mechanical properties	Silicon, providing flexibility and strength.
7.	Conductivity	CNTs can have higher electrical conductivity than silicon, making them suitable for specific high-performance applications.	Silicon has modest conductivity and is affected by temperature changes
8.	Fabrication Complexity	Large-scale production of CNTs is still an evolving challenge.	Silicon has well-established and mature fabrication processes

distinctions between carbon nanotubes and their applications in semiconductors such as silicon and germanium, focusing on factors including performance, heat resistance, and operation at ultra-high frequencies up to THz, as well as the result of this work paper.

REFERENCES

- [1] N. For and B. Applications, Jurnal Teknologi RECENT MODIFICATIONS OF CARBON, vol. 2, pp. 83100, 2023.
- [2] S. Kruss, A. J. Hilmer, J. Zhang, N. F. Reuel, B. Mu, and M. S. Strano, Carbon nanotubes as optical biomedical sensors, Adv. Drug Deliv. Rev., vol. 65, no. 15, pp. 19331950, 2013, doi: 10.1016/j.addr.2013.07.015.
- [3] V. Harish, D. Tewari, M. Gaur, A. B. Yadav, and S. Swaroop, Review on Nanoparticles and Nanostructured Materials: Bioimaging , Biosensing , Drug Delivery , Tissue Engineering , Antimicrobial , and Agro-Food Applications, 2022.
- [4] A. Barhoum et al., Review on Natural , Incidental , Bioinspired , and Engineered Nanomaterials: History , Definitions , Classifications , Synthesis , Properties , Market , Toxicities , Risks , and Regulations, 2022.
- [5] D. Maiti, X. Tong, X. Mou, K. Yang, and K. Yang, Carbon-Based Nanomaterials for Biomedical Applications: A Recent Study, vol. 9, no. March, pp. 116, 2019, doi: 10.3389/fphar.2018.01401.
- [6] D. Holmannova, P. Borsky, T. Svadlakova, and L. Borska, applied sciences Carbon Nanoparticles and Their Biomedical Applications, pp. 121, 2022.

- [7] A. Madni et al., Graphene-based nanocomposites: synthesis and their theranostic applications, *J. Drug Target.*, vol. 0, no. 0, pp. 126, 2018, doi: 10.1080/1061186X.2018.1437920.
- [8] S. Anwar et al., Recent Advances in Synthesis , Optical Properties , and Biomedical Applications of Carbon Dots, *ACS Appl. Bio Mater.*, vol. 2, pp. 23172338, 2019, doi: 10.1021/acsabm.9b00112.
- [9] E. T. Thostenson, Z. Ren, and T. Chou, Advances in the science and technology of carbon nanotubes and their composites: a review, vol. 61, pp. 18991912, 2001.
- [10] H. A. Owida, N. M. Turab, and J. Al-nabulsi, Carbon nanomaterials advancements for biomedical applications, vol. 12, no. 2, pp. 891901, 2023, doi: 10.11591/eei.v12i2.4310.
- [11] E. Ali et al., Carbon nanotubes: properties, synthesis, purification, and medical applications, *Nanoscale Res. Lett.*, vol. 9, no. 1, p. 393, 2014.
- [12] B. O. Murjani, P. S. Kadu, M. Bansod, S. S. Vaidya, and M. D. Yadav, Carbon nanotubes in biomedical applications: current status , promises , and challenges, *Carbon Lett.*, vol. 32, no. 5, pp. 12071226, 2022, doi: 10.1007/s42823-022-00364-4.
- [13] V. R. Rapphey, T. K. Henna, K. P. Nivitha, P. Mufeedha, C. Sabu, and K. Pramod, Advanced biomedical applications of carbon nanotube, *Mater. Sci. Eng. C*, vol. 100, no. July 2018, pp. 616630, 2019, doi: 10.1016/j.msec.2019.03.043.
- [14] F. Liang and B. Chen, A Review on Biomedical Applications of Single-Walled Carbon Nanotubes, *Curr. Med. Chem.*, vol. 17, no. 1, pp. 1024, 2009, doi: 10.2174/092986710789957742.
- [15] E. Vázquez and M. Prato, Carbon nanotubes and microwaves: Interactions, responses, and applications, *ACS Nano*, vol. 3, no. 12, pp. 38193824, 2009, doi: 10.1021/nn901604j.
- [16] J. Nandhini, E. Karthikeyan, and S. Rajeshkumar, Nanomaterials for wound healing: Current status and futuristic frontier, *Biomed. Technol.*, vol. 6, pp. 2645, 2024, doi: <https://doi.org/10.1016/j.bmt.2023.10.001>.
- [17] S. M. Asil et al., Theranostic applications of multifunctional carbon nanomaterials, no. January, pp. 124, 2023, doi: 10.1002/VIW.20220056.
- [18] R. S. Ruoff, D. C. Lorents, S. R. I. International, and M. Park, MECHANICAL AND THERMAL PROPERTIES OF CARBON NANOTUBES, vol. 33, no. 7, pp. 925930, 1995.
- [19] S. Keren, C. Zavaleta, Z. Cheng, A. De Zerda, O. Gheysens, and S. S. Gambhir, Noninvasive molecular imaging of small living subjects using Raman spectroscopy, 2008, doi: 10.1073/pnas.0710575105.
- [20] S. Gautam, D. Bhatnagar, D. Bansal, H. Batra, and N. Goyal, Recent advancements in nanomaterials for biomedical implants, *Biomed. Eng. Adv.*, vol. 3, p. 100029, 2022, doi: <https://doi.org/10.1016/j.bea.2022.100029>.
- [21] S. Beg, M. Rizwan, A. M. Sheikh, M. S. Hasnain, K. Anwer, and K. Kohli, Advancement in carbon nanotubes: Basics, biomedical applications and toxicity, *J. Pharm. Pharmacol.*, vol. 63, no. 2, pp. 141163, 2011, doi: 10.1111/j.2042-7158.2010.01167.x.
- [22] W. Yang, P. Thordarson, J. J. Gooding, S. P. Ringer, and F. Braet, Carbon nanotubes for biological and biomedical applications, *Nanotechnology*, vol. 18, no. 41, 2007, doi: 10.1088/0957-4484/18/41/412001.
- [23] J. Pang, A. Bachmatiuk, F. Yang, H. Liu, and W. Zhou, Applications of Carbon Nanotubes in the Internet of Things Era Carbon nanotubes based electronics, *Nano-Micro Lett.*, vol. 13, no. 1, pp. 115, 2021, doi: 10.1007/s40820-021-00721-4.
- [24] M. Asaftei et al., RSC Advances Fighting bacterial pathogens with carbon nanotubes: focused review of recent progress, pp. 1968219694, 2023, doi: 10.1039/d3ra01745a.
- [25] M. Ramezani, A. Dehghani, and M. M. Sherif, Carbon nanotube reinforced cementitious composites: A comprehensive review, *Constr. Build. Mater.*, vol. 315, p. 125100, 2022, doi: <https://doi.org/10.1016/j.conbuildmat.2021.125100>.
- [26] H. Cui, S. Zhao, and G. Hong, Wireless deep-brain neuromodulation using photovoltaics in the second near-infrared spectrum, *Device*, vol. 1, no. 4, p. 100113, 2023, doi: <https://doi.org/10.1016/j.device.2023.100113>.
- [27] D. S. G and M. B, A comprehensive review on current trends in greener and sustainable synthesis of ferrite nanoparticles and their promising applications, *Results Eng.*, vol. 21, p. 101702, 2024, doi: <https://doi.org/10.1016/j.rineng.2023.101702>.
- [28] Q. Dots, M. Bechelany, and A. Barhoum, Biomedical Applications of Carbon Nanomaterials: Fullerenes , 2021.
- [29] M. Ul-Islam, K. F. Alabbosh, S. Manan, S. Khan, F. Ahmad, and M. W. Ullah, Chitosan-based nanostructured biomaterials: Synthesis, properties, and biomedical applications, *Adv. Ind. Eng. Polym. Res.*, 2023, doi: <https://doi.org/10.1016/j.aiepr.2023.07.002>.
- [30] G. Rahman et al., An Overview of the Recent Progress in the Synthesis and Applications of Carbon Nanotubes, *C*, vol. 5, no. 1, p. 3, 2019, doi: 10.3390/c5010003.
- [31] M. R. Almeida et al., Carbon Nanotubes for Biomedical Applications, *Mater. Horizons From Nat. to Nanomater.*, vol. 4, no. 2, pp. 285331, 2022, doi: 10.1007/978-981-16-7483-9_14.
- [32] M. F. L. De Volder, S. H. Tawfick, R. H. Baughman, and A. J. Hart, Carbon nanotubes: Present and future commercial applications, *Science (80-.)*, vol. 339, no. 6119, pp. 535539, 2013, doi: 10.1126/science.1222453.
- [33] N. Gupta, S. M. Gupta, and S. K. Sharma, Carbon nanotubes: synthesis, properties and engineering applications, *Carbon Lett.*, vol. 29, no. 5, pp. 419447, 2019, doi: 10.1007/s42823-019-00068-2.
- [34] J. J. Shelke, A. R. , Roscoe, J. A. , Morrow, G. R. , Colman, L. K. , Banerjee, T. K. , & Kirshner and Perrine Susan, NIH Public Access, *Bone*, vol. 23, no. 1, pp. 17, 2008, doi: 10.1166/jbn.2005.004.Applications.
- [35] F. Kreupl, A. P. Graham, M. Liebau, G. S. Duesberg, R. Seidel, and E. Unger, Carbon nanotubes for interconnect applications, *Tech. Dig. - Int. Electron Devices Meet. IEDM*, pp. 683686, 2004, doi: 10.1109/iedm.2004.1419261.
- [36] N. Mehra, S. Pharma, A. Jain, R. Raj, D. Mishra, and S. Maharastra, *Th Au Se*, vol. 25, no. February, pp. 169206, 2008.
- [37] G. S. S. Clarence S Yah, The use of Carbon Nanotubes in Medical Applications - Is It a Success Story?, *Occup. Med. Heal. Aff.*, vol. 02, no. 01, pp. 1011, 2014, doi: 10.4172/2329-6879.1000147.
- [38] S. Mohanty and A. Misra, Sensors and Actuators B: Chemical Carbon nanotube based multifunctional flame sensor, *Sensors Actuators B. Chem.*, vol. 192, pp. 594600, 2014, Available: <http://dx.doi.org/10.1016/j.snb.2013.11.019>
- [39] S. Polizu, O. Savadogo, P. Poulin, and L. Yahia, Applications of carbon nanotubes-based biomaterials in biomedical nan-

- otechnology, *J. Nanosci. Nanotechnol.*, vol. 6, no. 7, pp. 18831904, 2006, doi: 10.1166/jnn.2006.197.
- [40] M. Roldo and D. G. Fatouros, Biomedical applications of carbon nanotubes, 2013, doi: 10.1039/c3pc90010j.
- [41] F. Karchoubi, R. Afshar Ghotli, H. Pahlevani, and M. Baghban Salehi, New insights into nanocomposite hydrogels; a review on recent advances in characteristics and applications, *Adv. Ind. Eng. Polym. Res.*, 2023, doi: <https://doi.org/10.1016/j.aiepr.2023.06.002>.
- [42] N. H. Solangi, R. R. Karri, N. M. Mubarak, and S. A. Mazari, Mechanism of polymer composite-based nanomaterial for biomedical applications, *Adv. Ind. Eng. Polym. Res.*, 2023, doi: <https://doi.org/10.1016/j.aiepr.2023.09.002>.
- [43] M. Bilal et al., Surface-coated magnetic nanostructured materials for robust bio-catalysis and biomedical applications-A review, *J. Adv. Res.*, vol. 38, pp. 157177, 2022, doi: <https://doi.org/10.1016/j.jare.2021.09.013>.
- [44] G. Gruner, Carbon nanotube transistors for biosensing applications, *Anal. Bioanal. Chem.*, vol. 384, no. 2, pp. 322335, 2006, doi: 10.1007/s00216-005-3400-4.
- [45] A. Mazzaglia and A. Piperno, Carbon Nanomaterials for Therapy , Diagnosis , and Biosensing.
- [46] Y. Charles, Nanoparticles with Raman Spectroscopic Fingerprints for DNA and RNA Detection, vol. 1536, no. 2002, 2012, doi: 10.1126/science.297.5586.1536.
- [47] S. Law, Mini-Review for an Electrocatalytic Application of Carbon Nanotube in Medical Fields Tissue Engineering , Drug Delivery , Cancer and SARS-CoV-2, vol. 13, no. 1, pp. 18, 2023.
- [48] B. K. Saikia, S. Maria, M. Bora, J. Tamuly, and M. Pandey, Review article A brief review on supercapacitor energy storage devices and utilization of natural carbon resources as their electrode materials, *Fuel*, vol. 282, no. April, p. 118796, 2020, doi: 10.1016/j.fuel.2020.118796.
- [49] A. Bianco, K. Kostarelos, D. Partidos, and M. Prato, Biomedical applications of functionalised carbon nanotubes, no. November 2004, pp. 571577, 2005, doi: 10.1039/b410943k.
- [50] S. Akgönüllü and A. Denizli, Recent advances in optical biosensing approaches for biomarkers detection, *Biosens. Bioelectron. X*, vol. 12, p. 100269, 2022, doi: <https://doi.org/10.1016/j.biosx.2022.100269>.

Edited by: Mustafa M Matalgah

Synergies of Neural Networks, Neurorobotics, and Brain-Computer Interface Technology:

Special issue on: Advancements and Applications

Received: Dec 16, 2023

Accepted: Feb 17, 2024



DRIVER DROWSINESS DETECTION

ANN ZEKI ABLAHD*, ALYAA QUSAY ALORAIBI† AND SUHAIR ABD DAWWOD‡

Abstract. The state of the driver of being extremely tired or sleepy through the operation of the vehicle is called driver drowsiness. Different factors caused this state such as alcohol, lack of sleep, and the side effect of some medication. The drowsiness of drivers is a serious safety lead to accidents or fatalities on external and internal roads. The increased number of road accidents resulted from drowsy driving. A special smart, reliable, and accurate system, Using Python language 3.6 for Windows, was designed to build an alert system for drivers in detecting drowsiness driver. This system is crucial in reducing accidents road by the ability to concentrate, react quickly, and produce sound decisions through driving. This system implements a real-time detector that can monitor the states of drivers through driving.

Smart cameras with 16-megapixel were used to ensure that capturing photos have a high quality. These cameras were used in gathering the driver's dataset in different alertness states, including both alert states and drowsy. The collected dataset is processed by extracting all relevant features such as head movement, yawning, and eye closure, which were used in identifying the driver's drowsiness. Python's libraries such as TensorFlow, OpenCV, Keras, and Pygame are used for extracting all the above features. Viola-Jones algorithm is used in face eye region detecting and extracting from the image of the face in the proposed system. A Support Vector Machine (SVM) algorithm was used in classifying between drowsy and non-drowsy drivers. The system is tested and evaluated in the real world, to ensure that the system is reliable and robust; it has high performance and accuracy, and the accuracy is about 99.1%. This system can be used in manufacturing vehicles.

Key words: sensors, driver, drowsiness, driving, accident, smart camera

1. Introduction and Preliminaries. Driver Drowsiness considered a significant contributor to road accidents worldwide. To enhance road safety a special detection system was developed to detect driver drowsiness. Such a system becomes very crucial in enhancing road safety.

The proposed system can monitor the physiological signs and behaviors of drivers to detect drowsiness and alert the driver early before the accident occurs. Various techniques have been used in a proposed system including facial expression analysis, The proposed system can be used in eye tracking, and Viola-Jones algorithm; future in the car industry to reduce road accidents resulting from drowsy drivers and keep a safe driver.

2. Literature Review. The research on driver drowsiness contributed significantly to improving the safety of external and internal roads to increase safety and reduce or prevent all drowsy driving risks. These researchers try to identify the warning signs that cause driver drowsiness. This information is used in developing a different technology for detecting and preventing driver drowsiness such as physiological sensors, steering behavior analysis, and eye-tracking systems. Such technologies alert the drowsy driver through becoming incapacitated, to take control vehicle. The benefit of this is to reduce accidents number and save people's lives. With the increasing of road accidents number caused by drowsy driving has become a driver drowsiness detection system very important field in transportation safety. Through the availability of digital cameras, it has been increasing the amount of archived recorded videos around the world and it became an effectively growing processing of these video data. Different studies have been conducted to develop and evaluate several techniques in detecting driver drowsiness.

The most popular technique is using electromyogram (EMG) [1], electroencephalogram (EEG), and electrocardiogram (ECG) signals in monitoring the driver's muscle activity, brainwaves, and heart rate respectively. Dong and colleagues (2015) introduced a study that used EEG signals in detecting driver [2] drossiness, with

*Technical College Kirkuk, Northern Technical University (drann@ntu.edu.iq)

†University of Mosul, College of Computer Science and Mathematics, Software Department (dr.alyaa@uomosul.edu.iq)

‡University of Mosul, College of Administration and Economics, Department Management Information Systems (suhair_abd_dawwod@uomosul.edu.iq)

an accuracy of 94%. Lee and colleagues (2017) introduce a study that used a combination of ECG, EEG signals in detecting driver drowsiness, with an accuracy of 98%. Chakraborty (2019) [3] used the data of eye tracking in the detection of driver's drowsiness with an accuracy of 85%. Wu and colleagues (2021) applied facial expression analysis in the detection of driver's drowsiness with an accuracy of 87%. Different machine language algorithms have been used in the detection of driver's drowsiness by using data pattern recognition and deciding the accurate predictions. Malik (2019) [4] applied a support vector machine (SVM) classifier in the detection of driver's drowsiness with an accuracy of 92%. Vishwakarma used a process called object tracking of the saved video in his survey [5]. M. Zhang [6] used extensive hardware sensors in detecting. But C. Anil used the generalizability of models in enhancing the detection of drowsy by increasing the dataset size [7]. M. Jafari introduced a new technology for enhancing the accuracy of system detection [8]. M. Aljasim suggested an expensive experiment detection method [9]. A. M. Leeuwenberg used more complex algorithms that used more variables and increased the tested samples [10]. The accuracy of the previous works is low in more time and most of these papers are surveys, that why it prepared a real-world practical system to protect people from the high numbers of accidents. The real-world, low-cost cost with larger data sets proposed system is reliable and robust and is used in different scenarios like Viola-Jones algorithm, face eye region detecting and extracting from the smart camera's images. After evaluation of the proposed system, the accuracy is 99.1%. This system is more practical and it is easy to use in manufacturing different vehicles.

3. Architecture of the Proposed System. The proposed system consists of different steps; Figure 1 shows the steps of the proposed system. Input data: Input picked videos and images as a dataset to identify the driver's physiological signals, such as eye movement, and facial expression. Face Detection and Extract eye region: The identification of the driver's physiological signals is preprocessed for extracting the relevant features of him(her), such as eye blinking which are used to detect the driver's drowsiness. The Viola-Jones algorithm is used to detect each object in videos or images. This step is very important in improving the accuracy of the proposed system. Classification: All the extracted features are fed to the SVM (Support Vector Machine) classifier. This binary classifier is a supervised learning algorithm used for pattern identification to distinguish between (Drowsy, and non-drowsy) drivers based on input features. A sound will be produced by the proposed system to alarm drivers as a notification to stop driving and prevent accidents.

4. Dataset and Preprocessing. After collecting different data set images, the data is extracted from videos by cutting the continuous streams into discrete frames to identify the objects (face, mouth, eye,..., etc). The OpenCV technique is used. OpenCV is a Python library (tool for image processing) with the deep learning algorithm Support Vector Machine. The dataset of the proposed system depends on sequences of videos and images captures by webcam, which is based on eye estimation motion.

The Viola-Jones algorithm was applied to detect the parts of the face and eyes [11], [12]. While the area of eye motions was estimated frame to frame by sparse track from optical flow. Different adaptive thresholding values were used to decide whether the eyes were closed or opened.

The data preprocessing is an essential step in the detection of driver drowsiness, and it helps to prepare and clean data for analysis. The preprocessing steps that applied to the proposed system are:

Collection of data: Collect all captured camera videos or images of drivers exhibiting different levels of fatigue or drowsiness, to detect the face and extract the eye region, facial expression, [13] head position from the face images. Figure 4.1 represents the preprocessing of data.

Cleaning of data: Remove all irrelevant noisy data points that affect the analysis accuracy [14], [15].

Normalization: Try to normalize the captured data according to a common scale to be easier to analyze and compare.

Features selecting and extracting: Selecting and extracting most of the relevant features is important, because, not all extracted features are useful in drowsiness detection. So the selection of the most important features can improve analysis accuracy. Examples of data extraction like determining the location of eyes, blinking of eyes (closed, opened), etc. The Python libraries (dlib, OpenCv) are used in performing feature extracting. From every captured video the landmarks eyes were detected by eye aspect ratio

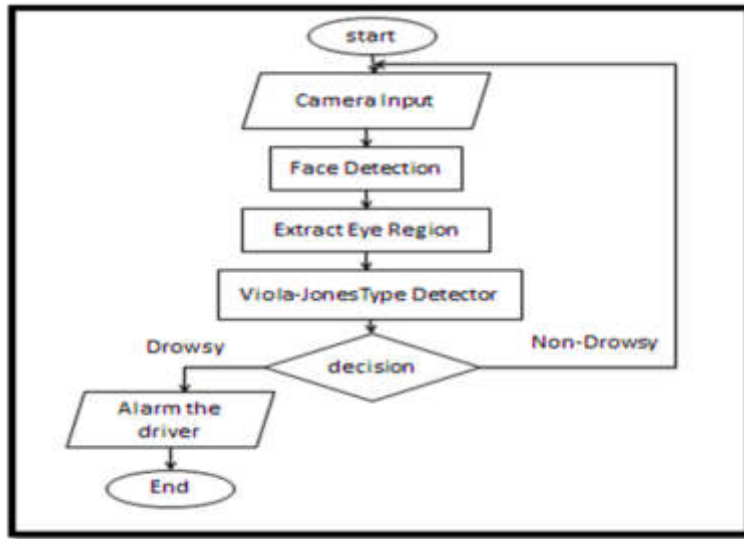


Fig. 4.1: Architecture of the Proposed System

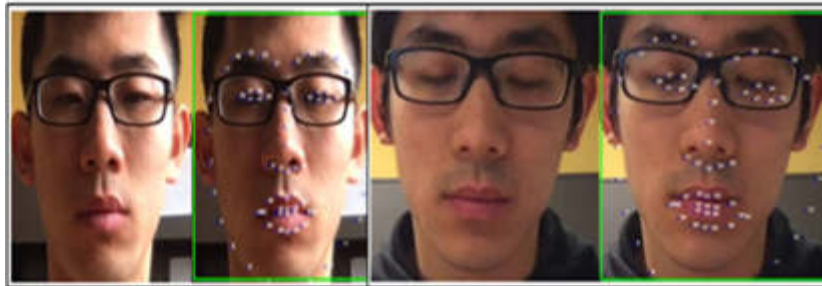


Fig. 4.2: The preprocessing of data for Face detection

(EAR) between the width and height of the eye will be computed by using equation 4.1.

$$EAR = \frac{||P2 - P6|| + ||P3 - P5||}{2||P1 - P4||} \tag{4.1}$$

where $p1, p2, p3, p4, p5, p6$ are the landmark location of two dimensions drawn in Figure 3 [9].

The EAR ratio is mostly constant when the eye is opened and EAR=0 when the eye is closed. EAR ratio for an open eye has a small variance, in plane rotation of the face, and is fixed to a uniform scaling for each image [16], [17]. The blinking of each eye is performed synchronously by both eyes, and the EAR of them is averaged.

Figure 4.2 represents an example of signals of EAR over the video sequence.

Classifying of data: Classifying the data into two cases drowsy or drowsy by using the Support Vector Machine (SVM) algorithm. The last step is creating special samples of data by applying small variations or adding noise to existing data. This step is done to improve the accuracy and performance of this proposed system [18].

Evaluate the proposed system: The evaluation is done in real words by testing the dataset using some metrics such as precision, accuracy, and recall. The Python library sklearn. metrics used in performing the evaluation. This system can detect drowsy drivers to alert them to stop driving and take a rest.

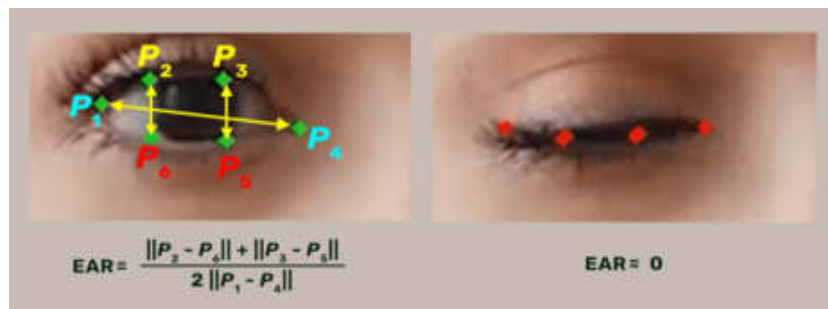


Fig. 4.3: Closed and open eyes with 2D landmark location

5. Viola-Jones algorithm. This is the most algorithms popular today in object detection for digital videos and images. This algorithm applies a series of image classifications [14], [15]. At first, this algorithm will convert each image into ("Haar-like features) a series of rectangular sub-images. This algorithm used Haar-like features, Figure 5.1 represents these features. These sub-images a binary images that have different highlights in brightness between adjacent pixels. This algorithm is used to determine and detect the features of the object. Like detection of eyes in the face image, which is very high real-time framework training [21], [22].

There are three ideas that used in this algorithm for face detection:

1. Integral Image: This idea represents an image where each pixel value represents the sum of all the above pixels and the left of it. This idea is very useful for quickly computing all sum values of pixels in any rectangle region of the image.
2. The classifier AdaBoost: This is an algorithm used in classification tasks. It is work combining different weak classifiers to form a strong one. It has multiple iterations, each one assigns higher weights to misclassified samples and trains a new one to get the final classifier from the weighted sum of all weak classifiers.
3. Attentional cascade structure: This is an algorithm in computer vision used for object detection. The idea of this algorithm is to break down the problem of object detection to a sequence of smaller sub-problems that are solved by a special detector. Each detector is applied to a small region of an image to form a successful detector in object detection [23].

The Viola-Jones used features of rectangles instead of pixels in face detection. Generally, the details of the eyes and face are detected by this algorithm automatically. The eye motion is estimated by the differences in optical flow intensity frame after frame, to decide if the eyes are covered or not by eyelids [24], [25].

This algorithm is used in a wide range of applications because it is the speed and accuracy in detecting objects, especially faces [26], [27].

6. Support Vector Machine Algorithm. Support Vector Machines (SVM) is a supervised machine learning algorithm that is used in classification of Driver Drowsiness Detection to identify drowsy or non-drowsy drivers depending on all based features [28]. After this classification, a sound will generate to alarm and prevent the driver from accidents caused by drowsy driving. SVM is used in regression analysis. In the context of the proposed system, the SVM is trained on different of dataset-labeled examples. Each example has a set of extracted features from a driver's face [29]. The most important features are head pose, eye closure duration, and all other factors of drowsiness indication.

The goal of SVM is creating the best boundary line for segregating the space of n-dimensional into classes [30]. To easily put the new data in the correct group. The hyperplane of the SVM algorithm called for the best decision boundary created by choosing the maximum points or vectors. These points are called support vectors. Figure 6 shows two different groups that are classified by using a hyperplane.

The model of drowsy or non-drowsy drivers is created using the SVM algorithm by training this model using a lot of images of the driver with drowsy and non-drowsy by learning different features of them. So the SVM algorithm will create a boundary between these features (support vectors) to simplify the decision of

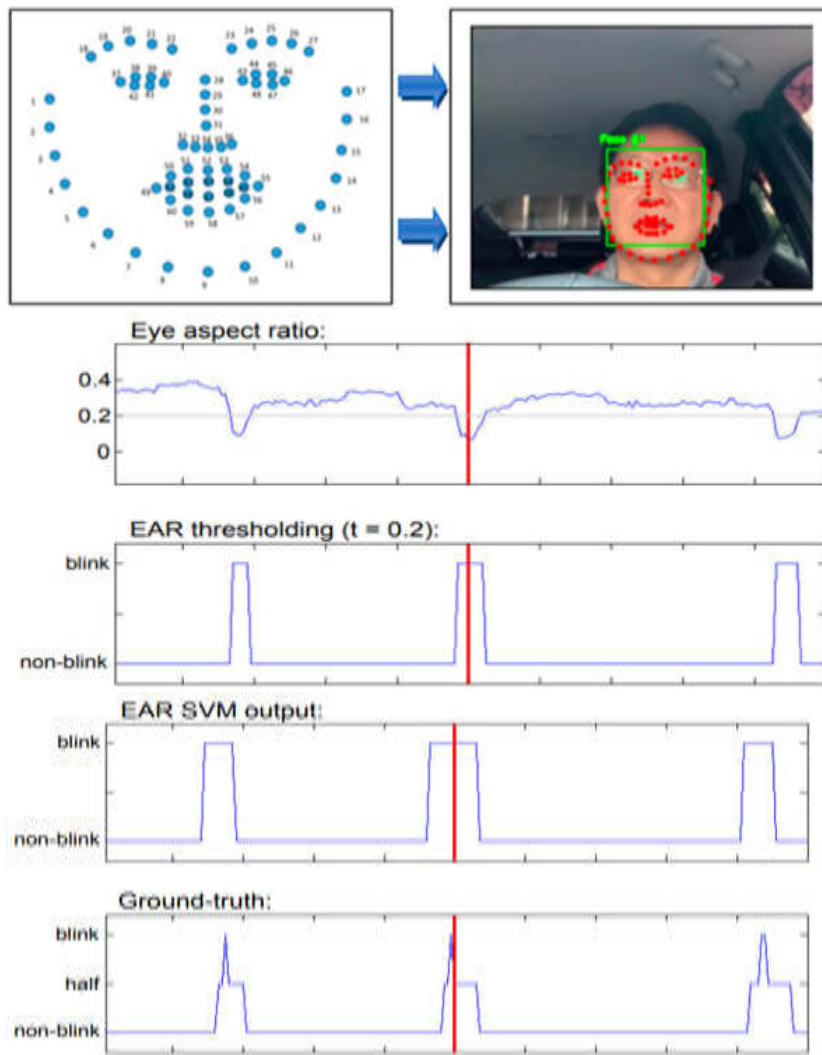


Fig. 5.1: An example of signals of EAR over the video sequence [19], [20]

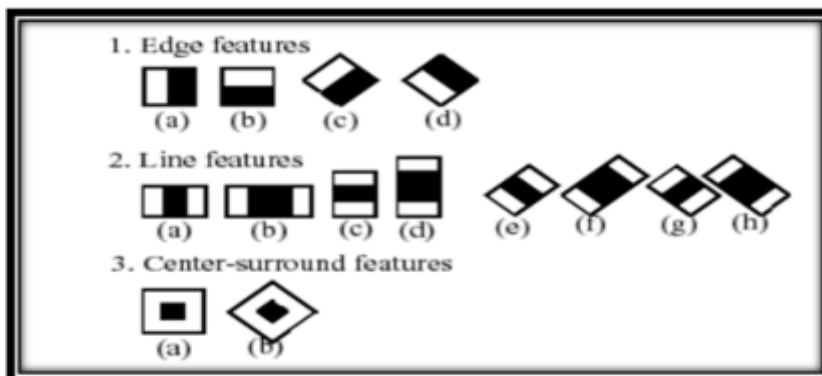


Fig. 5.2: The Features of Haar

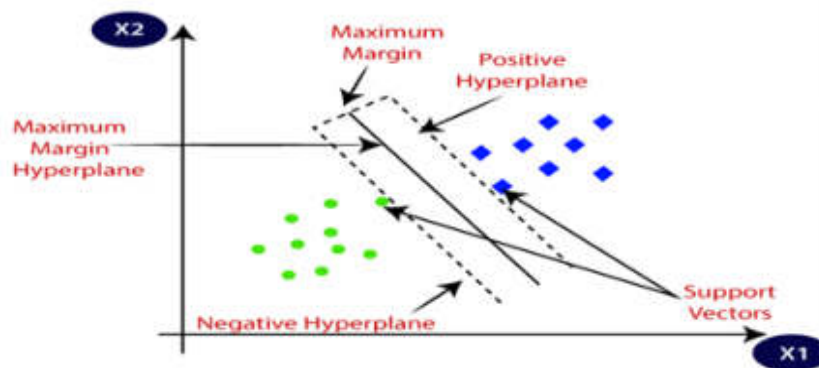


Fig. 6.1: Two different groups [31]

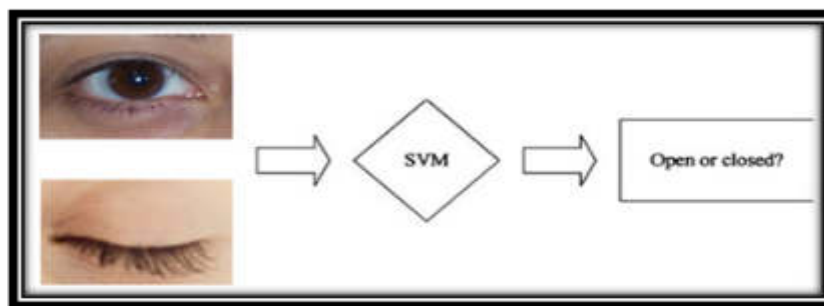


Fig. 6.2: Two different groups [31]

classification between drowsy and non-drowsy shown in Figure 6.1.

7. Python Language. This language is a powerful and easy, multi-programming language, often used for building an application of data science, machine learning, web applications, cyber security, and many development systems. The Python language is an Object Oriented Programming language that has an extensive standard modules library. The Python tools used in this proposed system are TensorFlow, OpenCV, eras, Pygame [32], [33]. These Python tools are used in detecting the closed eyes of drivers. TensorFlow is a free flexible open source library in Python. It was developed by specialists of the Google AI organization. This library has high support in deep neural networks like training and inference. OpenCV is a free open-source library in Python. It has a high performance in machine learning, computer vision, and image processing tasks such as object tracking, face detection, and many more tasks. It can recognize faces, eyes, and all objects from videos and images. This tool will monitor every image picked by the webcam and then feed this image into the proposed system model of deep learning to classify the eyes of the driver if it is opened or closed.

Keras is a free open-source library in Python. It is a built-in library that provides a high-performance Artificial Neural Network. This tool is used in building the proposed system classification model. Pygame is a free open-source library in Python. This tool is used to produce sound to alarm the drivers immediately to pay attention by detecting the driver's closed eyes. Figure 7.1 represents part of the Python code for the proposed system.

8. Smart cameras. Smart cameras are used in the proposed system to monitor the behavior of the driver and detect all the sign of drowsy or non-drowsy. 16-megapixel cameras were used for capturing the photos to be ensured photos with high quality. These cameras are equipped with the advanced of computer vision and image processing algorithms that responsible for analyzing different movement and facial features for detecting

```

import cv2
data = []
labels = []
for j in [60]:
for i in [10]:
vidcap = cv2.VideoCapture('drive/My Drive/Fold5_part2/' + str(j) + '/' + str(i) + '.mp4')
sec = 0
frameRate = 1
success, image = getFrame(sec)
count = 0
while success and count < 240:
landmarks = extract_face_landmarks(image)
if sum(sum(landmarks)) != 0:
count += 1
data.append(landmarks)
labels.append([i])
sec = sec + frameRate
sec = round(sec, 2)
success, image = getFrame(sec)
print(count)
else:
sec = sec + frameRate
sec = round(sec, 2)
success, image = getFrame(sec)
print("not detected")

```

Fig. 7.1: Part of Python code for the proposed system

all drowsiness signs. The most common features of driver are blinking eye rate, the duration of driver's eye closure and head pose. If the camera detect the closing of his eyes for period of time or his head was drooping, sound alarm will produce to stop driving and take a break. [34] [35] This type of cameras will combined with sensors and artificial intelligent or machine learning algorithms to increase the accuracy of this proposed system. Figure 8.1 shows the smart cameras.

Overall using of smart cameras in this proposed system can help in reducing the accidents risk that caused by drowsy driving and improve overall driver safety [36].

9. Alert device. The alert device is very important part in proposed system. The purpose of this part is to alert the driver when the system detects signs of fatigue and drowsiness. It is used in helping to avoid accidents caused by driver's inattention [37]. The alert device used in proposed system is Audible alerts that can include chimes, beep, or other sound that can triggered when the system detects signs of drowsiness. This type of alert is very effective at getting the attention of driver and can be customized and suitable for the driver's preferences. There are different factors the alert device used in this proposed system such as the type of vehicle, the preferences of driver, and the specific requirements of the system. The alert device that is choice is very powerful at getting the driver's attention to take action to avoid accident trough drowsiness.

10. Stacked deep convolution. This type of convolution called stacked con-volutional neural networks (CNNs) is used in the proposed system to improve the accuracy of this system. CNNs are a kind of deep learning algorithm was designed for processing all visual information such as videos and images. The stacking



Fig. 8.1: Smart cameras



Fig. 10.1: The images in different states Drowsy or None

with multiple CNN layers allows for more complex features to be detected and leads to high performance in classification. In the context of the proposed system this type of convolution used in analyzing video data and fed as CNNs input ,from the camera mounted on the vehicle dashboard , and the output of the proposed system is a prob-ability score indicating the driver's drowsy by producing a beep sound from an alert device. The CNNs are trained for detecting all drivers' features such as head nodding, changes in facial expression, and eye closure that are indicative of drowsiness [30]. Through building the proposed system using CNNs a dataset is gathered of images shows the drivers with open, closed eyes, and images that show drivers with different levels of drowsiness. In addition of that the images include different angles of driver position and lighting condition to ensure that CNN can recognize the drowsiness in different real world. Figure 10.1 shows the images in different states.

Also, CNN can be trained in learning all the features like (head nods, droopy eyelids, slower eye movements) that are indicative of drowsiness. CNN output has the ability of driver's classification as asleep, drowsy, or alert.

In this proposed system it is used a technique called data augmentation which include creating variations of the original images for increasing the dataset size.

Overall, CNN is a powerful tool for driver's drowsiness detection because it can be accurate in recognizing the patterns in real-time for indicating drowsiness and producing an alert to the driver to take action to prevent accidents.

Table 11.1: The experiment of the proposed system

Samples	Total	Training	Validation	Testing
Number	3000	1500	800	1500
Drowsy	1500	700	300	500
Non-Drowsy	1500	800	500	1000

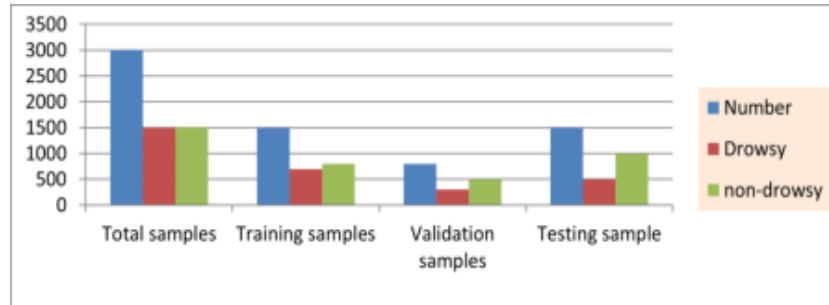


Fig. 11.1: Chart represents experiment data of the proposed system

11. Experiments and Analysis. The proposed system "Driver Drowsiness Detection" is an important area of research that aims to prevent accidents caused by drivers fatigued. It used a deep learning algorithm and some Python tools for feature extraction and classification (drowsy or not). This system is prepared to analyze the driver's faces depending on dynamic sequences of captured photos. There are two types of experiments were performed here. The first experiment is collecting a dataset, it generates a dataset with 3000 images see Table 11.1. The second experiment is performed in videos.

Out of 3000 images used for the experiment in the proposed system, 1500 images are drowsy and others are non-drowsy. For experimenting, 1500 images were used for training, 800 images were non-drowsy, and 700 images were drowsy. A total of 800 images were used for validation samples, 300 images were drowsy, and 500 images were non-drowsy it was diagnosed in several milliseconds. An 1500 images are used for testing, out of 500 images are drowsy, and 1000 images are non-drowsy Figure 11.1 shows a chart for experiment data. The accuracy of the proposed system is about 99.1% after testing the dataset. During the second experiment, the video frame was captured through smart cameras and generated an alarm by an alert device when the proposed system predicts drowsy. The static images are used in the training phase, but through the testing stage, the keyframe is extracted from continuous videos captured by smart cameras.

12. Conclusion. In this paper a new tool was proposed as a saver strategy for vehicle drivers through taking alcohol, drugs, and lack of sleep to protect them from expected accidents. In this state, it builds a monitoring system for detecting drowsiness. This system distinguishes between non-drowsy and drowsy and generates an alarm sound when the eyes are closed.

There are different algorithms and materials are used in such detection. A Viola-Jones detection algorithm is used for detecting the face and eye portion. The learning phase includes extracting all features of the driver's face by using a neural network algorithm called stacked deep convolution. The accuracy of the previous works is low in more time and most of these papers are surveys, which is why it prepared a real-world practical system to protect the people from the high numbers of accents. Support Vector Machines (SVM) is a supervised machine learning algorithm that is used in the classification of Driver Drowsiness Detection to identify drowsy or non-drowsy drivers depending on all based features. An alert device is used proposed system to alert the driver when the system detects signs of fatigue and drowsiness. Smart cameras with 16-megapixel were used for capturing the photos and videos to ensure that the photos were of high quality to monitor the behavior of the driver and detect all the signs of drowsy or non-drowsy. Python tools are used in this system like TensorFlow,

OpenCV, eras, and Pygame. These Python tools are used to detect the closed eyes of drivers. The accuracy of the proposed system is about 99.1%. Out of 3000 images used for the experiment in the proposed system, 1500 images are drowsy and others are non-drowsy. Generating alarm sound from the alert the device effectively when the proposed system identifies drivers' drowsiness.

13. Acknowledgment. The authors would like to show their gratitude to the Universities of Mosul and Technical College Kirkuk, Northern Technical University in Iraq for providing encouragement and support. There was no funding for this study.

REFERENCES

- [1] Amodio, A., Ermidoro, M., Maggi, D., Formentin, S., Savaresi, S.M., *Automatic detection of driver impairment based on pupillary light reflex*, IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 8, p.p. 3038-3048, 2018.
- [2] Yang, J.H., Mao, Z.H., Tijerina, L., Pilutti, T., Coughlin, J.F., Feron, E., *Detection of driver fatigue caused by sleep deprivation*, IEEE Transactions on systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 39, no. 4, pp. 694-705, 2009.
- [3] Hu, S., Zheng, G., *Driver drowsiness detection with eyelid related parameters by support vector machine*, Expert Systems with Applications, vol. 34, no. 4 pp. 7651-7658. . 2009.
- [4] Mardi, Z., Ashtiani, S.N., Mikaili, M., *EEG-based drowsiness detection for safe driving using chaotic features and statistical tests* Journal of Medical Signals and Sensors, vol. 1, no. 2. pp. 130-137. 2011.
- [5] S. Vishwakarma and A. Agrawal, *A survey on activity recognition and behavior understanding in video surveillance* The Visual Computer, vol. 29, pp. 983-1009, 2013.
- [6] M. Zhang, S. Xu, W. Song, Q. He, and Q. Wei, *Lightweight underwater object detection based on YOLO v4 and multi-scale attentional feature fusion* Remote Sensing, vol. 13, no. 22, p. 4706, 2021.
- [7] C. Anil, Y. Wu, A. Andreassen, A. Lewkowycz, V. Misra, V. Ramasesh, et al., *Exploring length generalization in large language models* In Advances in Neural Information Processing Systems, vol. 35, pp. 38546-38556, 2012.
- [8] M. Jafari, A. Kavousi-Fard, T. Chen, and M. Karimi, *A review on digital twin technology in smart grid*, Transportation system and smart city: Challenges and future, IEEE Access., 2023.
- [9] M. Aljasim and R. Kashef., *E2DR: a deep learning ensemble-based driver distraction detection with recommendations model*, Sensors, vol. 22, no. 5, p. 1858, 2022.
- [10] A. A. de Hond, A. M. Leeuwenberg, L. Hooft, I. M. Kant, S. W. Nijman, H. J. van Os, et al., *Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review*, NPJ digital medicine, vol. 5, no. 1, p. 2, 2022.
- [11] Abtahi, S., Hariri, B., Shirmohammadi, S., *Driver drowsiness monitoring based on yawning detection*, IEEE International Instrumentation and Measurement Technology Conference, Binjiang, pp. 1-4, 2011.
- [12] Dwivedi, K., Biswaranjan, K., Sethi, A., *Drowsy driver detection using representation learning*, Advance Computing Conference (IACC), IEEE, pp. 995-999, 2014.
- [13] Alshaquaqi, B., Baquhaizel, A.S., Amine Ouis, M.E., Boumehed, M., Ouamri, A., Keche, M., *Driver drowsiness detection system. 8th International Workshop on Systems, Signal Processing and Their Applications (WoSSPA)*, 2013.
- [14] Park, S., Pan, F., Kang, S., Yoo, C.D., *Driver drowsiness detection system based on feature representation learning using various deep networks*, In: Asian Conference on Computer Vision. Cham: Springer International Publishing, p. 154-164, 2016.
- [15] Tadesse, E., Sheng, W., Liu, M., *Driver drowsiness detection through HMM based dynamic modeling*, 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, pp. 4003-4008, 2014.
- [16] Said, S., AlKork, S., Beyrouthy, T., Hassan, M., Abdellatif, O., Abdraboo, M.F., *Real time eye tracking and detection- a driving assistance system. Advances in Science, Technology and Engineering Systems Journal*, vol.3, no. 6, p.p. 446-454, 2018.
- [17] SPicot, A., Charbonnier, S., Caplier, A., *On-Line Detection of drowsiness using brain and visual information.*, IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, vol. 42, no. 3, p.p. 764-775, 2012.
- [18] Mandal, B., Li, L., Wang, G.S., Lin, J., *Towards detection of bus driver fatigue based on robust visual analysis of eye state*, IEEE Transactions on Intelligent Transportation Systems, vol. 18, no. 3, pp. 545-557, 2017
- [19] Jabbar, R., Al-Khalifa, K., Kharbeche, M., Alhajyaseen, W., Jafari, M., Jiang, S., *Real-time driver drowsiness detection for android application using deep neural networks techniques*, Procedia Computer Science, vol. 130, p.p. 400-407, 2018.
- [20] Viola, P., Jones, M.J., *Robust real-time face detection*, International Journal of Computer Vision, vol. 57, no. 2, p.p. 137-154., 2004. <https://doi.org/10.1023/b:visi.0000013087.49260.fb>
- [21] Jensen, O.H., *Implementing the Viola-Jones face detection algorithm*, (Master's thesis, Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark), 2008.
- [22] O'Shea, K., Nash, R., *An introduction to convolutional neural networks*, arXiv preprint arXiv:1511.08458, 2015.
- [23] Kim, K., Hong, H., Nam, G., Park, K., *A study of deep CNN-based classification of open and closed eyes using a visible light camera sensor*, Sensors, vol. 17, no. 7, p.1534, 2017.
- [24] Lee, K., Yoon, H., Song, J., Park, K., *Convolutional neural network-based classification of driver's emotion during aggressive and smooth driving using multi-modal camera sensors*, Sensors, vol. 18, no. 4, p. 957, 2018.

- [25] Rukhsar Khan, Shruti Menon, Shivraj Patil, Suraj Anchan, Saritha L. R., *Human Drowsiness Detection System*, International Journal of Engineering and Advanced Technology (IJEAT), Vol. 8, no. 4, 2019.
- [26] B. Mohana and C. M. Sheela Rani, *Driver Drowsiness Detection Based on Yawning De-tecton and Eye Closure*, International Journal of Recent Technology and Engineering (IJRTE), Vol. 8, no. 4, 2019.
- [27] Shivani Sheth, Aditya Singhal, V.V. Ramalingam., *Driver Drowsiness Detection System using Machine Learning Algorithms*, International Journal of Recent Technology and En-gineering (IJRTE), Vol. 8, no. 6, 2020.
- [28] Y. Peng, Q. Xu, S. Lin, X. Wang, G. Xiang, S. Huang, et al., *The application of electroen-cephalogram in driving safety: current status and future prospects*, Frontiers in psychology, vol. 13, p. 919695, 2022.
- [29] S. Majid, F. Alenezi, S. Masood, M. Ahmad, E. S. Glundluz, and K. Polat., *Attention SIAM MACRO EXAMPLES 13 based CNN model for fire detection and localization in real-world images*, Expert Systems with Applications, vol. 189, p. 116114, 2022.
- [30] P. Naga, S. D. Marri, and R. Borreo, *Facial emotion recognition methods, datasets and technologies: A literature survey*, Materials Today: Proceedings, vol. 80, pp. 2824-2828, 2023.
- [31] Zainab Ali Abbood et al., *Driver Drowsy and Yawn System Alert Using Deep Cascade Con-volution Neural Network DCCNN*, Iraqi Journal for Computer Science and Mathematics, vol. 4, no. 4, p.p. 111-120, 2023.
- [32] M. K. Rusia and D. K. Singh, *A comprehensive survey on techniques to handle face identity threats: challenges and opportunities*, Multimedia Tools and Applications, vol. 82, no. 2, pp. 1669-1748, 2023.
- [33] T. Bhatnagar et al., *A Pixelated Interactions: Exploring Pixel Art for Graphical Primi-tives on a Pin Array Tactile Display*, Proceedings of the 2023 ACM Designing Interactive Systems Conference, pp. 1194-1208, 2023.
- [34] T. V. N. S. R. Sri Mounika, P. H. Phanindra, N. V. V. N. Sai Charan, Y. Kranthi Kumar Reddy, and S. Govindu., *Driver Drowsiness Detection Using Eye Aspect Ratio (EAR), Mouth Aspect Ratio (MAR), and Driver Distraction Using Head Pose Estimation*, ICT Systems and Sustainability: Proceedings of ICT4SD 2021, Vol. 1, pp. 619-627, Springer Singapore, 2023.
- [35] G. Gao, Q. Bai, C. Zhang, L. Zhang, and L. Yao., *Dualistic cascaded convolutional neural network dedicated to fully PolSAR image ship detection*, ISPRS Journal of Photogrammetry and Remote Sensing, vol. 202, pp. 663-681, 2023.
- [36] B. Dudi and V. Rajesh., *A computer-aided plant leaf classification based on optimal feature selection and enhanced recurrent neural network*, Journal of Experimental and Theoretical Artificial Intelligence, pp. 1-35, 2022.
- [37] Hybrid dual-channel convolution neural network (DCCNN) with spider monkey optimization (SMO) for cyber security threat detection in Internet of things, Measurement: Sensors, vol. 27, p. 100783, 2023.

Edited by: Mustafa M Matalgah

Special issue on: Synergies of Neural Networks, Neurorobotics, and Brain-Computer Interface Technology: Advancements and Applications

Received: Dec 17, 2023

Accepted: Feb 13, 2024



ENSEMBLE TRANSFER LEARNING FOR BOTNET DETECTION IN THE INTERNET OF THINGS

ALI AALSAUD*, SHAHAB WAHHAB KAREEM[†], RAGHAD ZUHAIR YOUSIF[‡] AND AHMED SALAHUDDIN MOHAMMED[§]

Abstract. Botnet attacks are just one security scalability problem that nearly comes as a default with each and every new IoT system launched into the real world. IoT devices, in particular, are tricky to locate on a network with standard methods of botnet detection due to their inherent volatility and system constraint developments. To this aim, we propose an ensemble method for botnet detection based on transfer learning that mitigates those drawbacks. The representation learning-based method is used to deliver a domain-adapt transfer of data between two domains (one that has traditional network data and other that contains IoT devices). Ensemble Method This technique improves the detection accuracy and robustness by employing pre-trained models and customizing them to the target IoT environment using many models working together. The ensemble transfer learning system includes low-level base classifiers (e.g., AlexNet, VGG16, inceptionV3, Mobile Net) that are trained on various IoT data and features. To utilize the domain-specific information effectively, the authors investigate model stacking and domain adaptation as two transfer learning strategies. The authors also consider feature engineering methods to determine signatures of IoT behavior and aid their models to distinguish between normal device behavior and botnet activities. The authors also perform extensive experiments on real-world IoT datasets to show the efficacy of the proposed ensemble transfer learning approach. In comparison to single-model techniques, the results show considerable gains in botnet detection accuracy, sensitivity, and specificity. The ensemble technique is also resilient to different IoT device types and network circumstances, making it appropriate for real-time deployment in various IoT contexts. In comparison to single-model techniques, the results show considerable gains in botnet detection accuracy, sensitivity, and specificity. The ensemble technique is also resilient to different IoT device types and network circumstances, making it appropriate for real-time deployment in various IoT contexts.

Key words: Deep Learning, Botnets, Detection, Transfer Learning, Internet of Things.

1. Introduction. The Internet of Things (IoT) has witnessed unprecedented growth, leading to an ever-expanding network of connected devices. This rapid expansion, while beneficial, introduces significant security vulnerabilities, particularly in the form of botnet attacks. Botnets, networks of compromised devices controlled by attackers, pose a substantial threat due to their capacity to execute coordinated cyberattacks, data breaches, or distributed denial-of-service (DDoS) operations. The IoT environment, characterized by its dynamic nature and diverse array of devices with varying capabilities, presents unique challenges for botnet detection. These challenges are compounded by the resource constraints inherent to many IoT devices, which limit the effectiveness of traditional botnet detection methods primarily designed for more static, homogeneous network environments [1][2].

Unlike traditional networks, IoT ecosystems comprise a wide variety of devices with different hardware configurations, communication protocols, and data generation patterns. This heterogeneity makes it particularly challenging for conventional botnet detection techniques to accurately identify malicious activities. The limitations of these methods in the context of IoT's diverse and dynamic nature necessitate an innovative approach to enhance the accuracy and efficiency of botnet detection. In response to these challenges, we propose a novel ensemble method that leverages the capabilities of transfer learning to improve botnet detection in IoT ecosystems. Transfer learning, a powerful deep learning paradigm, allows for the transfer and adaptation

*Computer Engineering Department, College of Engineering, Al-Mustansiriyah University Baghdad, Iraq
a.m.m.aalsaud@uomustansiriyah.edu.iq

[†]Information System Engineering Department, Technical Engineering College, Erbil Polytechnic University, Erbil 44001, Iraq.
Shahab.kareem@epu.edu.iq

[‡]Department of Physics, College of Science, Salahaddin University, Erbil, KRG, Iraq. raghad.yousif@su.edu.krd

[§]Department of Information Technology, College of Engineering and Computer Science, Lebanese French University, Erbil 44001, Iraq

of knowledge acquired in one domain (the source domain) to a new, relevant domain (the target domain), in this case, IoT environments [4][5]. Our ensemble approach intelligently utilizes pre-trained models from source domains, such as traditional network data, and optimizes them for effective functioning within the IoT domain. It incorporates multiple base classifiers, including renowned deep learning architectures like AlexNet, VGG16, InceptionV3, and MobileNet, each fine-tuned to the unique data and characteristics associated with IoT devices.

To effectively leverage knowledge from the source domain, our study explores various transfer learning strategies such as model stacking and domain adaptation. Additionally, we delve into feature engineering techniques specifically designed to capture the unique behaviour patterns exhibited by IoT devices. Existing botnet detection approaches in IoT environments often struggle with scalability, adaptability, and accuracy. These methods typically fail to account for the heterogeneous and evolving nature of IoT networks, resulting in suboptimal detection and increased false positives. There is a clear gap in developing detection techniques that can dynamically adapt to the IoT's varied landscape while maintaining high accuracy and low resource consumption. Our Contribution: Addressing these challenges, our research introduces a novel ensemble method leveraging transfer learning to improve botnet detection in IoT environments. This approach represents a significant advancement in several ways:

Adaptation to IoT Heterogeneity: By employing transfer learning, our method adeptly adapts knowledge from traditional network contexts (source domain) to the diverse IoT environment (target domain), a crucial step overlooked by existing methods.

Incorporation of Advanced Deep Learning Models: We use well-known architectures like AlexNet, VGG16, InceptionV3, and MobileNet, each fine-tuned to IoT-specific data characteristics, a strategy rarely adopted in conventional IoT security solutions.

Customized Feature Engineering: Our method involves developing feature engineering techniques tailored to the unique behavioral patterns of IoT devices, enhancing the precision in differentiating between normal operations and botnet activities.

We empirically validate our approach using a full validation to demonstrate the strength of our method in detecting anomalies at a higher level compared to existing approaches available with a single model, on a number of IoT datasets from our industry partners. **Flexibility and Range of Use:** Our ensemble method is extremely well-suited for real-time deployment across a wide range of IoT settings, and it is robust to different types of IoT devices and network conditions. This study aims to solve a critical deficiency in IoT security by building the new, efficient, and scalable botnet detection methodology. This work allowed us to set new baselines in Internet of Things (IoT) security and will continue to accelerate progress in this key area. **Large-scale Trials on Real-world Internet of Things Datasets to Evaluate Ensemble Transfer Learning Method Results** obtained illustrate significant improvements in botnet detection accuracy, sensitivity and specificity in comparison with common single-model methods. What's more, our ensemble approach is robust and can resistant to a wide range of IoT devices and network environments, which renders it a suitable choice for real-time transmission in a variety of IoT scenarios.

2. Literature Review. Traditional botnet detection algorithms do encounter some limitations when applied in IoT scenarios, simply because they can be stemming from more conventional network data. The diversity of IoT devices in their hardware configurations, connection protocols and data patterns make it difficult for traditional methods to properly detect botnet activities. In this paper, we provide solutions to these problems through presenting an ensemble-based transfer learning technique to achieve higher accuracy on IoT Botnet detection [6]. One major concern is that IoT systems are not built with security in mind, and this poses significant concerns for botnet attacks. The Internet of Things (IoT) is so diverse and constantly developing that it represents an issue for traditional botnet detection technology, because this technology was created to operate in environments that are more stable and homogenous. Such solutions are typically too inflexible to manage the huge array of hardware configurations, communication protocols, and the vagaries of data patterns in a IoT network. To this end, the current research gap calls for the development of detection techniques that are more flexible and reliable, which are able to take the unique characteristics of IoT devices into the account.

Our study proposes this new ensemble method which leverage transfer learning to advance the botnet detection in IoT ecosystems to address these problems. Transfer learning is a powerful deep learning technique

that allows knowledge gained in more general network environments (source domain) to be transferred to the IoT (target domain). This is must do for remediation of issues that are in mainstream practice which do not account with IoT network of hundreds of other protocols. applying an advanced ensemble-learning approach to reinforce the security of IoT devices. Bandara addressed the concern over increasing security threats, in particular scale-based botnet attacks, in the changing IoT landscape. How does a botnet work, and how might cybercriminals use botnets to carry out DDoS attacks, or cyber-attacks and data breaches? Traditional intrusion detection Approaches which are designed for static network data, find it hard to keep pace with a constantly changing and resource constrained nature of IoT devices. The authors propose a conceptual model for transfer learning in ensemble learning to address these issues.

Transfer learning is really the ability to generalize from one area to another. In this transference of data between classical networks and IoT devices, an advantage is that the capability of detecting botnets is enhanced [7]. This is echoed in the further security challenge in IoT contexts raised by the authors, namely the detection of attacks which could have significant impact on the available linked devices [8]. The model improves the efficiency and accuracy of identifying cyberattacks by modifying transfer learning towards the constrained nature of IoT networks. The collaborative part of the model of many Internets of Things (IoT) devices then shares to teach other devices to recognize threats. Each device adds its own strengths to the collective model, by tapping into the pool of knowledge accumulated by all other contributors. Zhang and his team used this type of collaborative learning to fine-tune the model to work with various IoT networks and devices. This research article uses realistic IoT network data going through the Deep Transfer Learning model. In [9] this research aims to improve the security of IoT networks in terms of an easy way to detect potential security vulnerabilities and attacks. Deep Transfer Learning: One DNN based approach, which uses transfer learning techniques and the idea is to shift from one domain to the other, one can be some common network traffic, but the other will be IoT devices. This helps the model to get better at detecting based on how the IoT devices behave and its unique properties. The performance of Deep Transfer Learning model is validated with the results from the experiments in this study. Results of statistical investigations using real-world IoT datasets show that our model can attain improved intrusion detection accuracy when compared to most traditional single-domain techniques.

Transfer learning is employed in the proposed IDS which refers of transfer learning and Optimized Convolutional Neural Networks (CNN), two vital methods. However, due to the number of threats in the IoV domain, transferring previously learned models to IoV environment improves the DL model. The data from the IoV can be easily and accurately extracted as features through the optimized CNN architecture. This paper examines the Transfer Learning and Optimized CNN-based IDS in the IEEE International Conference on Communications (ICC) 2022 and then evaluates its performance. In this section, we evaluate the intrusion detection performance of our model using IoV datasets from the real world. The findings indicate how efficient the design is in monitoring and blocking attacks in IoV networks. This is the brief introduction about the detection by deep learning according to the tabulated form Table 2.1. For example, article [11] likely expresses the depths of progressive neural networks can be applied to enhance the mechanism of IIoT security defense. This is may be an opening paragraph about how IIoT systems need better detection principles and a more complex threat landscape The next paragraph will provide examples of the neural network topologies of the most convenient architecture, consider in which patterns they are operating, their low-level responsiveness and accuracy of threat detection. Authors of this study [22] may have some experience detecting malware on IoT devices with machine learning. P1: New malware advancements make them harder for untrained eyes to know where they are, and security threats in IoT are always changing. Then the next paragraph could explain how IoT networks are deployed, which machine learning models are working and how well those models detect different types of malwares. The authors probably explore potential transfer learning uses in the domain of Internet of Things intrusion detection.

Next section will then discuss a definition of transfer learning, its traditional applications, and how it can be applied to the area of Security in the Internet of Things (IoT). The next part will likely detail the ways that transfer learning might raise robustness of anomaly detection of intrusion attacks against zero-day attacks and decrease the time needed for retraining take place [13]. This article provides an ensemble tree model which might help them to detect intrusions occurred in IIoT. Previously an explanation of why ensemble methods

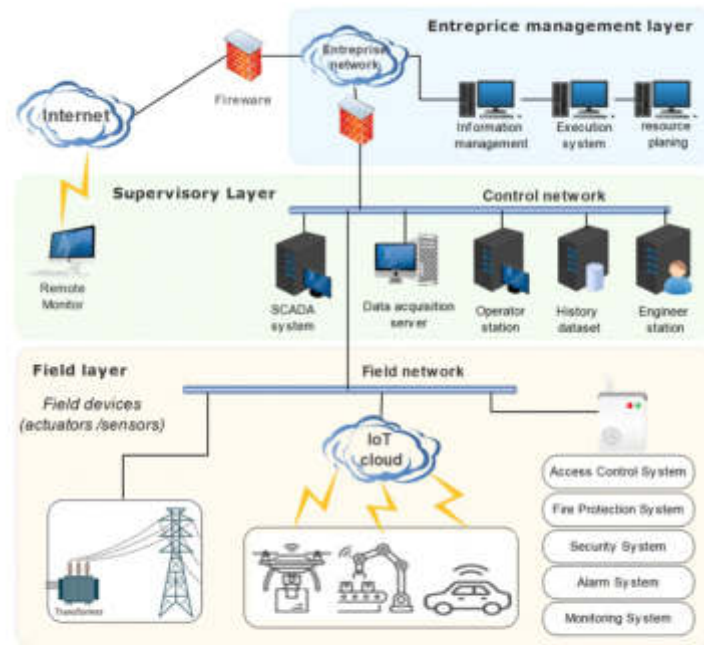


Figure 1 The general layout of an ICN.

Fig. 2.1: The general layout of an ICN

are so powerful for detection, please check the first paragraph. For example, paragraph one could cover the benefits of using an ensemble model rather than single model approach, the architecture of the model, and the performance of the model in different IIoT scenarios [14]. Threat Model for Smart Home Attack Detection using Transfer Learning is depicted in [15], hence, this article likely discusses SALT, as an approach to danger prediction in smart home settings, which is based on transfer learning. Perhaps even some notes on the use of transfer learning and on some of those security considerations that smart homes have. However, as you will see, this opens the door to a detailed description of SALT design and how it extends previous work, and where SALT is useful in a smart home scenario presents a hybrid IDS based on different techniques for the protection of IoT [16]. The first paragraph can list the advantages of hybrid systems, as well as provide an overview of how it works. You can expand more about the hybrid strategy which was used and how the different methodologies have been incorporated and categorize it more effective in the overall system in second paragraph [17].

3. Methodology. The Internet of Things (IoT) is an oven of problems the minute it grows into such rapid proportions where wreaking havoc through botnet attacks is a recurring theme from the security risks to devices that come about. This will include ensemble approach using transfer learning for to leverage IoT botnet discovery. By taking an ensemble approach to combine the intelligence of multiple models, the approach devised improves the robustness and accuracy of botnet detection. The goal of this approach it to combine the base classifiers so that the mistakes of one base classifier are corrected by the another in order to detect Internet of Things (IoT) devices participating in botnet activities.

Figure 3.1 shows the complete proposed model Transfer Learning is one such paradigm of deep learning in which the knowledge from one domain (source domain) is transferred to adapt and be used in the other domain. In this work, they propose to transfer from the source domain (conventional network traffic) to the target domain (IoT devices). This allows the models to adjust their conduct to carry out effectively in the IoT environment based on what they learned previously. Basic classifiers such as AlexNet, VGG16, inceptionV3, and MobileNet are employed to enhance the transfer learning process. These models have been adapted to

Table 2.1: Comparison of some related work.

Ref	Methods	Dataset	Evaluation	Achievement	Analysis
11	Deep Progressive Neural Networks	Industrial IoT	Mobile Networks and Applications	Improved security in IIoT using DPNs	Advancements in IIoT security using deep progressive NNs
12	Machine Learning	IoT Devices	-	Understanding IoT malware and protection strategies	Understanding IoT malware and strategies for protection
13	Transfer Learning	IoT	2022 IEEE ICETCI	Improved IoT Intrusion Detection based on Transfer Learning	Effective IoT IDS based on transfer learning
14	Ensemble Tree-Based Model	Industrial IoT	Applied Sciences	Enhanced intrusion detection in IIoT networks	Improved IDS in IIoT using an ensemble tree-based model
15	SALT: Transfer Learning	Smart Home	Scientific Reports	Transfer learning-based attack detection in smart homes	Effective attack detection in smart homes using transfer learning
16	Ensemble Hybrid IDS	IoT Attacks	Electronics	Efficient ensemble-based IDS for IoT attacks	Effective ensemble-based IDS for detecting IoT attacks
17	Transformers-based Transfer	Malware Detection	Sensors	Explainable malware detection system	Effective malware detection using transformers and visuals
18	Ensemble-Based IDS	Internet of Things	Arabian Journal for Science and Engineering	Improved ensemble-based IDS for IoT	Effective IDS for IoT using an ensemble approach
19	Deep Transfer Learning	Internet of Medical Things	2022 ICATIECE	EEG Signal Classification using deep transfer learning	Effective EEG signal classification in IoMT using DTL
20	Hybrid Deep Learning Model	Internet of Things	Computer Communications	IoT attack detection using hybrid deep learning	Effective attack detection in IoT using hybrid DL model
21	Feature Selection	Intrusion Detection in IoT	ICT Express	Effective feature selection for IoT IDS	Improved feature selection for IoT intrusion detection
22	Enhanced Flower Pollination	IoT Network	Concurrency and Computation	Enhanced IDS for IoT using EFP algorithm	Improved IDS in IoT networks using enhanced FP algorithm
23	Transfer Learning, MobileNetV2	Internet of Vehicles	Multimedia Tools and Applications	Lightweight IDS for IoV using TL and MobileNetV2	Effective IDS for IoV with lightweight TL and MobileNetV2

deal with data that is suitable for IoT and the IoT botnet-originating features. To fully utilize the learned information from the original domain, researchers are also exploring model stacking and domain adaptation-based transfer learning approaches, which stack multiple models and transfer one (trained) or few top layers respectively.

These techniques are aimed at enhancing the pre-trained ones to detect botnet activities on Internet of Things devices. The Technique, in its suggestion, solves the security issues arose by the rapid growth of Internet of Things (IoT) and the botnet attacks. The technique tackles these challenges by achieving significant improvement in botnet detection performance using feature engineering, pre-trained model adaptation, ensemble learning, and transfer learning. This versatility in how it handles various classes of IoT devices and networking environments gives the method greater practical value for IoT environments that companies will encounter in reality. The work offers a valuable perspective on IoT security from the aspect of botnet identification using deep learning. In this paper, an ensemble approach with transfer learning is recommended for detecting botnets in IoT environments. Traditional methods are ineffective at identifying botnets across diverse IoT use cases.

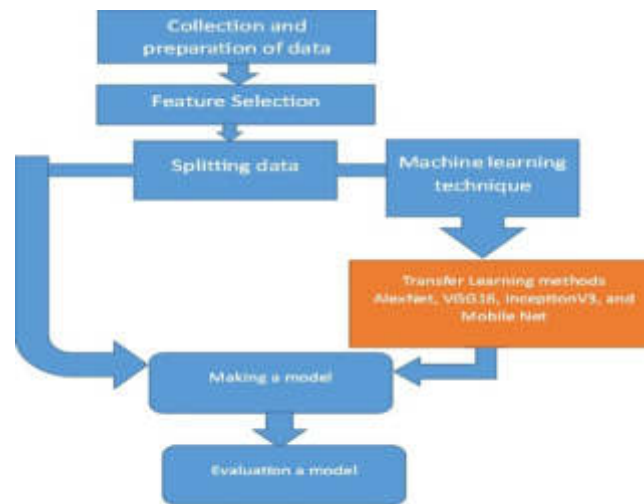


Fig. 3.1: Overall Proposed models

This strategy is designed to combat those limitations.

1. The ensemble learning method is used to increase the accuracy and robustness of botnet detection by aggregating different model intelligences. Ensemble learning is used to detect this botnet behaviour in IoT devices as ensemble learning is the process in which the strong model is built by combining multiple base classifiers.
2. This is a knowledge transfer and adaption approach A data can be transferred from one domain as traditional network data and can be adapted in to another domain, i.e. the domain of Internet of Things devices. Using their earlier training, the models can adjust to the IoT environment.
3. Use Of Pre-Trained Models: For peak performance with IoT data, leverage pre-trained models like AlexNet, VGG16, InceptionV3, MobileNet. These models have been further fine-tuned to detect botnets under IoST environment. Step Four: Repair Data Imbalances Which can easily fixed by using something like over- or under-sampling, or generating fake data in data preprocessing. Data normalization and feature engineering comes under cleaning and preparation for analysis phases
4. Model Training: Multiple Resources for training 75 To a quire model training use Transfer Learning technique because we do have to train a model on a training data. All the measurements e.g. accuracy, sensitivity, and specificity are executed in a testing set;
5. For the 5th step, the data has been separated into two groups 75 to be used for training and 25 for testing and apply Transfer Learning transfer learning to train the model using the training data. Evaluate the model on testing set using measures in terms of accuracy, sensitivity and specificity.
6. Model Optimization and Validation: Apply model stacking and domain adaptation techniques for optimization. Conduct extensive tests with the iot23 dataset to assess the models performance.

Experimental Flowchart:

- Step 1: Gather data from the iot23 dataset.
- Step 2: Address data imbalance and preprocess data.
- Step 3: Divide data into training and testing sets.
- Step 4: Implement Transfer Learning and train the model.
- Step 5: Test and evaluate the model.
- Step 6: Optimize the model using advanced transfer learning methodologies.

Contribution of the Methodology. Our methodology addresses the security issues posed by botnet attacks and the rapid growth of IoT. It significantly enhances botnet detection accuracy using ensemble learning and transfer learning techniques.

The methods adaptability makes it suitable for real-time deployment in various IoT environments.

The following is the architecture of AlexNet: AlexNet model is a deep convolutional neural network (CNN) based on 5 convolutional layers and 3 fully connected layers. The first convolutional layer applies 96 11x11 filters with stride 4, followed by a ReLU activation function and max pooling with 3x3 size and 2 stride. 2nd CONVOLUTIONAL LAYER -> RELU -> MAX POOLING (3x3 strides 2) Convolution Layer 2:256 5x5 filters, stride-1 the fourth and fifth convolutional layers have 256 filters each but they have a size of 3x3 and stride of 1. Each fully-connected layer has 4096 units after a dropout to curb overfitting, followed by a ReLU activation function. The last layer, with an sigmoid activation used for binary classification.

VGG16 is made up of 16 layers, of which there are 3 fully connected layers and 13 are convolutional. And a CNN which have deep learning functionality. Each convolutional layer is followed by a ReLU activation function and a 2x2 max-pool with a stride of 2. Similarly for each layer: 3 by 3 filters and stride of 1. Every fully connected layer has 4096 units following a dropout to prevent overfitting (and a ReLU activation function). Zeros layer: Binary classification with sigmoid activation The inception module, a dense layer containing filters of all possible sizes, is used in deep CNNs (eg, InceptionV3) to identify patterns when featurized data is streamed into our computation frames. By stacking these Inception modules we allow the model to learn even more complicated features. Global Average Pooling Layer is being used that instead of reduce the parameter counts and avoid overfitting i.e. replacing fully connected layers with average pooling. The simple example would be a single unit with sigmoid activation, as the last layer for binary classification. These models are each optimized using the Adam optimizer and trained using a 32-person batch size over 10 epochs. Accuracy is employed as the evaluation metric, while binary cross-entropy is the applied loss function. During the training process, the training data is divided into two portions: 75% for training and 25% for validation. With the iot23 dataset, the objective is to identify botnet activities as accurately as feasible.

4. Discussion and Result. The iot23 dataset is a benchmark dataset created especially for analyzing the performance of machine learning models in the context of intrusion detection and IoT (Internet of Things) network traffic analysis. To answer the demand for standardized and varied datasets for IoT-related security research, a team of researchers created it. The iot23 dataset is very useful for researching the security issues and dangers that IoT settings must deal with because it is made up of network traffic data that was gathered from actual IoT devices and scenarios. The dataset offers a thorough depiction of IoT network traffic because it covers a variety of IoT device types, communication protocols, and traffic patterns. The iot23 datasets accessibility has considerably advanced IoT security research, particularly in the areas of intrusion detection and network traffic analysis. This dataset is made available to network and IoT researchers, developers and the like to enable the advancement of robust and secure machine learning models and algorithms against the threats on networks and IoT devices. The ensemble approach is evaluated for botnet detection in IoT23 dataset, and the performance in terms of the accuracy, Precision, Recall, F1-Score are the performance criteria.

Check the Ensemble model how much good predicting overall. This calculates the number of times the event is accurately predicted (ie true positive and true negative) as a percentage of the total number of cases in the dataset. It means our ensemble model is predicting some substantial amount of our dataset correctly which lead to a high accuracy score. The accuracy of the ensemble model means how much it can spot the botnet instances while expecting them. It is the proportion of prediction cases in which botnet detection has been predicted to the total number of detection cases even wrong ones. Higher precision scores means that the ensemble is more likely to treat instances as a botnet accurately. Recall (also known as sensitivity, or true positive rate) is a metric which tells, what is the ensembled models ability to find all the botnet instances out of all the true botnet instances (labeled as true by the data owner). This is done by computing the number of true positives divided by the total number of botnet instances in the dataset (including the ones which were false negative). A good recall score means, the ensemble model is detecting most of the instances of botnet correctly.

The harmonic mean of recall and precision is known as the F1-Score. It offers a balanced measurement that accounts for both recall and precision. When the distribution of the classes is unbalanced, the F1-Score is helpful. It has a value between 0 and 1, with a higher number indicating better performance. Achieved results compared with related work [26][27] and [28].

The performance outcomes of various botnet detection algorithms using Accuracy, Precision, Recall, and

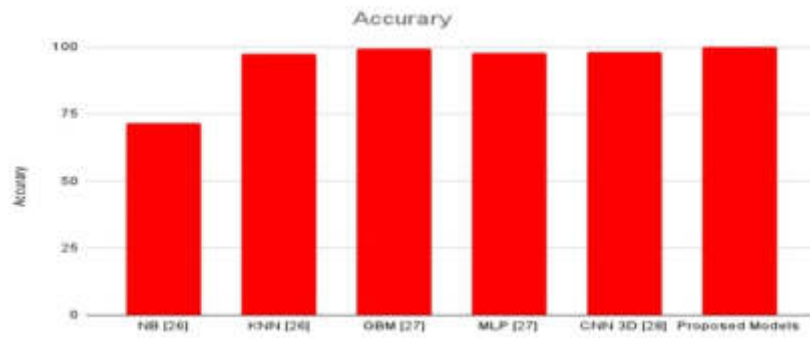


Fig. 4.1: Accuracy comparison of the proposed model

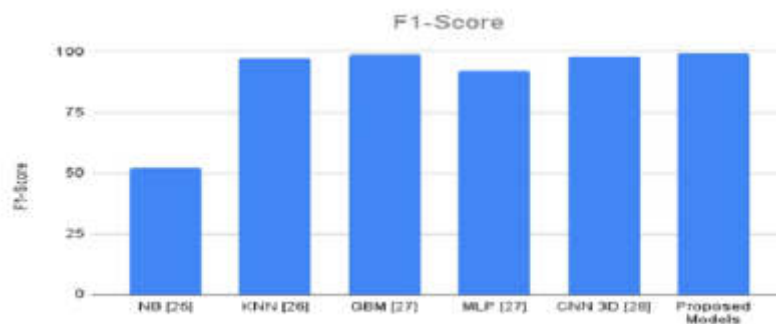


Fig. 4.2: F1_Score comparison of the proposed model

F1-Score as evaluation criteria. The accuracy is shown in Figure 4.1. It is the proportion of cases in the dataset that were successfully predicted to all other instances. The model is better able to produce accurate predictions the higher the accuracy. Accuracy = 71.72 KNN (K-Nearest Neighbours) for NB (Naive Bayes) [26] [26]: GBM (Gradient Boosting Machine) Accuracy = 97.51 [27]: MLP (Multi-Layer Perceptron) Accuracy = 99.452 [27]: Precision is 97.842. Accuracy of CNN 3D (3D Convolutional Neural Network) [28]: 98.13 Accuracy of proposed models: 99.95.

Figure 4.2 shows the f1-score, The F1-Score provides a balanced measurement that takes into account both measures because it is the harmonic mean of precision and recall. When the distribution of classes is unbalanced, it is helpful. Naive Bayes (NB) F1-Score is 52.01 in [26]. K-Nearest Neighbors (KNN) F1-Score is 97.05 in [26]. Machine for gradient boosting [27]: MLP (Multi-Layer Perceptron) F1-Score = Not Available (-) F1-Score = Not Available (-), [27] 3D Convolutional Neural Network or CNN 3D F1-Score is 98.1 in [28]. Models suggested: F1-Score = 99.25.

Figure 4.3 shows recall, The capacity of the model to accurately identify positive cases among all of the real positive examples in the dataset is measured by recall (also known as sensitivity or true positive rate). It measures the proportion of real positives to all actual positives. Naive Bayes (NB) [26]: KNN (K-Nearest Neighbours) Recall = 36.11 Recall = 96.44 [26], 3D Convolutional Neural Network, or CNN 3D Recall = 98.09 [28], suggested models 99.2 of the time.

Precision is a measure of the model's ability to reliably detect positive cases (such as instances of botnets) among those that it expected to be positive. It is the proportion of actual positive results to all expected positive results. Naive Bayes, or NB [26]: Exactness = 92.89 the K-Nearest Neighbours method [26]: Precision GBM (Gradient Boosting Machine) = 97.67 [27]: MLP (Multi-Layer Perceptron) Precision = Not Available (-)

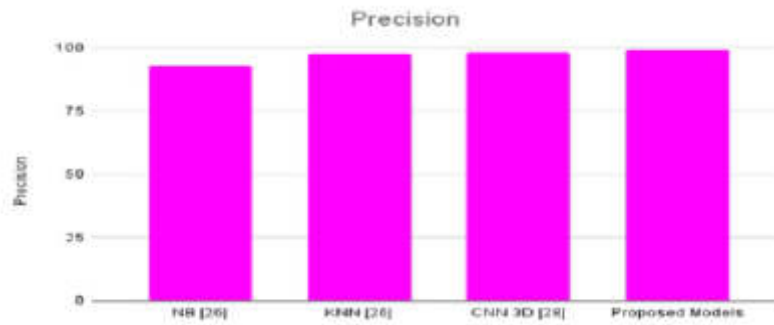


Fig. 4.3: Precision comparison of the proposed model

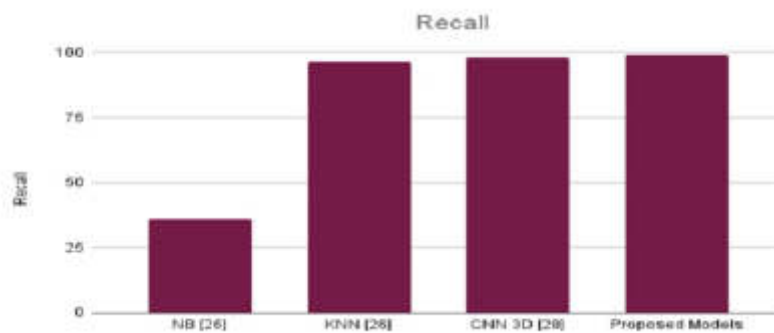


Fig. 4.4: Recall the comparison of the proposed model

Precision = Not Available (-) [27], Using a 3D convolutional neural network, CNN 3D [28]: Exactness = 98.1, Precision = 99.08 for the proposed models, as shown in Figure 4.4.

5. Conclusion. An extensive and practical solution for botnet identification in IoT contexts is provided by the suggested ensemble transfer learning model. The suggested strategy provides a potential method for enhancing IoT security and reducing botnet threats in the changing environment of connected devices by utilising the power of transfer learning, customising previously trained models, and utilising ensemble methodologies. Instead of all that work, the ensemble-methods approach has been proposed to resolve the safety problems that throw up when Internet of Things (IoT) is getting into Its stride. There is a method that greatly increases botnet detection accuracy: this strives after feature engineering and then fits the pre-trained model in particular for your purpose in addition, it is suitable for real-world IoT applications as it can fit with diverse IoT device types and network settings. Numerical results for the method proposed are given, together with comparisons to relevant previous studies. The model's performance is evaluated by using accuracy, precision, recall and f1-score as metrics. Out> The results indicate that the proposed ensemble transfer learning approach is better than any of the other classic machine learning or deep learning models> mentioned in the study, with respect to accuracy, precision, recall and f1-score. In a word, for pure performance the suggestion is best: transferring learning approach It achieves good results on accuracy, precision, recall and F1-Score in a proof of its efficiency for detecting botnet action within the IoT scenario of context. Finally, there is the conclusion. The iot 23 dataset, which is used as a standard comparison data set for machine learning models in network traffic analysis and intrusion detection among the Internet of Things (IoT), has great significance here. The depiction in the dataset of a variety of different IoT device types, communication protocols, and traffic patterns benefits IoT security research greatly. By accurately identifying botnet activity, the ensemble approach with transfer

learning for deep learning-based botnet identification greatly improves IoT security.

REFERENCES

- [1] Q. K. KADHIM, A. S. AL-SUDANI, I.A. ALMANI, T.LGHAZALI, H. K.DABIS, A. T.MOHAMMED, Y. MEZAA *IOT-MDEDTL: IoT Malware Detection based on Ensemble Deep Transfer Learning*, *Majlesi Journal of Electrical Engineering*, 16(3), 47-54.
- [2] F. YAN, G. ZHANG, D. ZHANG, X. SUN, B. HOU, N. YU , *TL-CNN-IDS: transfer learning-based intrusion detection system using convolutional neural network*, *The Journal of Supercomputing*, 1-23, 2023.
- [3] HAMZA KHEDDARA, YASSINE HIMEURB AND ALI ISMAIL AWADC, *Deep Transfer Learning Applications in Intrusion Detection Systems: A Comprehensive Review*, arXiv:2304.10550v1 [cs.CR] 19 Apr 2023.
- [4] P. PANDA, O. K. CU, S.MARAPPAN, S.MA, D. VEESANI NANDI, *Transfer Learning for Image-Based Malware Detection for IoT. Sensors*, 23(6), 3253.
- [5] K.RAMBABU, N. VENKATRAM,*Ensemble classification using traffic flow metrics to predict distributed denial of service scope in the Internet of Things (IoT) networks*. *Computers and Electrical Engineering*, 96, 107444..
- [6] L.VU, Q. U.NGUYEN, D. T.HOANG, D. N.NGUYEN, E. DUTKIEWICZ *A Novel Transfer Learning Model for Intrusion Detection Systems in IoT Networks**In Emerging Trends in Cybersecurity Applications (pp. 45-65), 2023. Cham: Springer International Publishing.*, *Transfer Learning for Image-Based Malware Detection for IoT. Sensors*, 23(6), 3253.
- [7] Y. ALOTAIBI,M. ILYAS*Ensemble-Learning Framework for Intrusion Detection to Enhance Internet of Things Devices Security. Sensors*, 23(12), 5568.2023.
- [8] T. V.KHOA, D.T. HOANG, N.L. TRUNG, C.T. NGUYEN, T.T.T. QUYNH, D. N. NGUYEN, E. DUTKIEWICZ. *Deep transfer learning: A novel collaborative learning model for cyberattack detection systems in IoT networks. IEEE Internet of Things Journal*.2022.
- [9] B. XUE, H. ZHAO, W. YAO, . *Deep Transfer Learning for IoT Intrusion Detection. In 2022 3rd International Conference on Computing, Networks and Internet of Things (CNIOT) (pp. 88-94). IEEE. 2022.*
- [10] L. YANG, A.SHAMIA *transfer learning and optimized CNN based intrusion detection system for Internet of Vehicles. In ICC 2022-IEEE International Conference on Communications (pp. 2774-2779). IEEE 2022.*
- [11] M. SHARMA, S. PANT, P. YADAV,D. SHARMA, N.GUPTA, G. SRIVASTAVA *Advancing security in the industrial internet of things using deep progressive neural networks. Mobile Networks and Applications*, 1-13.2023.
- [12] M.M.ALI, F.MAQSOOD, W.HOU, Z. WANG,K. HAMEED, Q. ZIA.*Machine Learning-Based Malware Detection for IoT Devices: Understanding the Evolving Threat Landscape and Strategies for Protection*.2023.
- [13] W.YUTAO, L. ZHONGTIAN, B.YI, L.JIE, X.FANGZHENG, B.YU*Internet of Things Intrusion Detection System based on Transfer Learning. In 2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI) (pp. 25-30). IEEE.2022.*
- [14] J.B. AWOTUNDE, S.O. FOLORUNSO, A.L. IMOIZE,J.O. ODUNUGA, C.C. LEE, C.T. LI, D.T. DO*An Ensemble Tree-Based Model for Intrusion Detection in Industrial Internet of Things Networks. Applied Sciences*, 13(4), 2479, 2023.
- [15] P. ANAND, Y.SINGH, H. SINGH, M.D.ALSHEHRI, S. TANWAR*SALT: transfer learning-based threat model for attack detection in smart home. Scientific Reports*, 12(1), 12247.2022.
- [16] A.KHRAISAT, I.GONDAL, P.VAMPLEW, J. KAMRUZZAMAN,A. ALAZABA *novel ensemble of hybrid intrusion detection system for detecting internet of things attacks. Electronics*, 8(11), 1210.2019.
- [17] F. ULLAH,A. ALSIRHANI,M.M. ALSHAHRANI,A. ALOMARI, H. NAEEM, S.A. SHAH*Explainable malware detection system using transformers-based transfer learning and multi-model visual representation. Sensors*, 22(18), 6766. 2022.
- [18] A.ABBAS, M.A. KHAN, S. LATIF, M. AJAZ, A.A. SHAH, J. AHMADA *new ensemble-based intrusion detection system for internet of things. Arabian Journal for Science and Engineering*, 1-15.2021.
- [19] P.R.SAXENA, D. CHAUHAN*EEG Signal Classification using Deep Transfer Learning Technique in an Internet of Medical Things Environment. In 2022 Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing and Communication Engineering (ICATIECE) (pp. 1-6). IEEE.2022.*
- [20] A.K. SAHU, S. SHARMA,M. TANVEER, R. RAJA*Internet of Things attack detection using hybrid Deep Learning Model. Computer Communications*, 176, 146-154.2021.
- [21] P. NIMBALKAR, D. KSHIRSAGAR*Feature selection for intrusion detection system in Internet-of-Things (IoT). ICT Express*, 7(2), 177-181.2021.
- [22] R. GANGULA, V, M. M. *Network intrusion detection system for Internet of Things based on enhanced flower pollination algorithm and ensemble classifier. Concurrency and Computation: Practice and Experience*, 34(21), e7103.2022.
- [23] Y. WANG, G. QIN, M. ZOU, Y. LIANG, G. WANG, K. WANG, Z. ZHANG *A lightweight intrusion detection system for internet of vehicles based on transfer learning and MobileNetV2 with hyper-parameter optimization. Multimedia Tools and Applications*, 1-23..2023.
- [24] D.Y.MIKHAIL, R.S. HAWEZI, S.W. KAREEM*An Ensemble Transfer Learning Model for Detecting Stego Images. Appl. Sci.* 13, 7021.2023.
- [25] P. H.Q.AWLA, S.W.KAREEM,A.S. MOHAMMEDA *Comparative Evaluation of Bayesian Networks Structure Learning Using Falcon Optimization Algorithm. International Journal of Interactive Multimedia and Artificial Intelligence*, 527.2023.
- [26] F. HUSSAIN, S. G. ABBAS, U. U. FAYYAZ, G. A. SHAH, A. TOQEER AND A. ALI*Towards a Universal Features Set for IoT Botnet Attacks Detection," 2020 IEEE 23rd International Multitopic Conference (INMIC), Bahawalpur, Pakistan, 2020, pp. 1-6, 2020.*
- [27] T.M. BOOJI, I. CHISCOP,E. MEEUWISSEN, N. MOUSTAFA, F.T. DEN HARTOG, *ToN-IoT: The role of heterogeneity and the need for standardization of features and attack types in IoT network intrusion data sets. IEEE Internet of Things Journal*,

9(1), 485-496.2021.

- [28] STRATOSPHERE LABORATORY. A LABELED DATASET WITH MALICIOUS AND BENIGN IOT NETWORK TRAFFIC. AGUSTIN PARMISANO, SEBASTIAN GARCIA, MARIA JOSE ERQUIAGA, (ACCESSED APRIL 26, 2020). ONLINE. AVAILABLE: [HTTPS://WWW.STRATOSPHEREIPS.ORG/DATASETS-IOT23](https://www.stratosphereips.org/datasets-iot23).

Edited by: Mustafa M Matalgah

Special issue on: Synergies of Neural Networks, Neurorobotics, and Brain-Computer Interface Technology:
Advancements and Applications

Received: Dec 18, 2023

Accepted: Mar 18, 2024



SECURE MEDICAL IMAGE RETRIEVAL USING FAST IMAGE PROCESSING ALGORITHMS

SAMEER ABDULSTTAR LAFTA * AMAAL GHAZI HAMAD RAFASH † NOAMAN AHMED YASEEN AL-FALAH ‡
HUSSEIN ABDULQADER HUSSEIN § AND MOHANAD MAHDI ABDULKAREEM ¶

Abstract. Content Based Image Retrieval (CBIR) is a relatively new idea in the field of real-time image retrieval applications; it is a framework for retrieving pictures from diverse medical imaging sources using a variety of image-related attributes, such as color, texture, and form. Using both single and multiple input queries, CBIR processes semantic data or the same object for various class labels in the context of medical image retrieval. Due to the ambiguity of image search, optimizing the retrieval of a query picture by comparing it across numerous image sources may be problematic. The goal is to find a way to optimize the process by which requested images are retrieved from various storage locations. To effectively extract medical images, we propose a hybrid framework (consisting of deep convolution neural networks (DCNN) and the Pareto Optimization technique). In order to obtain medical pictures, a DCNN is trained on them, and then its properties and classification results are employed. Explore enhanced effective medical picture retrieval by using a Pareto optimization strategy to eliminate superfluous and dominant characteristics. When it comes to retrieving images by query from various picture archives, our method outperforms more conventional methods. Use the jargon of machine learning to propose a Novel Unsupervised Label Indexing (NULI) strategy for retrieving picture labels. To enhance the effectiveness of picture retrieval, we characterize machine learning as a matrix convex optimization using a cluster rebased matrix representation. We describe an empirical investigation on many medical picture datasets, finding that the search-based image annotation (SBIA) schema benefits from our suggested method. As a result, CT images of the lung region are explored in this study by constructing a content-based image retrieval system using various machine learning and Artificial Intelligence techniques. Real-world applications of medical imaging are becoming more significant. Medical research facilities acquire and archive a wide variety of medical pictures digitally.

Key words: Medical image, Image retrieval, Image processing

1. Introduction. Raw images captured by spacecraft, satellites, and cameras in our everyday environments may have their usefulness greatly enhanced by the use of image processing techniques. In the last ten years, several image processing programs have been created. Image processing systems are currently the most popular due to the ease with which personal computers can be maintained, the wide availability of graphics software, and the large capacity of memory devices, etc. [1]. Most of these methods were developed to make use of images obtained from unidentified space probes in real time. Image Processing relates Analogy: It describes the alteration of image via electrical data representation, example for this type of data representation is television, and television signal represents various amplitude to access brightness of image with significant pixel extraction [2].

Processing of digital image. Digital image computer processing with respect to different pixel dimensions, Image can be directed into different dimensions, it defines parallel data objects to serious of pixel operations to retrieve efficient results from original picture representation with image notations. The main advantage of digital image processing is to extract original data precision [3].

The main presentation of this approach is to define magnification of image for effective pixel identification and image. Image to image with different pixel factors in recent formations. Analysis of image is concerned with different measurements from image to image to extract image description with representation of image with pixels. Analysis of image approach describes the features of finding different objects on semantic image

*Middle Technical University, Technical Instructors Training Institute, Baghdad, Iraq (Sameer.abdsattar@mtu.edu.iq),

†Middle Technical University, Technical Instructors Training Institute, Baghdad, Iraq

‡Digital Transformation Department, Senior Chief of Programmers, Iraqi Ministry of Communications Baghdad, Iraq

§Director of Data Centers Management Department, Assistant chief engineer, Iraqi Ministry of Communications Baghdad-Iraq

¶5Director of Data Centers Management Department, Assistant chief engineer, Iraqi Ministry of Communications Baghdad-Iraq

feature representation [4].

In image segmentation process, divide image into different equal parts from input image, segmentation follows isolated applications with respect to interest of objects based on pixel value presentation of original image into sequential feature extraction from original data evaluation in pre-processing with autonomous contents in real time application development. Classification is the label of similar group pixel based on its grey value presentation, in information retrieval information classification is the main effective and mostly used method. Classification is set of pixels with multiple features of particular images [5].

Removable and reduction of degradation of image is called image restoration, it includes de-blurring of images, filtering noise pixel information and data presentation for efficient quality of image. Compression is an essential framework to achieve picture data and transfer to network maintenance in reliable pixel formation and presentation that uses Discrete Cosine Transformation (DCT) based on compression feature extraction. Based on these approaches present in image processing, different types of applications were developed in real time with preferable operation presentation [6].

After evaluation of image processing introduction with developing techniques and approaches with different pixel values. Our research mainly focuses image retrieval in image processing. Image retrieval is a foundational concept in the field of image processing, used to locate specific information using either a search query or an image's metadata. Various real-time applications, including healthcare, satellite data, video surveillance, and digital forensics, have made use of the vast amounts of multimedia data that have become available with the spread of multimedia and internet technologies. They were maintaining that multimedia-related data may be stored efficiently with varied characteristics [7] due to the specific needs of these domains. Text Based Image Retrieval (TBIR) is the most used method for retrieving information. Automatic and human picture retrieval from a variety of image sources form the basis of that search. Content Based Image Retrieval (CBIR) is a user-friendly image search retrieval approach that can extract data from many picture sources, unlike TBIR's human effort and time requirements for image retrieval. This is the fundamental architecture for retrieving images sequentially from many sources, with properties like color, shape, texture, and position supplied as feature vectors in several places. This retrieval method will manifest as indexed visual results in response to user queries. Finally, the indexing process is followed by an efficient searching technique of the picture database, and on the basis of this procedure, relevant user input is collected using a variety of visual processes [8].

Medical picture retrieval and searching using that term to get matching information from several medical image sources. The primary goal of this content-based method to medical image retrieval is to sift through a great quantity of data sources, each of which is characterized in terms of the query picture. In addition to the major component of medical images, features are also a key component to investigate utilizing feature matrix vectors to compare with medical image sources, contrasting various relevant and irrelevant characteristics based on original image sources with medical query picture. Challenges in retrieving medical images using various visual criteria, indexing, and clustering methods. For the purposes of this study, this is the primary issue statement for retrieving appropriate medical images from medical image databases [9]. A range of imaging modalities, such as CT, MRI, and X-ray, can be used to identify lung cancer. Due to decreased distortion and noise, the CT scan captures the features seen in distinct areas of the Lungs better than any other imaging modality, allowing radiologists to grasp and identify the occurrence of sickness [10].

Content-based picture retrieval has advanced significantly over the last decade, allowing for faster and more accurate image searches. Despite these advances, many issues remain unanswered. Semantic gap (occurs due to poor degree of pixel quality in feature presentation of pictures and also visual dimensional representation of image with varying index values) is the first challenge to extract data from multiple image sources. Some writers and academics have joined forces to find ways to close the semantic gap in picture retrieval. The large issue of the semantic gap in image retrieval may be broken down into a variety of smaller ones. Therefore, in this work, we single out such issues and provide adequate and practical remedies for them in the field of image retrieval [10]. The literature review is discussed in Part 2, the research methodology is outlined in Section 3, the study's findings and discussion are discussed in Part 4, and the study's conclusion and directions for future research are discussed in Part 5.

2. Literature review. Literature review in respect to Study of Secure Medical Image Retrieval Using Fast Image Processing Algorithms.

To improve picture recovery and recognizability evidence with content-based picture recognition, a method for extracting highlights through linearization of images is described in [11]. The developers tested their approach with 3688 images culled from two publicly available datasets. Regardless of the size of the image, this method reduced the number of highlights to 12. The factual measurements (with respect to correctness and review outcomes) were obtained for evaluation purposes. Misclassification of inquiry images might hinder the strategy's execution as compared to currently available alternatives for data recovery.

Band-based feature extraction and representation was suggested in [12]. If the image is altered, this method will dependably recover the data from the central (most important) objects. Fake neural networks were used for image recovery, with system performance and success measured using three publicly available informative indices (Coil, Corel, and Caltech 101), and recovery proficiency measured via exactness and review values.

Using statistical methods like Welch's t-tests and the F-ratio, [13] suggested a method for image recovery. The two completed visual information questions were reviewed for quality. While the full image is taken into account in the final product, the form is broken down into its component parts according to its orientation in the organized version. The F-ratio test is the first stage in the aforementioned procedure, with successful images moving on to the dynamic range test. The photographs were determined to be comparable if they passed both conditions. If nothing else, they are remarkable. The execution was approved and verified using a Mean Average Precision score. This enables us to create a system that isn't reliant on hand-crafted characteristics, which are typically necessary for other machine learning approaches.

So that surface and shading highlight extraction might have the same effect on CBIR, [14] developed a picture descriptor (Global Correlation Descriptor). The benefits of the structural component connection and insights from the histogram were included into the proposals for the Global Connection Vector and the Directional Global Correlation Vector, which are used to show surface and shading highlights, respectively. Approval was conducted on the Corel-10 K and Corel-5 K datasets, and performance was evaluated based on review and correctness.

In [15] a neighbourhood structure descriptor is proposed for picture recovery. Neighbourhood structure descriptor is made in light of the neighbourhood structures hidden hues; it has consolidated the shading, shape, and surface as one unit for recovery of pictures. Likewise, they proposed a calculation for include extraction which can separate nearby structure histogram utilizing neighbourhood structure descriptor.

In [16] proposed an approach known as picture recovery utilizing an intuitive hereditary calculation for figuring a high number of particular highlights at that point contrasting of related pictures for these highlights. The approach was tried on a gathering of 10,000 general pictures to demonstrate the effectiveness of the proposed approach.

In [17], a CBIR strategy was presented that combines Faster-Up Robust Features (SURF) and Scale Invariant Feature Transform (SIFT). Because SIFT is robust to rotation and scale shift and SURF is more robust to light fluctuations, depictions of these neighborhood highlights are used for recovery. The success of CBIR is enhanced when SURF and SIFT work together. All tests and evaluations were conducted on Corel-1500, Corel-2000, and Corel-1000 computers.

After settling on a visual list of capabilities, the next question is how to point them in the direction of precise image recovery. Over the last several decades, many novel architectural concepts have been put forward at the most fundamental level. Here, we will gloss over the techniques discussed in [18] and instead present a small subset of the more recent approaches.

In [19], a semantically-sensitive approach to content-based image recovery is presented. For effective image matching, a semantic organization (such diagram vs. photo vs. completed vs. non-textured) for extracting relevant elements is necessary, as is a location-based universal comparability measure. The speed with which this architecture can recover is crucial. Using region highlight bunching and the Most Similar Highest Priority (MSHP) guideline, the coordinating measure Integrated Region Matching (IRM) has been developed for faster recovery.

It has been proposed to use a highlight coordinating system for area-based picture recovery with the help of district codebooks and learned locale weights, and efforts have been made to fuse spatial similarity using the Hausdorff remove on limited measured point sets [20].

In [21], an alternative illustration is shown for protest recovery in chaotic images that does not need on

perfect division. A different approach to image restoration is area-based querying, which employs homogenous shading surface parts termed blobs. If a user recognizes at least one segmented blob as being similar to the concept "tiger," then her search may expand to include looking for tigers in other images, perhaps with different backgrounds. All things considered, this may lead to a more semantically accurate depiction of the client's inquiry objections, but it also involves more significant involvement from and dependence on her. Recovery may also be carried out without the client's explicit location marking for the purpose of locating images containing scaled or decrypted forms of inquiry items [22].

The use of multi-leveled perceptual collecting of primitive image highlights and their inter connections to characterize structure has been proposed as an alternative to picture division for the purpose of recovery [23].

Another idea, inspired by data compression and content-based methods, is to use vector quantization (VQ) on image squares to generate codebooks for depiction and recovery [24].

Protest-based image recovery using a windowed search over area and scale has been shown to be more effective than solutions based on erroneous division [25]. The client's inquiry Region-of-interest (ROI) is divided into rectangular parts for a coarser closer view/foundation, and then the frontal portions are used in a database search. Division is not fundamental for whole images. The Kullback-Leibler method for quantitatively analyzing models has been presented as a means of surface recovery through a combined presentation of highlight extraction and proximity estimate.

In [26], we find a proposal for yet another wavelet-based recovery method that makes advantage of striking focuses. It has been shown that image histograms based on fractal square codes are effective in recovering lost data from completed image databases.

In [27], it is explored how MPEG-7 content descriptors might be used to generate self-organizing maps (SOM) for the purpose of image recovery. A secure image recovery framework is one of the recent developments in the field. When tying things down, it's important to discover a group of agent "stay" images and choose the semantic proximity between a self-assertive picture match and these stays in terms of their comparability.

The evaluations for the standard photo recovery task are excellent. For literature published in the 1990s, please refer to [28]. While early frameworks saw widespread use of more elementary features like shade and surface, more advanced features like Significance and Scale Invariant Feature Transform (SIFT) have gained traction in recent years.

In this study, we choose the widely used Bag-of-Words (BoW) representation according to the neighborhood invariant SIFT features. The effectiveness of this component representation has been shown in a number of contexts. Since the focus of this study is on efficient research, this section provides an overview of the state of the art in terms of adept hunt systems, which may be roughly categorized into three groups: updated document, tree-based ordering, and hashing. It's still widely used for record recovery in the data recovery community that the changed file was initially offered [29].

BoW, for example, is quite similar to the sack of words representation of literary records, thus it was familiar with the area of image recovery. In this setup, a list of references to each record (image) for every content (visual) word is created, allowing for quick retrieval of relevant reports (images) in response to questions using just a few words. In any case, the written inquiries often include not very many words, which is a major difference between archive recovery and visual inquiry. For instance, Google online searches often only provide four-word answers.² In contrast to the BoW depiction, a single Medical Image may include several visual words, resulting in a large number of potentially useful images (from the revised data) that need pre-checking, a process often based on similarities to the original BoW highlights. Because of this, the utility of reorganized documents for a wide-ranging visual examination is severely limited. The number of applications may be reduced by increasing the visual vocabulary measure in BoW, which will also significantly raise memory use [30].

2.1. Research methodology. Here we offer a Novel Unsupervised Label Indexing (NULI) method for retrieving image labels, which is a term from the field of machine learning. In order to enhance the effectiveness of image retrieval frameworks, we describe machine learning as matrix convex optimization using cluster based matrix representation. Using the Search Based picture Annotation (SBIA) schema, we outline an empirical investigation on many different kinds of medical picture data sets, finding that our suggested technique provides superior outcomes.

2.2. Methodology. Web oriented medical image retrieval is the most efficient approach to handle processing of image with different structural analysis in real time medical healthcare systems. As World-Wide Web develops at a detonating rate, web crawlers end up noticeably imperative devices for any clients who look for data on the Internet, and web picture look is no special case. Web picture recovery has been investigated and created by scholastic analysts and business organizations; including scholastic models extra hunt measurement of existing web indexes.

The effective extraction of medical images from medical image sources has prompted the development of a number of machine learning-related methods. In real-time applications, such as various medical research identification of approximately matched medical pictures related to input query medical image, medical image annotation is a useful notion. Annotating medical images is a superior idea for retrieving near-perfect matches to a query medical picture. It takes a lot of time and effort to gather various sorts of label medical pictures from huge medical image data sets, which is why traditional medical image annotation systems were established.

Since a huge number of poorly labeled face medical photos are readily accessible on the World Wide Web (WWW), some recent research has attempted to develop an attractive search-based annotation design for facial medical picture annotation. The search-based medical image annotation (SBIA) design is meant to handle the automated face annotation process by utilizing content-based medical picture retrieval (CBIR) techniques, as opposed to coaching explicit classification designs by the standard model-based medical image annotation methods. The primary goal of the SBIA method is to properly align the input medical image's name labels. In particular, given a novel medical image for annotation, we first recover a narrow your search of top K most identical medical pictures from a weakly marked medical image data source, and then annotate the medical image by performing voting on appearance associated with the top K similar medical pictures. In this study, we offer a Novel Unsupervised Label Indexing (NULI) method for retrieving labels of medical pictures utilizing language from the field of machine learning so that we may access these characteristics in medical image retrieval from various medical image sources. The efficient image retrieval framework may be enhanced by defining machine learning as matrix convex optimization using cluster based matrix representation. Our experimental findings show improved performance compared to the status quo when it comes to real-time medical picture retrieval applications using traditional methods.

2.3. Semantic Signatures. Medical image re ranking in web based medical image retrieval with offline and online stages perform medical image reference classes operations to extract medical images automatically from different medical image pools. To avoid ambiguity in medical image query search retrieval from different medical image sources, our proposed approach follows semantic signatures for reference class verification to automatically retrieve medical images. Procedure of the semantic signatures presentation explained with following example.

For example implementation of N reference class labels from input query image q and then pre-processing those images based on sequential selection of trained medical images, multiple class reference classifiers on visual features of pictures are trained and then give M-dimensions vector p , which indicates the probability of newly generated picture I related to different class labels. P is used to describe the semantic image features of input query image Q and calculate the distance between each pixel from I_a and I_b with different pixel notations P^a and P^b .

$$d(I_a, I_b) = \|p^a - p^b\| \quad (2.1)$$

2.4. Separate Features. To separate different medical images based on image features are extracted and then pre-processed them using SVM classifier with following visual features like signature of the colour, Spatiality colour, wavelet pixel formation, invariant notation of histogram and gradient based histogram and GIST. Those medical images are characterized from different features like shape, colour and texture on combined M-dimensions.

A characteristic thought is to join a wide range of visual highlights to prepare a solitary intense SVM classifier which better recognizes diverse reference classes. In any case, the motivation behind utilizing semantic marks is to catch the visual substance of a picture, which may have a place with none of the reference classes, rather than ordering it into one of the reference classes. On the off chance that there are N sorts of autonomous visual highlights, it is in reality more successful to prepare to isolate SVM classifiers on various sorts of highlights

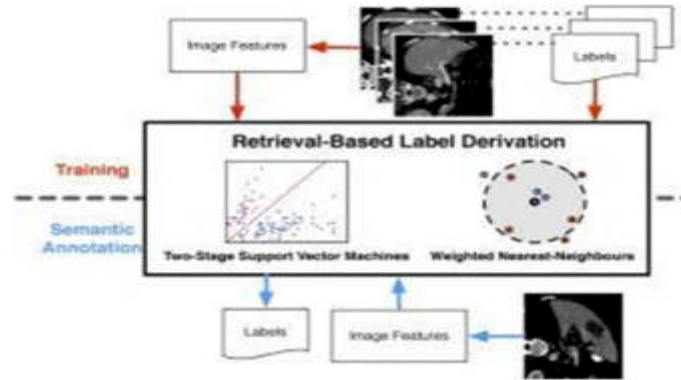


Fig. 2.1: Reference classes between input medical images with different dimensions.

and to consolidate the N semantic signatures $\{p^n\}^N$ = from n classifier present in N semantic signatures based on semantic features n 1 like shape, texture and dimension which describe different class labels in sharing of different images.

For instance, in Figure 2.1, given input as medical image relates to brain, liver with two reference class labels, if another image i.e. matched with source images based on verification of semantic signatures. If high amount of semantic signatures are matched with brain then retrieve relevant matched images with top matched brain images.

2.5. Medical image Reference Class Discovery. For each presented input query q , classify different image index labels. To describe this procedure, arrange images in sequential order i.e. $S(q)$ and describe the retrieved images based on index using query q describe the view of literary data. Query expansion found from different sources with sequential presentation of referable class labels with visual substance of image representation. Also contain subset of images explored with different instances matched with pixels of image at different dimensions.

Every idiom expansion e is used to obtain photos from the internet searcher and matched top $-k$ results with matching semantic keyword expansion for automated learning and retrieval of training medical images with reference classes. First the keyword q , retrieval relevant images sub sequentially extract relevant images matched by q . to find the similarity between images, use k-means clustering approach which are the images consists referable class labels. In this methodology similar group of average matched images are arranged in cluster and describe the exceptions from original images from medical image sources.

Here represents some of the keyword expansions like brain, liver and different keywords which consists identical and semantic visual features to increase the performance. To increase the computational cost in representation of image with different discriminative spaces between pixels in image. Estimate the referable class label index to learn parameters using SVM classifier to classify data into specified data relates to keyword with different pixel notations to find relevancy.

The first two basic training phases, Ai1 and Ai2, are used to extract m reference classes from the preceding procedures. Reference classes $D(i, j)$ will be obtained by using SVM classification learned from Ai1 and Ai2 to distinguish between reference classes i and j . The SVM calculates the likelihood of classifier score for the i class for each reference class. $D(i, j) = h((p_i + p_j)/2)$, where h is an increasing function, if the average score of Ai1 is p_i and the average score of the PJ across A2j is also determined. The production of a single binary bit is described as follows:

$$h(p_i) = 1 - e^{-1}(p_j) \quad (2.2)$$

While and remain fixed, the ratio of $(p_i + p_j)/2$ goes up as $h(p_i)$ goes down, and this trend holds true across all meaningful reference classes.

We describe the different referral class labels from n no. of users. Keyword expansion is used to extract reference class labels which explore mostly matched results with input image. Meanwhile, choose different referable class labels with different functions based on expansion of keyword. Distinct matrix $m \times n$ represents and its procedure in next sections with different parameters. Qualifying measures to solve optimization of image annotation based referable class label representation.

3. Medical image Indexing Implementation. This section describes general implementation NULI for accessing relevant images based on initial sequence factors in image retrieval in medical sources, conversation about problem development in medical picture annotation, criteria execution to catalog medical picture annotations, approximation collection process on function removal to determine medical picture recovery.

We describe $X = md$ is explored different medical features which consist different dimensions with pixels. $= m_1, m_2, \dots, m_n$ defines image labels with annotated pixel representations, m is the label of image. be the labelled matrix which consists weak label data which presents i th and j th rows and columns represents in sequential pixel Y_i formation of medical image $Y = [1, 0]^{m \times n}$. In NULI, individual medical query image from image source to gather relevant images based on label index.

Sequential matrix with class labels is used to illustrate the NULI method. Different values for x and y indicate the content of the label in matrix y . Use convex optimization based on the important features of the class labels to efficiently index photos for labelling. These procedures are used to achieve optimum performance when retrieving medical images based on relevance:

$$E_S(F, W) = \frac{1}{2} \sum_{i,j=1}^n W_{ij} \|F_{i^*} - F_{j^*}\|_F^2 = \text{tr}(F^T L F) \quad (3.1)$$

Matrix weight measure i.e., which comprises of optimal functions to build both "normal" and "fantastic" weight matrices. The following is a description of the representation of matrix regulations:

$$F^* = \arg \min_{F \geq 0} E_s(F, W) + \alpha \cdot E_p(F, Y) \quad (3.2)$$

Non-zero regulatory elements based on feature dimensions are specified by the following matrix parameter:

$$E_p(F, Y) = \|F_{i^*} - F_{j^*} \cdot S\|_F^2 \quad (3.3)$$

We supply an effective label index for each picture so that the development of a matrix of functions may be achieved. To maximize the evaluation in terms of accuracy and recall and others in real-time image processing applications, we will discuss the implementation of the NULI approach, the index label representation for various images, and the automated picture annotation.

3.1. Performance of Experiments. This section user interface implementation procedure of proposed NULI with hybrid approach in re-rank based image retrieval from image pools. Medical image sets are collected from different search engines defined in table 3.1. It describes medical image search engines and sample keywords with how many medical images retrieved from search engines with different keywords. photos. Anisotropic diffusion reduces noise while preserving significant sections of a picture. The lung region is segmented using morphological erosion.

As shown in table 3.1, discuss three publicly available data sets to test performance of proposed approach at various representations. Google image search contain 1000-10000 images to arrange in re-rank with searching relates to search optimizations. This search images spread search procedures with different objects like pixel dimension, calculation of pixel length, time, image patch and sequentially representation of image simultaneously. Data set 2 arrange results in re-rank procedure extracted images from image search of the label 2. Images are combined and get image data from Google search engine at different time frames, re-positioning based image retrieval with different label formation check whether it is present or not. As shown in table 3.1, collect medical images and then semantic signatures with reference classes labelled with different presentation shown in figure 3.1.

Table 3.1: Medical image data sets description with different search engines

Medical image Collection Procedure			
Medical image Data sets	Keywords	Medical images	Search Engine
I	50	1000	Bing Medical image Search
II	50	1000	
III	20	500	Bing Medical image Search



Fig. 3.1: Relevant medical image for input query medical image based on visual semantic features.

4. Results and discussion.

Data sets. This section describes the set up environment of efficient image retrieval for different real time medical image processing applications. Different medical images taken from <http://www.imdb.com> URL which consist different feature related images, those images consists different labels i.e. data, path and image name. User search images based on name of an image which consist different label procedures. Based on collected data, using JAVA and Net beans software construct search engine to retrieve relevant images automatically whenever input image query matched with source image based on visual features in real time medical image network system. Using a variety of test cases, this section compares and contrasts the NULI method with the conventional method, i.e. a hybrid image retrieval framework. To demonstrate NULI’s efficacy in medical picture retrieval, this illustration contrasts NULI’s performance with that of the conventional method across a variety of metrics, including precision, recall, accuracy, and time. The following is a description of the quantitative analysis used to obtain medical pictures from several medical image sources:

$$\begin{aligned}
 precision &= \frac{No.of\ relevant\ images\ retrieved}{Total\ no.\ of\ images\ retrivd} \\
 Recall &= \frac{No.of\ relevant\ images\ retrieved}{Total\ no.\ of\ relevant\ images\ in\ database} \\
 Accuracy &= 2 \frac{precision * recall}{precision + recall}
 \end{aligned}$$

To obtain weak label medical pictures from many medical image sources, medical image sources include various medical images with various criteria such as labels and features. The findings of the NLUI methodology provide greater accuracy with weak label indexing of each picture from diverse image sources than the traditional approaches and procedures done on medical sources to obtain efficient images. Accuracy compared to conventional methods is shown in Figure 4.1.

Table 4.1: Average error rate and average computation time

Iterations	Average error rate			Average consumption time (sec)		
	Precision	Recall	False positive	F-Measure	Miss Rate	False negative
10	0.5375	0.4	0.35	0.018162	0.012237	0.014621
20	0.4	0.35	0.275	0.019874	0.014943	0.010203
30	0.3875	0.325	0.3	0.021312	0.011617	0.013204
40	0.375	0.3	0.325	0.022683	0.010491	0.014435
50	0.375	0.3	0.3	0.023986	0.01065	0.010167

Table 4.2: Error rate deviation and computation time deviation

Iterations	Error rate deviation			Computation time deviation(sec)		
	Precision	Recall	False positive	F-Measure	Miss Rate	False negative
10	0.158607	0.229129	0.122474	0.004236	0.00342	0.006892
20	0.122474	0.122474	0.075	0.001454	0.009629	0.000887
30	0.117925	0.114564	0.1	0.002176	0.003505	0.008579
40	0.136931	0.1	0.114564	0.003372	0.000816	0.011351
50	0.125	0.1	0.1	0.004663	0.000629	0.001296

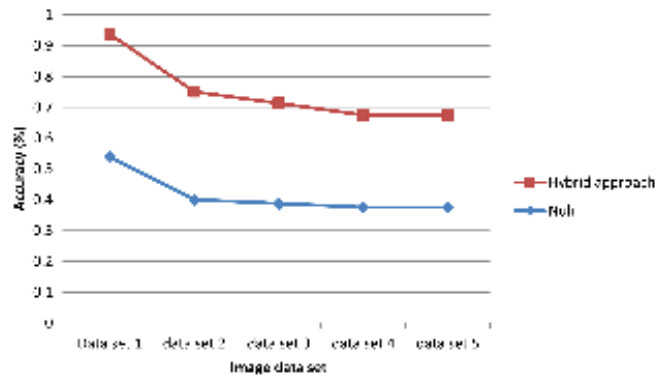


Fig. 4.1: Different types of medical imaging accuracy

For both NULI and conventional methods, i.e. the hybrid strategy shown in Figure 4.2, accuracy is the primary metric for efficiently retrieving similarly matched images from medical sources.

In healthcare related image retrieval related application, recall for weak label image retrieval from medical image sources described in Figure 4.3.

Figure 4.4 shows the accuracy presentation of proposed approach with different image databases.

5. Conclusion and future work. Discussed in this article is a research proposal entitled "Novel Unsupervised Label Indexing for Efficient Image Retrieval from Medical Image Sources Based on Label Indexing Using a Re-rank Process Established Using a Search-Based Annotation Methodology." Various image notations are used to describe the arrangement of pictures in a convex optimization representation of image pixels. The findings indicate that the characteristics derived from the deep learning model point in the direction of developing an efficient CBIR system. This study will be extended in the future by training on a real-time dataset. The processing of various keywords for medical picture retrieval presentations is shown to improve accuracy, recall, and time efficiency in experimental findings. Weak label medical image classification is a potential future feature in medical image retrieval from all medical image sources, together with the further advancement of

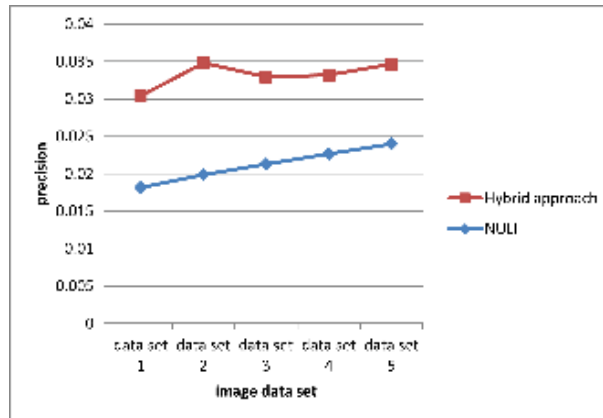


Fig. 4.2: Precision of different medical image with different techniques

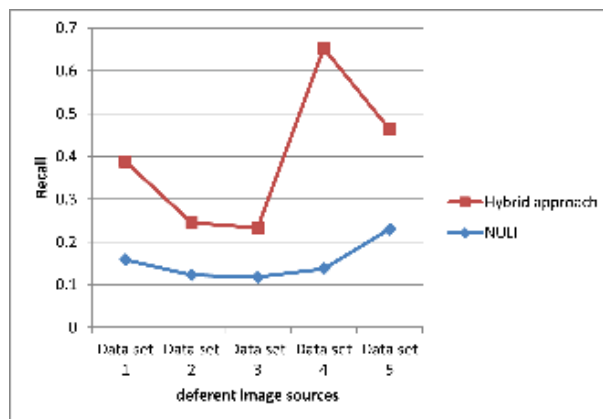


Fig. 4.3: Recall values of different medical images with different data sets

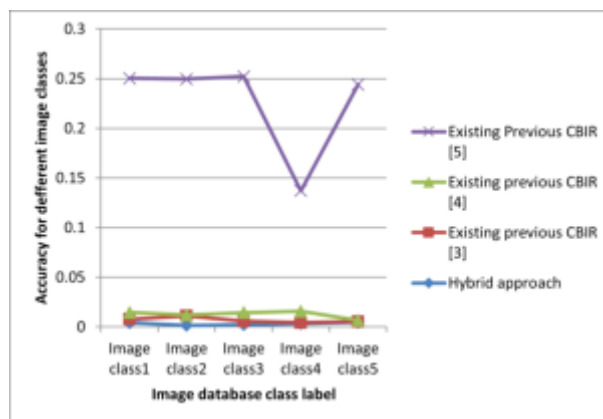


Fig. 4.4: Accuracy values with respect to different image database

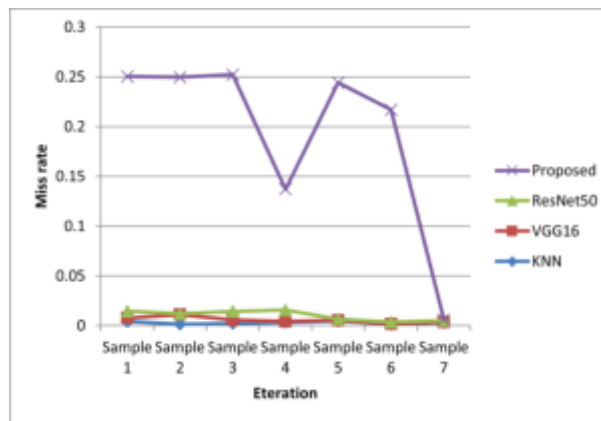


Fig. 4.5: Miss Rate Analysis of various Proposed Algorithms

effective CBIR from multiple medical image sources. Our suggestions for the future of medical image retrieval centre on the use of weak label generation to improve accuracy, recall, and throughput.

REFERENCES

- [1] CHENG, B., ZHUO, L., BAI, Y., *Secure Index Construction for Privacy-Preserving Large-Scale Image Retrieval*, In Proceedings of the 2014 IEEE Fourth International Conference on Big Data and Cloud Computing, Sydney, NSW, Australia, 4 December 2014, pp. 116120.
- [2] FERREIRA, B., RODRIGUES, J., LEITAO, J., DOMINGOS, H., *Practical Privacy-Preserving Content-Based Retrieval in Cloud Image Repositories*, IEEE Trans. Cloud Comput., vol. 7, pp. 784798, 2019
- [3] XIA, Z., XIONG, N.N., VASILAKOS, A.V., SUN, X., *EPCBIR: An efficient and privacy-preserving content-based image retrieval scheme in cloud computing*, Inf. Sci., vol. 387, pp. 195204, 2017.
- [4] ZHU, X., LI, H., GUO, Z., *Privacy-preserving query over the encrypted image in cloud computing*, J. XiDian Univ., vol. 41, pp.151158, 2014.
- [5] IBRAHIM, A., JIN, H., YASSIN, A.A., ZOU, D., XU, P., *Towards Efficient Yet Privacy-Preserving Approximate Search in Cloud Computing*, Comput. J., vol. 57, pp. 241254, 2014.
- [6] FAN, K., WANG, X., SUTO, K., LI, H., YANG, Y., *Secure and Efficient Privacy-Preserving Ciphertext Retrieval in Connected Vehicular Cloud Computing*, IEEE Netw., vol. 32, pp. 5257, 2018.
- [7] XIA, Z., WANG, X., ZHANG, L., QIN, Z., SUN, X., REN, K., *A Privacy-Preserving and Copy-Deterrence Content-Based Image Retrieval Scheme in Cloud Computing*, IEEE Trans. Inf. Forensics Secur., vol. 11, pp. 25942608, 2016.
- [8] FENG, W., HE, Y., *Cryptanalysis and Improvement of the Hyper-Chaotic Image Encryption Scheme Based on DNA Encoding and Scrambling*, IEEE Photon. J., vol. 10, pp. 115, 2018.
- [9] ZHU, Z.-L., ZHANG, W., WONG, K.-W., YU, H., *A chaos-based symmetric image encryption scheme using a bit-level permutation*, Inf. Sci., vol. 181, pp. 11711186, 2011.
- [10] RAVICHANDRAN, D., PRAVEENKUMAR, P., RAYAPPAN, J.B.B., AMIRTHARAJAN, R., *PChaos based crossover and mutation for securing DICOM image*, Comput. Boil. Med., 72, pp. 170184, 2016.
- [11] CHEN, J.-X., ZHU, Z.-L., FU, C., YU, H., ZHANG, L.-B., *A fast chaos-based image encryption scheme with a dynamic state variables selection mechanism*. Commun, Nonlinear Sci. Numer. Simul., vol. 20, pp. 846860, 2015.
- [12] CHAI, X., CHEN, Y., BROYDE, L., *A novel chaos-based image encryption algorithm using DNA sequence operations*, Opt. Lasers Eng., vol. 88, pp. 197213, 2017.
- [13] XU, L., LI, Z., LI, J., HUA, W., *A novel bit-level image encryption algorithm based on chaotic maps*, Opt. Lasers Eng., vol. 78, pp. 1725, 2016.
- [14] WANG, J., LONG, F., *CNN-based colour image encryption algorithm using DNA sequence operations*, In Proceedings of the 2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), Shenzhen, China, 1517 December 2017, pp. 730736.
- [15] ENAYATIFAR, R., ABDULLAH, A.H., ISNIN, I.F., ALTAMEEM, A., LEE, M., *Image encryption using a synchronous permutation-diffusion technique*, Opt. Lasers Eng., vol. 90, pp. 146154, 2017.
- [16] LIU, W., SUN, K., ZHU, C., *A fast image encryption algorithm based on chaotic map*, Opt. Lasers Eng., vol. 84, pp. 2636, 2016.
- [17] DATAR, M., IMMORLICA, N., INDYK, P., MIRROKNI, V.S., *Locality-sensitive hashing scheme based on p -stable distributions*, In Proceedings of the 20th Annual Symposium on Computational Geometry, Brooklyn, NY, USA, 911 June 2004, pp. 253262.

- [18] XIA, P., PAN, Y., LAI, H., LIU, C., YAN, S., *Supervised hashing for image retrieval via image representation learning*, In Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence(AAAI), Québec City, QC, USA, 2731 July 2014, pp. 21562162.
- [19] SHEN, F., SHEN, C., LIU, W., SHEN, H.T., *Supervised Discrete Hashing*, In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 712 June 2015, pp. 3745.
- [20] LI, W.-J., WANG, S., KANG, W.-C., *Feature Learning Based Deep Supervised Hashing with Pairwise Labels*, IJCAI: New York, NY, USA, pp. 32703278, 2016.
- [21] LI, Q., SUN, Z., HE, R., TAN, T., *Deep supervised discrete hashing*, In Advances in Neural Information Processing Systems, NIPS: Long Beach, CA, USA, pp. 24822491, 2017.
- [22] LI, N., LI, C., DENG, C., LIU, X., GAO, G., *Deep joint semantic-embedding hashing.*, In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), Stockholm, Sweden, 1319 July 2018, pp. 23972403.
- [23] JIANG, Q., CUI, X., LI, W., *Deep Discrete Supervised Hashing*, IEEE Trans. Image Process, vol. 27, pp. 59966009, 2018.
- [24] NAZARIMEHR, F., RAJAGOPAL, K., KENGNE, J., JAFARI, S., PHAM, V.T., *A new four-dimensional system containing chaotic or hyper-chaotic attractors with no equilibrium, a line of equilibria and unstable equilibria.* Chaos Solitons Fractals, vol. 111, pp. 108118, 2018.
- [25] ZHANG, Y., ZHANG, Q., LIAO, H., WU, W., LI, X., NIU, H., *A Fast Image Encryption Scheme Based on Public Image and Chaos*, In Proceedings of the 2017 International Conference on Computing Intelligence and Information System (CIIS), Nanjing, China, 2123 April 2017, pp. 270276.
- [26] DESHMUKH, P., KOLHE, V., *Modified AES based algorithm for MPEG video encryption*, In Proceedings of the International Conference on Information Communication and Embedded Systems (ICICES2014), Chennai, India, 2728 February 2014, pp. 15.
- [27] CISSE, I.I., KIM, H., HA, T., *A rule of seven in Watson-Crick base-pairing of mismatched sequences*, Nat. Struct. Mol. Biol., vol. 19, pp. 623627, 2012.
- [28] ZHANG, X.Q., WANG, X.S., *A Multiple-image encryption algorithm based on DNA encoding and chaotic system*, Multimed. Tools Appl., vol. 77, pp. 129, 2018.
- [29] WANG, X.Y., ZHANG, Y.Q., BAO, X.M., *A novel chaotic image encryption scheme using DNA sequence operations*, Opt. Lasers Eng., vol. 73, pp. 5361, 2015.
- [30] MURAT, H., ZHANG, S., YAN, C., *Classification of Xinjiang Uygur medicine image based on KNN Classifier*, J. Xinjiang Med. Univ., vol. 38, pp. 800804, 2015.

Edited by: Mustafa M Matalgah

Special issue on: Synergies of Neural Networks, Neurorobotics, and Brain-Computer Interface Technology: Advancements and Applications

Received: Jan 2, 2024

Accepted: Mar 28, 2024



ON SOFT STRONGLY B^* –COMPACTNESS AND SOFT STRONGLY B^* –CONNECTEDNESS IN SOFT TOPOLOGICAL SPACES

SAIF Z. HAMEED* ABDELAZIZ E. RADWAN† AND ESSAM EL-SEIDY‡

Abstract. In this research article, we present a new class of soft compact spaces and soft Lindelöf spaces, we identify the idea of soft strongly b^* –compact and soft strongly b^* –Lindelöf spaces and we supply multiple interesting examples. As well as we mention that the inaugurated spaces are conserved under soft strongly b^* –irresolute mappings and we look into definite of results which connect an extensive soft topology with the showing soft spaces. As well as we inquiry the features and attributive of soft strongly b^* –connected spaces and discuss and identify its relationship with soft connectedness.

Key words: soft strongly b^* –closed set, soft strongly b^* –open set, soft strongly b^* –compact, soft strongly b^* –Lindelöf spaces, soft strongly b^* –connected space

1. Introduction and Preliminaries. Molodtsov [1] used an acceptable parametrization. In 1999, he introduced the soft set theorem’s basic idea and disclosed the theorem’s first result. He had many experimenters working on the proposal. Topology is eminent in colorful divaricate of mathematics. Therefore, Shabir and Naz [2] were the pioneers who introduced the concept of soft topological spaces. Kannan [3] assigned soft generalized closed and soft generalized open sets in soft topological spaces. Akdag and Ozkan ([4], [5]) presented a conception of soft α –open, the soft b –open, and their respective continuous functions. Zorlutuna et al. inquiry soft interior point and soft neighbourhood and he first examined the compactness of soft topological spaces [6]. Connectedness [7] is an effective tool for topology introduced by Porter J. and Woods R.. Hussain [8] assigned and take a look at the features of soft connected space. Saif Z. et al. [9] introduced the soft bc –open set. The soft b^* –closed are introduced by Hameed, Saif Z. et al. [10]. Soft b^* –continuous functions, soft strongly b^* –closed and soft strongly b^* –continuous functions are studied by Hameed, Saif Z. et al. [11], [12].

In the present work, we define the soft strongly b^* –compact and soft strongly b^* –Lindelöf spaces. Also, we introduce the soft strongly b^* –connected spaces. The details of the properties, examples, and counterexamples that substantiate the concept are thoroughly discussed.

In this study, consider \mathcal{W} as an initial universe and $P(\mathcal{W})$ as the power set of \mathcal{W} . In addition, $\tilde{E} \neq \phi$ stands for the family of parameters that are being considered and $\phi \notin \varphi \subseteq \tilde{E}$.

DEFINITION 1.1. [1] (Ψ, φ) is referred to be a soft set over \mathcal{W} if Ψ is a map from φ to $P(\mathcal{W})$.

DEFINITION 1.2. [13] The soft set $(\mathcal{S}, \varphi) \in \mathcal{SS}(\mathcal{W}, \varphi)$, where $S(\nabla) = \phi$, for every $\nabla \in \varphi$ is stated A-null soft set of $\mathcal{SS}(\mathcal{W}, \varphi)$ and symbolize by $\tilde{\phi}$. The soft set $(\mathcal{S}, \varphi) \in \mathcal{SS}(\mathcal{W}, \varphi)$, where $S(\nabla) = \mathcal{W}$, for every $\nabla \in \varphi$ is stated the A-absolute soft set of $\mathcal{SS}(\mathcal{W}, \varphi)$ and symbolize by $\tilde{\mathcal{W}}$.

DEFINITION 1.3. [13] For two sets $(\Psi, \varphi), (\mathcal{S}, \Theta) \in \mathcal{SS}(\mathcal{W}, \varphi)$, then (Ψ, φ) is a soft subset of (\mathcal{S}, Θ) symbolize by $(\Psi, \varphi) \subseteq (\mathcal{S}, \Theta)$, if

1. $\varphi \subseteq \Theta$.
2. $\psi(\nabla) \subseteq S(\nabla), \forall \nabla \in \varphi$.

Then, (Ψ, φ) is stated to be a soft superset of (\mathcal{S}, Θ) , if (\mathcal{S}, Θ) is a soft sub-set of (Ψ, φ) , $(\mathcal{S}, \Theta) \subseteq (\Psi, \varphi)$.

DEFINITION 1.4. [2] Let (Ψ, φ) be soft set over \mathcal{W} , $z \in \mathcal{W}$. that’s what we call $z \in (\Psi, \varphi)$, whenever $z \in \psi(\nabla)$ for all $\nabla \in \varphi$. The soft set (Ψ, φ) over \mathcal{W} such that $\psi(\nabla) = \{z\}, \forall \nabla \in \varphi$ is stated singleton soft point and symbolize by z_φ or (z, φ) .

*Department of Mathematics, College of Education, Mustansiriyah University, Baghdad, Iraq (saif.zuhar.edbs@uomustansiriyah.edu.iq),

†Department of Mathematics, Faculty of Science, Ain Shams University, Cairo, Egypt (zezoradwan@yahoo.com).

‡Department of Mathematics, Faculty of Science, Ain Shams University, Cairo, Egypt (esam_elsedy@hotmail.com).

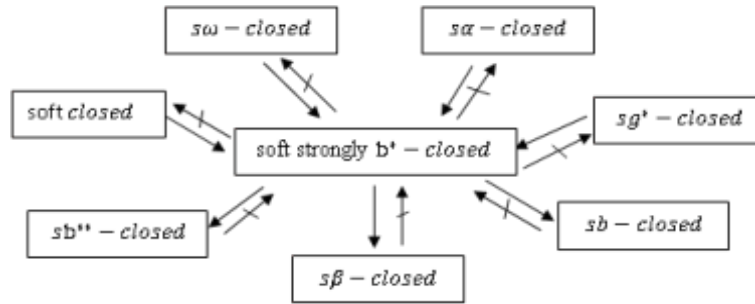


Fig. 1.1: Relationships of soft strongly b^* -closed

DEFINITION 1.5. [2] Let $\mathcal{Q} \subseteq \mathcal{SS}(\mathcal{W}, \wp)$. Then \mathcal{Q} is stated to be soft topological space (STS) if

1. $\tilde{\phi}$ and \tilde{W} belong to \mathcal{Q} .
2. Arbitrary unions of members \mathcal{Q} belongs to \mathcal{Q} .
3. Finite intersections of members \mathcal{Q} belongs to \mathcal{Q} .

It is symbolize by $(\mathcal{W}, \mathcal{Q}, \wp)$ (briefly \mathcal{W}).

DEFINITION 1.6. [2] Let $(\mathcal{W}, \mathcal{Q}, \wp)$ be a STS over \mathcal{W} , then the organ of \mathcal{Q} are stated to be soft open sets in \mathcal{Q} .

DEFINITION 1.7. [2] Let $(\mathcal{W}, \mathcal{Q}, \wp)$ be a STS over \mathcal{W} . A soft set (Ψ, \wp) over \mathcal{W} is stated to be a soft closed set in \mathcal{W} , if its relative complement (Ψ, \wp) belongs to \mathcal{Q} .

DEFINITION 1.8. [6] Let $(\mathcal{W}, \mathcal{Q}, \wp)$ be a STS and $(\Psi, \wp) \in \mathcal{SS}(\mathcal{W}, \wp)$. Then

1. The soft closure of (Ψ, \wp) is the soft set $cl(\Psi, \wp) = \cap\{(\mathcal{S}, \wp) : (\mathcal{S}, \wp) \in \mathcal{Q}^c, (\Psi, \wp) \subseteq (\mathcal{S}, \wp)\}$.

2. The soft interior of (Ψ, \wp) is the soft set $int(\Psi, \wp) = \cup\{(\mathcal{S}, \wp) : (\mathcal{S}, \wp) \in \mathcal{Q}, (\mathcal{S}, \wp) \subseteq (\Psi, \wp)\}$.

DEFINITION 1.9. A soft set (Ψ, \wp) of a STS $(\mathcal{W}, \mathcal{Q}, \wp)$ is stated to be

1. soft α -open [4] if $(\Psi, \wp) \subset int(cl(int((\Psi, \wp))))$,
2. soft pre-open [14] if $(\Psi, \wp) \subset int(cl((\Psi, \wp)))$,
3. soft semi-open [15] if $(\Psi, \wp) \subset cl(int((\Psi, \wp)))$,
4. soft β -open [14] if $(\Psi, \wp) \subset cl(int(cl((\Psi, \wp))))$,
5. soft b -open [5] if $(\Psi, \wp) \subset int(cl((\Psi, \wp))) \cup cl(int((\Psi, \wp)))$.

DEFINITION 1.10. [16] A soft set (Ψ, \wp) is called soft ω -closed in a STS $(\mathcal{W}, \mathcal{Q}, \wp)$, if $cl(\Psi, \wp) \subseteq (\mathcal{S}, \wp)$ whenever $(\Psi, \wp) \subseteq (\mathcal{S}, \wp)$ and (\mathcal{S}, \wp) is soft semi-open set in \mathcal{W} . The relative complement of (Ψ, \wp) is called soft ω -open in \mathcal{W} .

DEFINITION 1.11. [12] A soft set (Ψ, \wp) of a STS $(\mathcal{W}, \mathcal{Q}, \wp)$ is called a soft strongly b^* -closed (briefly sSb^* -closed) if $cl(int(\Psi, \wp)) \subseteq (\mathcal{S}, \wp)$, whenever $(\Psi, \wp) \subset (\mathcal{S}, \wp)$ and (\mathcal{S}, \wp) is sb -open. The complement of a sSb^* -closed set is stated to be sSb^* -open set.

THEOREM 1.12. [12] The following statements are correct:

1. Every soft open is sSb^* -open.
2. Every $s\alpha$ -open is sSb^* -open.
3. Every sSb^* -open set is sb -open.
4. Every $s\omega$ -open is sSb^* -open.

DEFINITION 1.13. [12] Let $(\mathcal{W}, \mathcal{Q}, \wp)$ be a STS. a subset $(\Psi, \wp) \subseteq \mathcal{W}$ is called a soft strongly b^* -neighbourhood (briefly sSb^* -nbd) of point $\nu \in \mathcal{W}$ if \exists an sSb^* -open set (Ψ, \wp) where $\nu \in \mathcal{W} \subseteq (\Psi, \wp)$.

DEFINITION 1.14. [12] Let $(\mathcal{O}, \wp) \in \mathcal{SS}(\mathcal{W}, \wp)$. Then $sSb^*int(\mathcal{O}, \wp) = \cup\{(\mathcal{L}, \wp) : (\mathcal{L}, \wp) \text{ is a } sSb^*\text{-open set and } (\mathcal{L}, \wp) \subset (\mathcal{O}, \wp)\}$.

DEFINITION 1.15. [12] Let $(\mathcal{L}, \wp) \in \mathcal{SS}(\mathcal{W}, \wp)$. Then $sSb^*cl(\mathcal{L}, \wp) = \cap\{(\Psi, \wp) : (\Psi, \wp) \text{ is a } sSb^*\text{-closed set and } (\mathcal{L}, \wp) \subset (\Psi, \wp)\}$.

DEFINITION 1.16. [12] A soft mapping $\Pi : \mathcal{W} \rightarrow \Sigma$, from $STS (\mathcal{W}, \mathcal{Q}, \wp)$ into $STS (\Sigma, \Omega, \Theta)$, is stated to be soft strongly b^* -continuous (briefly sSb^* -continuous) if the inverse image of every soft open set in Σ is a sSb^* -open set in \mathcal{W} .

DEFINITION 1.17. [12] A soft mapping $\Pi : \mathcal{W} \rightarrow \Sigma$ is stated to be soft strongly b^* -irresolute (briefly sSb^* -irresolute) if the inverse image of every sSb^* -closed set in Σ is a sSb^* -closed set in \mathcal{W} .

For are details, we refer to [12], [6], [7].

2. Soft strongly b^* -compact spaces. In this section, We offer the conception of soft strongly b^* -compact and soft strongly b^* -Lindelöf spaces and The significant structural properties.

DEFINITION 2.1. A collection $\{(\psi_\epsilon, \wp) : \epsilon \in \zeta\}$ of soft strongly b^* -open sets is called a soft strongly b^* -open cover of $(\mathcal{W}, \mathcal{Q}, \wp)$, if $\widetilde{\mathcal{W}} = \bigcup_{\epsilon \in \zeta} (\psi_\epsilon, \wp)$.

DEFINITION 2.2. A $STS (\mathcal{W}, \mathcal{Q}, \wp)$ is called soft strongly b^* -compact (resp. soft strongly b^* -Lindelöf), if each sSb^* -open cover of $\widetilde{\mathcal{W}}$ has a finite (resp. countable) soft subcover of $\widetilde{\mathcal{W}}$.

DEFINITION 2.3. A soft subset $(, \wp)$ of a $STS (\mathcal{W}, \mathcal{Q}, \wp)$ is called soft strongly b^* -compact in \mathcal{W} determined by for every collection $\{(\psi_\epsilon, \wp) : \epsilon \in \zeta\}$ of soft strongly b^* -open sets of \mathcal{W} where $(, \wp) \subset \cup\{(\psi_\epsilon, \wp) : \epsilon \in \zeta\} \exists$ finite subset ζ_0 of ζ where $(, \wp) \subset \cup\{(\psi_\epsilon, \wp) : \epsilon \in \zeta_0\}$

DEFINITION 2.4. A $STS (\mathcal{W}, \mathcal{Q}, \wp)$ is called soft strongly b^* -space if every sSb^* -open set of \mathcal{W} is soft open set in \mathcal{W} .

COROLLARY 2.5. If $STS (\mathcal{W}, \mathcal{Q}, \wp)$ is a sSb^* -compact space and soft strongly b^* -space, then \mathcal{W} is soft compact space.

Proof. Assume that $\{(\psi_\epsilon, \wp) : \epsilon \in \zeta\}$ be soft open cover of \mathcal{W} . For each soft open set is sSb^* -open set, $\{(\psi_\epsilon, \wp) : \epsilon \in \zeta\}$ is sSb^* -open cover of \mathcal{W} . For \mathcal{W} is sSb^* -compact space and sSb^* -space, \exists finite subset ζ_0 of ζ where $\mathcal{W} \subset \{(\psi_\epsilon, \wp) : \epsilon \in \zeta_0\}$. Therefore, \mathcal{W} is soft compact space. \square

COROLLARY 2.6. If $\Pi : \mathcal{W} \rightarrow \Sigma$ is a sSb^* -continuous function and sSb^* -space, then Π is soft continuous function.

Proof. Assume $\{(\psi_\epsilon, \wp) : \epsilon \in \zeta\}$ be soft open set of Σ . whereas Π is sSb^* -continuous, $\{\Pi^{-1}((\psi_\epsilon, \wp)) : \epsilon \in \zeta\}$ is sSb^* -open set of \mathcal{W} and whereas \mathcal{W} is sSb^* -space, $\{\Pi^{-1}((\psi_\epsilon, \wp)) : \epsilon \in \zeta\}$ forms soft open set of \mathcal{W} . Thus, Π is soft continuous. \square

COROLLARY 2.7. Assume $(\mathcal{W}, \mathcal{Q}, \wp)$ be STS . If $(\mathcal{W}, \mathcal{Q}_\nabla)$ is a sSb^* -compact space, for each $\nabla \in \wp$, then $(\mathcal{W}, \mathcal{Q}, \wp)$ is a sSb^* -compact space.

Proof. Assume that $\wp = \{\nabla_1, \nabla_1, \dots, \nabla_n\}$ be a set of parameter and $(\mathcal{W}, \mathcal{Q}_\nabla)$ is sSb^* -compact space, for each $\epsilon = \overline{1, n}$. Suppose $\{(\psi_\epsilon, \wp) : \epsilon \in \zeta\}$ be sSb^* -open cover of \mathcal{W} . Since $\cup_{\epsilon \in \zeta} (\psi_\epsilon, \wp)(\nabla) = \widetilde{\mathcal{W}}$, for each $\nabla \in \wp$, and $(\mathcal{W}, \mathcal{Q}_\nabla)$ is a sSb^* -compact, \exists finite subset ζ_0 of ζ where $\cup_{\epsilon \in \zeta_0} (\psi_\epsilon, \wp)(\nabla) = \widetilde{\mathcal{W}}$. Hence, $\{(\psi_\epsilon, \wp) : \epsilon \in \zeta_0\}$ is a finite subcover of $\{(\psi_\epsilon, \wp) : \epsilon \in \zeta\}$. Hence, $(\mathcal{W}, \mathcal{Q}, \wp)$ is a sSb^* -compact space. \square

COROLLARY 2.8. Every sSb^* -compact (resp. sSb^* -Lindelöf) space is soft compact (resp. soft Lindelöf).

In the next example, indicates that the inclusions of the Corollary 2.8 is not necessarily correct.

Example 1. Consider $\wp = Q^c$ is the set of irrational numbers. Let $\mathcal{Q} = \{\widetilde{\phi}, \widetilde{\mathcal{W}}, (, \wp)$ when $(\nabla) = \{1\}, \forall \nabla \in \wp$ be a STS on $\mathcal{W} = \{1, 2\}$. clearly, $(\mathcal{W}, \mathcal{Q}, \wp)$ is soft compact. furthermore, a family $\{(\delta, \Theta) : \delta(v) = \{1\}, \forall v \neq \nabla\}$ is a sSb^* -open cover of $\widetilde{\mathcal{W}}$. For has not a soft countable subcover of $\widetilde{\mathcal{W}}$. Thus, $(\mathcal{W}, \mathcal{Q}, \wp)$ is not a sSb^* -Lindelöf space.

THEOREM 2.9. Every sSb^* -compact space is a sSb^* -Lindelöf.

Proof. Clear. \square

In the next example, indicates that the inclusions of the Theorem 2.9 and Figure 2.1 is not necessarily correct.

Example 2. Let $\mathcal{Q} = \{\widetilde{\phi}, \widetilde{\aleph}, (\delta, \wp)\}$ and $\wp = \{\nabla_1, \nabla_1, \dots, \nabla_n\}$. such that $\delta(\nabla) = \{1\}, \forall \nabla \in \wp$ be a STS on the set of natural numbers \aleph . Since \wp and \aleph are soft countable, then $(\aleph, \mathcal{Q}, \wp)$ is a sSb^* -Lindelöf. furthermore, a family $\{(\mathcal{S}, \Theta) : \mathcal{S}(v) = \{1, x\}, \text{ for each } v \in \Theta, x \in \aleph\}$ is a sSb^* -open cover of $\widetilde{\aleph}$. For has not soft finite subcover of $\widetilde{\aleph}$. Thus, $(\aleph, \mathcal{Q}, \wp)$ is not a sSb^* -compact.

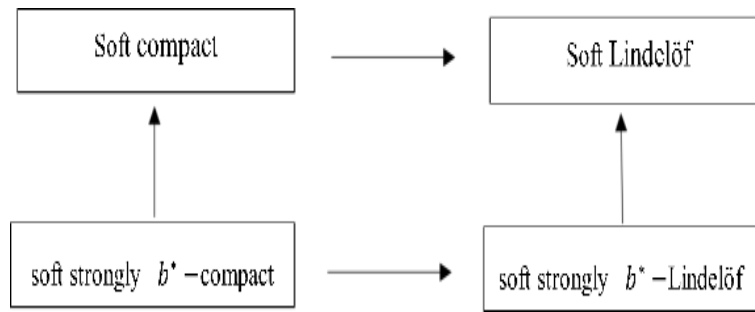


Fig. 2.1: Relationships

THEOREM 2.10. *The soft union of two sSb^* -compact (resp. sSb^* -Lindelöf) sets is sSb^* -compact (resp. sSb^* -Lindelöf).*

Proof. Let (ψ, \wp) and $(, \wp)$ be two sSb^* -compact sets. Assume that $\{(\psi_\epsilon, \wp) : \epsilon \in \zeta\}$ is a sSb^* -open cover of $(\psi, \wp) \cup (, \wp)$. Then, $\{(\psi_\epsilon, \wp) : \epsilon \in \zeta\}$ is a sSb^* -open cover of (ψ, \wp) and $(, \wp)$. Since (ψ, \wp) and $(, \wp)$ are sSb^* -compact, there exist finite subfamilies ζ_0 and ζ_1 of ζ such that $(\psi, \wp) \subseteq \{(\psi_\epsilon, \wp) : \epsilon \in \zeta_0\}$ and $(, \wp) \subseteq \{(\psi_\epsilon, \wp) : \epsilon \in \zeta_1\}$. Hence, $(\psi, \wp) \cup (, \wp) \subseteq (\cup\{(\psi_\epsilon, \wp) : \epsilon \in \zeta_0\}) \cup (\cup\{(\psi_\epsilon, \wp) : \epsilon \in \zeta_1\})$. It follows that, $(\psi, \wp) \cup (, \wp) \subseteq \cup\{(\psi_\epsilon, \wp) : \epsilon \in \zeta_0 \cup \zeta_1\}$. Thus, $(\psi, \wp) \cup (, \wp)$ is a sSb^* -compact.

The proof of the case of sSb^* -Lindelöfness is similar. \square

THEOREM 2.11. *Every sSb^* -closed subset (\mathcal{L}, \wp) of sSb^* -compact $(\mathcal{W}, \mathcal{Q}, \wp)$ is a sSb^* -compact.*

Proof. Assume that (\mathcal{L}, \wp) be a sSb^* -closed subset of sSb^* -compact space $(\mathcal{W}, \mathcal{Q}, \wp)$. Then (\mathcal{L}^c, \wp) is a sSb^* -open. Let $\{(\underline{\epsilon}, \wp) : \underline{\epsilon} \in \ell\}$ be a sSb^* -open cover of (\mathcal{L}, \wp) . Therefore, $\{(\underline{\epsilon}, \wp) : \underline{\epsilon} \in \ell\} \cup (\mathcal{L}^c, \wp)$ is sSb^* -open cover of \mathcal{W} . For \mathcal{W} is sSb^* -compact space, \exists finite subcover $\{(\underline{\epsilon}, \wp) : \underline{\epsilon} \in \zeta_0\} \cup (\mathcal{L}^c, \wp)$ for \mathcal{W} . Now, $\{(\underline{\epsilon}, \wp) : \underline{\epsilon} \in \zeta_0\} \cup (\mathcal{L}^c, \wp) - (\mathcal{L}^c, \wp)$ is a finite subcover of $\{(\underline{\epsilon}, \wp) : \underline{\epsilon} \in \zeta\}$ for (\mathcal{L}, \wp) . So, (\mathcal{L}, \wp) is sSb^* -compact. \square

THEOREM 2.12. *Every sSb^* -closed subset (\mathcal{S}, \wp) of sSb^* -Lindelöf space $(\mathcal{W}, \mathcal{Q}, \wp)$ is sSb^* -Lindelöf.*

Proof. Assume that (\mathcal{S}, \wp) be sSb^* -closed subset (\mathcal{S}, \wp) of sSb^* -compact space $(\mathcal{W}, \mathcal{Q}, \wp)$ and $\{(\Psi_\epsilon, \wp) : \epsilon \in \zeta\}$ be sSb^* -open cover of (\mathcal{S}, \wp) . Therefore, (\mathcal{S}^c, \wp) is a sSb^* -open and $(\mathcal{S}^c, \wp) \subseteq \cup_{\epsilon \in \zeta} (\Psi_\epsilon, \wp)$. Therefore, $\widetilde{\mathcal{W}} = (\Psi_\epsilon, \wp) \cup (\mathcal{S}^c, \wp)$. Since $\widetilde{\mathcal{W}}$ is a sSb^* -Lindelöf space, then $\widetilde{\mathcal{W}} = \cup_{\epsilon \in \zeta} (\Psi_\epsilon, \wp) \cup (\mathcal{S}^c, \wp)$. This implies that $(\mathcal{S}, \wp) \subseteq \cup_{\epsilon \in \zeta} (\Psi_\epsilon, \wp)$. Hence, (\mathcal{S}, \wp) is a sSb^* -Lindelöf. \square

COROLLARY 2.13. *If (δ, \wp) is a sSb^* -closed subset of $\widetilde{\mathcal{W}}$ and (Ψ, \wp) is a sSb^* -compact (resp. sSb^* -Lindelöf) subset of $\widetilde{\mathcal{W}}$. Then, $(\Psi, \wp) \cap (\delta, \wp)$ is a sSb^* -compact (resp. sSb^* -Lindelöf).*

Proof. Let (Ψ, \wp) be a sSb^* -compact set, consider $\{(G_\epsilon, \wp) : \epsilon \in \zeta\}$ is sSb^* -open cover of $(\Psi, \wp) \cap (\delta, \wp)$. Then $\{(G_\epsilon, \wp) : \epsilon \in \zeta\} \cup (\delta^c, \wp)$ is sSb^* -open cover of (Ψ, \wp) . For (Ψ, \wp) is a sSb^* -compact. So, \exists a soft finite subfamily ζ_0 of $\zeta \ni (\Psi, \wp) \subseteq \cup_{\epsilon \in \zeta_0} (G_\epsilon, \wp) \cup (\delta^c, \wp)$. Hence, $(\Psi, \wp) \cap (\delta, \wp) \subseteq \cup_{\epsilon \in \zeta_0} (G_\epsilon, \wp) \cap (\delta, \wp) \subseteq \cup_{\epsilon \in \zeta_0} (G_\epsilon, \wp)$. Therefore, $(\Psi, \wp) \cap (\delta, \wp)$ is sSb^* -compact.

The same evidence applies to sSb^* -Lindelöf space. \square

In the next example, indicates that the inclusions of the Theorem 2.12 is not necessary correct.

Example 3. Let $\mathcal{W} = \{h_1, h_2\}$ and $\wp = \{\nabla_1, \nabla_2\}$. Consider

$\mathcal{Q} = \{\widetilde{\mathcal{W}}, \phi, (\Psi_1, \wp), (\Psi_2, \wp), (\Psi_3, \wp)\}$ where $(\Psi_1, \wp), (\Psi_2, \wp)$ and (Ψ_3, \wp) defined as following manner:

$(\Psi_1, \wp) = \{(\nabla_1, \{h_1\}), (\nabla_2, \phi)\},$

$(\Psi_2, \wp) = \{(\nabla_1, \phi), (\nabla_2, \{h_2\})\}$ and

$(\Psi_3, \wp) = \{(\nabla_1, \{h_1\}), (\nabla_2, \{h_2\})\}.$

Then $(\mathcal{W}, \mathcal{Q}, \wp)$ is a *STS* over \mathcal{W} . Obviously, $(\mathcal{W}, \mathcal{Q}, \wp)$ is sSb^* -compact. furthermore, a soft set $(\Pi, \wp) = \{(\nabla_1, \{h_1\}), (\nabla_2, \mathcal{W})\}$ is a sSb^* -compact, even so it is not a sSb^* -closed.

THEOREM 2.14. *A $(\mathcal{W}, \mathcal{Q}, \wp)$ is sSb^* -compact (resp. sSb^* -Lindelöf) if and only if each collection of sSb^* -closed subsets of $(\mathcal{W}, \mathcal{Q}, \wp)$, satisfying the soft finite (resp. soft countable) intersection property, $\cap_{\epsilon \in \ell} (\Psi_\epsilon, \wp) \neq \phi$.*

Proof. Let $(\mathcal{W}, \mathcal{Q}, \wp)$ is a sSb^* -compact, and $\{(\xi_\epsilon, \wp) : \epsilon \in \ell\}$ be a family of sSb^* -closed subsets of $\widetilde{\mathcal{W}}$. Let $\cap_{\epsilon \in \ell} (\xi_\epsilon, \wp) = \phi$. Then $\widetilde{\mathcal{W}} = \cup_{\epsilon \in \ell} (\xi_\epsilon^c, \wp)$. For $\cup_{\epsilon \in \ell} (\xi_\epsilon^c, \wp)$ is a collection of sSb^* -open sets covering $\widetilde{\mathcal{W}}$. As $(\mathcal{W}, \mathcal{Q}, \wp)$ is sSb^* -compact, then \exists a soft finite subset ℓ_0 of $\ell \ni \cup_{\epsilon \in \ell_0} (\xi_\epsilon^c, \wp) = \widetilde{\mathcal{W}}$ then $\cap_{\epsilon \in \ell_0} (\xi_\epsilon, \wp) = \phi$. Which gives contradictions. Therefore, $\cap_{\epsilon \in \ell} (\xi_\epsilon, \wp) \neq \phi$. Conversely, let $\{(\gamma_\epsilon, \wp) : \epsilon \in \ell\}$ be a family of sSb^* -open cover of $\widetilde{\mathcal{W}}$. Let for every finite subset $\ell_0 \subset \ell$, we have $\cup_{\epsilon \in \ell_0} (\gamma_\epsilon^c, \wp) \neq \widetilde{\mathcal{W}}$. Then $\cap_{\epsilon \in \ell} (\gamma_\epsilon^c, \wp) \neq \phi$. Thus, $\{(\gamma_\epsilon^c, \wp) : \epsilon \in \ell\}$ satisfies the finite intersection property. By definition get $\cap_{\epsilon \in \ell} (\gamma_\epsilon^c, \wp) \neq \phi$ which implies $\cup_{\epsilon \in \ell_0} (\gamma_\epsilon, \wp) \neq \widetilde{\mathcal{W}}$ and this contradicts that $\{(\gamma_\epsilon, \wp) : \epsilon \in \ell\}$ is a sSb^* -open cover of $\widetilde{\mathcal{W}}$. Hence, $(\mathcal{W}, \mathcal{Q}, \wp)$ is a sSb^* -compact space. \square

THEOREM 2.15. *Let $\Pi : \mathcal{W} \rightarrow \Sigma$ be a sSb^* -continuous function. If \mathcal{W} is a sSb^* -compact space, then the image of \mathcal{W} under the Π is a soft compact.*

Proof. Assume $\Pi : \mathcal{W} \rightarrow \Sigma$ is a sSb^* -continuous, $\{(G_\epsilon, \wp) : \epsilon \in \ell\}$ is a soft cover of Σ . For Π is a sSb^* -continuous, therefore $\{\Pi^{-1}((G_\epsilon, \wp)) : \epsilon \in \ell\}$ is a sSb^* -open cover of $\widetilde{\mathcal{W}}$ and \mathcal{W} is a sSb^* -compact, \exists a soft finite sub-set ℓ_0 of $\ell \ni \{\Pi^{-1}((G_\epsilon, \wp)) : \epsilon \in \ell_0\}$ composes a sSb^* -open cover of $\widetilde{\mathcal{W}}$. Thus, $\{\Pi^{-1}((G_\epsilon, \wp)) : \epsilon \in \ell_0\}$ composes a soft finite soft open cover of $\widetilde{\Sigma}$. Therefore, Σ is a soft compact. \square

THEOREM 2.16. *Let $\Pi : \mathcal{W} \rightarrow \Sigma$ be a sSb^* -irresolute surjection and \mathcal{W} is a sSb^* -compact space, then Σ is a sSb^* -compact.*

Proof. Suppose $\Pi : \mathcal{W} \rightarrow \Sigma$ be a sSb^* -irresolute surjection, \mathcal{W} be a sSb^* -compact $(\mathcal{W}, \mathcal{Q}, \wp)$ to (Σ, Ω, Θ) . A soft open cover $\{(\delta_\epsilon, \wp) : \epsilon \in \zeta\}$ of Σ . Then $\{\Pi^{-1}((\delta_\epsilon, \wp)) : \epsilon \in \zeta\}$ is a sSb^* -open cover of $\widetilde{\mathcal{W}}$. For \mathcal{W} is a sSb^* -compact, then \exists a finite subset ζ_0 of ζ such that $\{\Pi^{-1}((\delta_\epsilon, \wp)) : \epsilon \in \zeta_0\}$ composes a sSb^* -open cover of $\widetilde{\mathcal{W}}$. Therefore, $\{\Pi^{-1}((\delta_\epsilon, \wp)) : \epsilon \in \zeta_0\}$ composes a finite sSb^* -open cover of $\widetilde{\Sigma}$. Hence, Σ is a sSb^* -compact. \square

THEOREM 2.17. *The sSb^* -irresolute image of a sSb^* -compact (resp. sSb^* -Lindelöf) set is a sSb^* -compact (resp. sSb^* -Lindelöf).*

Proof. Assume that $\Pi : \mathcal{W} \rightarrow \Sigma$ be a sSb^* -irresolute and let $(, \wp)$ be a sSb^* -Lindelöf subset of $\widetilde{\mathcal{W}}$. Let $\{(\Psi_\epsilon, \wp) : \epsilon \in \zeta\}$ is sSb^* -open cover of $\Pi(, \wp)$. Then $\Pi(, \wp) \subseteq \cup_{\epsilon \in \zeta} (\Psi_\epsilon, \wp)$. Then, $(, \wp) \subseteq \cup_{\epsilon \in \zeta} \Pi^{-1}(\Psi_\epsilon, \wp)$ and $\Pi^{-1}(\Psi_\epsilon, \wp)$ is sSb^* -open, for every $\epsilon \in \zeta$. by assumption, $(, \wp)$ is a sSb^* -Lindelöf, then $(, \wp) \subseteq \cup_{\epsilon \in \zeta} \Pi^{-1}(\Psi_\epsilon, \wp)$. Therefore, $\Pi(, \wp) \subseteq \cup_{\epsilon \in \zeta} \Pi(\Pi^{-1}(\Psi_\epsilon, \wp)) \subseteq \cup_{\epsilon \in \zeta} (\Psi_\epsilon, \wp)$. Thus, $\Pi(, \wp)$ is sSb^* -Lindelöf space. The same proof in case sSb^* -compact space. \square

3. Soft strongly b^* -connected spaces. One of the most important properties of soft strongly b^* -connected space is discussed and explored in this section.

DEFINITION 3.1. *Let $(\mathcal{W}, \mathcal{Q}, \wp)$ be STS, and $(\Psi, \wp), (\mathcal{L}, \wp)$ are sSb^* -open sets over $\widetilde{\mathcal{W}}$. Then, (Ψ, \wp) and (\mathcal{L}, \wp) are stated to be soft strongly b^* -separated sets iff $sSb^*cl(\Psi, \wp) \cap (\mathcal{L}, \wp) = \phi$ and $(\Psi, \wp) \cap sSb^*cl(\mathcal{L}, \wp) = \phi$.*

THEOREM 3.2. *If (Ψ, \wp) and (\mathcal{L}, \wp) are sSb^* -separated sets then they are disjoint.*

Proof. $(\Psi, \wp) \cap (\mathcal{L}, \wp) \subseteq sSb^*cl(\Psi, \wp) \cap (\mathcal{L}, \wp) = \phi$. \square

THEOREM 3.3. *If (Ψ, \wp) and (\mathcal{L}, \wp) are sSb^* -separated subsets of \mathcal{W} and $(\Gamma, \wp) \subseteq (\Psi, \wp)$ and $(\Upsilon, \wp) \subseteq (\mathcal{L}, \wp)$ then (Γ, \wp) and (Υ, \wp) are also sSb^* -separated.*

Proof. Suppose (Ψ, \wp) and (\mathcal{L}, \wp) are sSb^* -separated subsets of a space \mathcal{W} , by definition 3.1; $sSb^*cl(\Psi, \wp) \cap (\mathcal{L}, \wp) = \phi$ and $(\Psi, \wp) \cap sSb^*cl(\mathcal{L}, \wp) = \phi$. Since $(\Gamma, \wp) \subseteq (\Psi, \wp)$, we have $sSb^*cl(\Gamma, \wp) \subseteq sSb^*cl(\Psi, \wp)$ and since $(\Upsilon, \wp) \subseteq (\mathcal{L}, \wp)$, then $sSb^*cl(\Upsilon, \wp) \subseteq sSb^*cl(\mathcal{L}, \wp)$. Hence, $(\Gamma, \wp) \cap sSb^*cl(\Upsilon, \wp) = (\Psi, \wp) \cap sSb^*cl(\mathcal{L}, \wp) = \phi$ and $sSb^*cl(\Gamma, \wp) \cap (\Upsilon, \wp) = sSb^*cl(\Psi, \wp) \cap (\mathcal{L}, \wp) = \phi$. Therefore, (Γ, \wp) and (Υ, \wp) are also sSb^* -separated. \square

THEOREM 3.4. *Two soft separated sets are soft sSb^* -separated sets.*

Proof. Assume (ϑ, \wp) and $(, \wp)$ be two soft separated sets over \mathcal{W} , so $sSb^*cl(\vartheta, \wp) \cap (, \wp) = \phi$ and $(\vartheta, \wp) \cap sSb^*cl(, \wp) = \phi$.

As

$$sSb^*cl(\vartheta, \wp) \subseteq cl(\vartheta, \wp).$$

$$sSb^*cl(\vartheta, \wp) \cap (, \wp) \subseteq cl(\vartheta, \wp) \cap (, \wp) = \phi.$$

$$\text{and similarly, } (\vartheta, \wp) \cap sSb^*cl(, \wp) \subseteq (\vartheta, \wp) \cap cl(, \wp) = \phi.$$

Hence, (ϑ, \wp) and $(, \wp)$ are sSb^* -separated sets.

\square

Remark 1. If (ϑ, \wp) and $(, \wp)$ are disjoint. Then, require not be sSb^* -separated.

Example 4. Consider $\mathcal{W} = \{\varsigma_1, \varsigma_2\}$ and $\wp = \{\nabla_1, \nabla_2\}$. Consider $\mathcal{Q} = \{\widetilde{\mathcal{W}}, \widetilde{\phi}, (\vartheta, \wp)\}$ where $(\vartheta, \wp) = \{(\nabla_1, \{\varsigma_1\}), (\nabla_2, \{\varsigma_2\})\}$, let $(, \wp) = \{(\nabla_1, \{\varsigma_1\}), (\nabla_2, \{\varsigma_2\})\}$ and $(\mathcal{S}, \wp) = \{(\nabla_1, \{\varsigma_2\}), (\nabla_2, \{\varsigma_1\})\}$ be two soft sets over \mathcal{Q} . Then $(, \wp)$ and (\mathcal{S}, \wp) are soft disjoint sets but they are not sSb^* -separated.

DEFINITION 3.5. Let $(\mathcal{W}, \mathcal{Q}, \wp)$ be STS over \mathcal{W} . Then $(\mathcal{W}, \mathcal{Q}, \wp)$ is stated to be sSb^* -connected, if \mathcal{W} cannot be intimated as the union of two sSb^* -open sets. Else, $(\mathcal{W}, \mathcal{Q}, \wp)$ is stated to be a sSb^* -disconnected.

Example 5. Consider $\mathcal{W} = \{r, t, d\}$ and $\wp = \{\nabla_1, \nabla_2\}$. Consider

$$\mathcal{Q} = \{\widetilde{\mathcal{W}}, \widetilde{\phi}, (\Gamma_1, \wp), (\Gamma_2, \wp), (\Gamma_3, \wp), (\Gamma_4, \wp), (\Gamma_5, \wp)\}$$

where $(\Gamma_1, \wp), (\Gamma_2, \wp), (\Gamma_3, \wp), (\Gamma_4, \wp)$ and (Γ_5, \wp) are sSb^* -open sets over \mathcal{W} , define as follows:

$$(\Gamma_1, \wp) = \{(\nabla_1, \{t\}), (\nabla_2, \{r\})\},$$

$$(\Gamma_2, \wp) = \{(\nabla_1, \{t, d\}), (\nabla_2, \{r, t\})\},$$

$$(\Gamma_3, \wp) = \{(\nabla_1, \{r, t\}), (\nabla_2, \mathcal{W})\},$$

$$(\Gamma_4, \wp) = \{(\nabla_1, \{r, t\}), (\nabla_2, \{r, d\})\} \text{ and}$$

$$(\Gamma_5, \wp) = \{(\nabla_1, \{t\}), (\nabla_2, \{r, t\})\}.$$

Then $(\mathcal{W}, \mathcal{Q}, \wp)$ is STS on \mathcal{W} . Thus, $(\mathcal{W}, \mathcal{Q}, \wp)$ is a STS over \mathcal{W} . Intelligibly, \mathcal{W} is a sSb^* -connected.

THEOREM 3.6. Let $(\mathcal{W}, \mathcal{Q}, \wp)$ be a STS and (Ψ, \wp) is a sSb^* -connected. Let (\mathcal{L}, \wp) and (\mathcal{S}, \wp) are sSb^* -separated sets. If $(\Psi, \wp) \subseteq (\mathcal{L}, \wp) \cup (\mathcal{S}, \wp)$. Then either $(\Psi, \wp) \subseteq (\mathcal{L}, \wp)$ or $(\Psi, \wp) \subseteq (\mathcal{S}, \wp)$.

Proof. Suppose (Ψ, \wp) be a sSb^* -connected and $(\mathcal{L}, \wp), (\mathcal{S}, \wp)$ are sSb^* -separated sets such that $(\Psi, \wp) \subseteq (\mathcal{L}, \wp) \cup (\mathcal{S}, \wp)$. Let (Ψ, \wp) not subset of (\mathcal{L}, \wp) and (Ψ, \wp) not subset of (\mathcal{S}, \wp) . Suppose $(P_1, \wp) \subseteq (\mathcal{L}, \wp) \cap (\Psi, \wp) \neq \phi$ and $(P_2, \wp) \subseteq (\mathcal{S}, \wp) \cap (\Psi, \wp) \neq \phi$. Then $(\Psi, \wp) = (P_1, \wp) \cup (P_2, \wp)$. Since $(P_1, \wp) \subseteq (\mathcal{L}, \wp)$, hence $sSb^*cl(P_1, \wp) \subseteq sSb^*cl(\mathcal{L}, \wp)$. Since $sSb^*cl(\mathcal{L}, \wp) \cap (\mathcal{S}, \wp) = \phi$ then $sSb^*cl(P_1, \wp) \cap (P_2, \wp) = \phi$. Since $(P_2, \wp) \subseteq (\mathcal{S}, \wp)$, hence $sSb^*cl(P_2, \wp) \subseteq sSb^*cl(\mathcal{S}, \wp)$. Since $sSb^*cl(\mathcal{S}, \wp) \cap (\mathcal{L}, \wp) = \phi$, then $sSb^*cl(P_2, \wp) \cap (P_1, \wp) = \phi$. But $(\Psi, \wp) = (P_1, \wp) \cup (P_2, \wp)$. Therefore, (Ψ, \wp) is not a sSb^* -connected. This is a contradiction. Then either $(\Psi, \wp) \subseteq (\mathcal{L}, \wp)$ or $(\Psi, \wp) \subseteq (\mathcal{S}, \wp)$. \square

THEOREM 3.7. Let (Ψ, \wp) be a sSb^* -connected set. If $(\Psi, \wp) \subseteq (\mathcal{L}, \wp) \subseteq sSb^*cl(\Psi, \wp)$ then (\mathcal{L}, \wp) is also a sSb^* -connected.

Proof. If (Ψ, \wp) be not a sSb^* -connected, then \exists two soft sets $(\mathcal{S}, \wp) \subseteq (G, \wp)$ such that $sSb^*cl(\mathcal{S}, \wp) \cap (G, \wp) = (G, \wp) \cap sSb^*cl(\mathcal{S}, \wp) = \phi$ and $(\mathcal{L}, \wp) = (\mathcal{S}, \wp) \cup (G, \wp)$. Since $(\Psi, \wp) \subseteq (\mathcal{L}, \wp)$, thus either $(\Psi, \wp) \subseteq (\mathcal{S}, \wp)$ or $(\Psi, \wp) \subseteq (G, \wp)$. Suppose $(\Psi, \wp) \subseteq (\mathcal{S}, \wp)$ then $sSb^*cl(\Psi, \wp) \subseteq sSb^*cl(\mathcal{S}, \wp)$, thus $sSb^*cl(\Psi, \wp) \subseteq (G, \wp) = sSb^*cl(\mathcal{S}, \wp) \cap (G, \wp) = \phi$. But $(G, \wp) \subseteq (\mathcal{L}, \wp) \subseteq sSb^*cl(\Psi, \wp)$ thus $sSb^*cl(\Psi, \wp) \cap (G, \wp) = (G, \wp)$. Therefore, $(G, \wp) = \phi$, so is a contradiction. Hence, (\mathcal{L}, \wp) is a sSb^* -connected. Similarly, if $(\Psi, \wp) \subseteq (\mathcal{L}, \wp)$, then $(\mathcal{S}, \wp) = \phi$. Which again a contradiction. Hence, (\mathcal{L}, \wp) is a sSb^* -connected. \square

THEOREM 3.8. If (Ψ, \wp) is a sSb^* -connected set then $sSb^*cl(\Psi, \wp)$ is a sSb^* -connected.

Proof. Let (Ψ, \wp) is a sSb^* -connected set then $sSb^*cl(\Psi, \wp)$ is not. Then there exists two sSb^* -separation sets (\mathcal{S}, \wp) and (δ, \wp) such that $sSb^*cl(\Psi, \wp) = (\mathcal{S}, \wp) \cup (\delta, \wp)$. But $(\Psi, \wp) \subseteq sSb^*cl(\Psi, \wp)$, then $(\Psi, \wp) = (\mathcal{S}, \wp) \cup (\delta, \wp)$ and since (Ψ, \wp) is sSb^* -connected set, then by theorem 3.6 either $(\Psi, \wp) \subseteq (\mathcal{S}, \wp)$ or $(\Psi, \wp) \subseteq (G, \wp)$. If $(\Psi, \wp) \subseteq (\mathcal{S}, \wp)$ then $sSb^*cl(\Psi, \wp) \subseteq sSb^*cl(\mathcal{S}, \wp)$. But $sSb^*cl(\mathcal{S}, \wp) \cap (\delta, \wp) = \phi$. Hence, $sSb^*cl(\Psi, \wp) \cap (\delta, \wp) = \phi$. Since $(\delta, \wp) \subseteq sSb^*cl(\Psi, \wp)$, then $(\delta, \wp) = \phi$. So is a contradiction. If $(\Psi, \wp) \subseteq (\mathcal{S}, \wp)$ we can prove $(\mathcal{S}, \wp) = \phi$ as the same, that is a contradiction. Hence, $sSb^*cl(\Psi, \wp)$ is a sSb^* -connected. \square

THEOREM 3.9. If (Ψ, \wp) and (\mathcal{S}, \wp) are two sSb^* -connected sets where $(\Psi, \wp) \cap (\mathcal{S}, \wp) \neq \phi$. Therefore $(\Psi, \wp) \cup (\mathcal{S}, \wp)$ is also a sSb^* -connected set.

Proof. Assume that, if possible, $(\Psi, \wp) \cup (\mathcal{S}, \wp)$ be sSb^* -disconnected set, then $(\Psi, \wp) \cup (\mathcal{S}, \wp) = (\vartheta, \wp) \cup (\delta, \wp)$, where $(\vartheta, \wp) \neq \phi, (\delta, \wp) \neq \phi \ni (\vartheta, \wp)$ and (δ, \wp) are sSb^* -separation. Since $(\Psi, \wp) \subseteq (\Psi, \wp) \cup (\mathcal{S}, \wp) = (\vartheta, \wp) \cup (\delta, \wp)$, Therefore, $(\Psi, \wp) \subseteq (\vartheta, \wp) \cup (\delta, \wp)$. Hence, by Theorem 3.6, we have either $(\Psi, \wp) \subseteq (\vartheta, \wp)$ or $(\Psi, \wp) \subseteq (\delta, \wp)$. Again, either $(\mathcal{S}, \wp) \subseteq (\vartheta, \wp)$ or $(\mathcal{S}, \wp) \subseteq (\delta, \wp)$. Thus, we have four choices either $(\Psi, \wp) \subseteq (\vartheta, \wp)$ and $(\mathcal{S}, \wp) \subseteq (\vartheta, \wp)$ or $(\Psi, \wp) \subseteq (\vartheta, \wp)$ and $(\mathcal{S}, \wp) \subseteq (\delta, \wp)$ or $(\Psi, \wp) \subseteq (\delta, \wp)$ and $(\mathcal{S}, \wp) \subseteq (\vartheta, \wp)$ or $(\Psi, \wp) \subseteq (\delta, \wp)$ and $(\mathcal{S}, \wp) \subseteq (\delta, \wp)$. If $(\Psi, \wp) \subseteq (\vartheta, \wp)$ and $(\mathcal{S}, \wp) \subseteq (\vartheta, \wp)$ or $(\Psi, \wp) \subseteq (\delta, \wp)$ and $(\mathcal{S}, \wp) \subseteq (\delta, \wp)$, then $(\Psi, \wp) \cup (\mathcal{S}, \wp) \subseteq (\vartheta, \wp)$ or $(\Psi, \wp) \cup (\mathcal{S}, \wp) \subseteq (\delta, \wp) \Rightarrow (\vartheta, \wp) \cup (\delta, \wp) \subseteq (\vartheta, \wp)$ or $(\vartheta, \wp) \cup (\delta, \wp) \subseteq (\delta, \wp) \Rightarrow (\vartheta, \wp) \cup (\delta, \wp) = (\vartheta, \wp)$ or $(\vartheta, \wp) \cup (\delta, \wp) = (\delta, \wp) \Rightarrow (\delta, \wp) = \phi$ or $(\vartheta, \wp) = \phi$, then is a contradiction. So $(\Psi, \wp) \subseteq (\vartheta, \wp)$ and $(\mathcal{S}, \wp) \subseteq (\delta, \wp)$ or $(\Psi, \wp) \subseteq (\delta, \wp)$ and $(\mathcal{S}, \wp) \subseteq (\vartheta, \wp)$, then in both the cases, $(\Psi, \wp) \cap (\mathcal{S}, \wp) \subseteq (\vartheta, \wp) \cap (\delta, \wp) = \phi \Rightarrow (\Psi, \wp) \cap (\mathcal{S}, \wp) = \phi$. So, is contradiction again to the given supposition that $(\Psi, \wp) \cap (\mathcal{S}, \wp) \neq \phi$. Hence,

we have $(\Psi, \wp) \cup (\mathcal{S}, \wp)$ is also a sSb^* -connected set. \square

THEOREM 3.10. *If $(\mathcal{W}, \mathcal{Q}, \wp)$ is a sSb^* -connected space, then it is a soft connected.*

Proof. Suppose $(\mathcal{W}, \mathcal{Q}, \wp)$ not sSb^* -disconnected. Therefore, \exists nonnull soft sets (\mathcal{L}, \wp) and (\mathcal{S}, \wp) , where $\mathcal{W} = (\mathcal{L}, \wp) \cup (\mathcal{S}, \wp) \ni cl(\mathcal{L}, \wp) \cap (\mathcal{S}, \wp) = \phi$ and $(\mathcal{L}, \wp) \cap cl(\mathcal{S}, \wp) = \phi$. Since $sSb^*cl(\mathcal{L}, \wp) \subseteq cl(\mathcal{L}, \wp)$. Thus, $sSb^*cl(\mathcal{L}, \wp) \cap (\mathcal{S}, \wp) \subseteq cl(\mathcal{L}, \wp) \cap (\mathcal{S}, \wp) = \phi$. Hence, $sSb^*cl(\mathcal{L}, \wp) \cap (\mathcal{S}, \wp) = \phi$. Similarly, $(\mathcal{L}, \wp) \cap sSb^*cl(\mathcal{S}, \wp) = \phi$. Hence, $(\mathcal{W}, \mathcal{Q}, \wp)$ is a sSb^* -disconnected. So, is a contradiction. Therefore, $(\mathcal{W}, \mathcal{Q}, \wp)$ is a soft connected. \square

THEOREM 3.11. *If $\Omega : \mathcal{W} \rightarrow \Sigma$ be a sSb^* -continuous surjection and \mathcal{W} is a sSb^* -connected space, then Σ is soft connected.*

Proof. Assume that Σ is not soft connected. Let $\Sigma = (\Psi, \wp) \cup (\mathcal{L}, \wp)$ where (Ψ, \wp) and (\mathcal{L}, \wp) are disjoint nonempty soft open sets in Σ . Since Ω is a sSb^* -continuous and onto, $\mathcal{W} = \Omega^{-1}(\Psi, \wp) \cup \Omega^{-1}(\mathcal{L}, \wp)$ where $\Omega^{-1}(\Psi, \wp)$ and $\Omega^{-1}(\mathcal{L}, \wp)$ are disjoint nonempty sSb^* -open sets in \mathcal{W} , which is contradiction to \mathcal{W} is a sSb^* -connected. Therefore, Σ is a soft connected. \square

THEOREM 3.12. *If $\Omega : \mathcal{W} \rightarrow \Sigma$ is a sSb^* -irresolute surjection and \mathcal{W} is a sSb^* -connected, then Σ is a sSb^* -connected.*

Proof. Assume Σ is not sSb^* -connected and $\Sigma = (\Psi, \wp) \cup (\mathcal{L}, \wp)$ where (Ψ, \wp) and (\mathcal{L}, \wp) are disjoint nonempty sSb^* -open sets in Σ . Since Ω is a sSb^* -irresolute and onto, $\mathcal{W} = \Omega^{-1}(\Psi, \wp) \cup \Omega^{-1}(\mathcal{L}, \wp)$ where $\Omega^{-1}(\Psi, \wp)$ and $\Omega^{-1}(\mathcal{L}, \wp)$ are disjoint nonempty sSb^* -open sets in \mathcal{W} , which is contradiction to \mathcal{W} is a sSb^* -connected. Therefore, Σ is a sSb^* -connected. \square

4. Conclusion. In this article, we presented some of conception of soft sets and soft topological spaces are investigated. The basis of paper is to establish and introduce soft compactness and soft Lindelöfness, namely, sSb^* -compactness, sSb^* -Lindelöfness. Examining some properties of these spaces allows us to prove some of our results and varied introduce the relationship between spaces and illustrate our main findings. Moreover, We define and explore the soft strongly b^* -connected spaces and discuss its relation with soft connectedness spaces. Also, the properties of sSb^* -connected and sSb^* -disconnected with examples are studied.

5. Acknowledgements. The authors would like to thank Mustansiriya University, (<https://uomustansiriya.edu.iq/>), Baghdad, Iraq, for its support to the present work. Also, the authors would like to thank the Reviewers for their valuable comments which help us to improve the manuscript.

REFERENCES

- [1] D. MOLODTSOV, *Soft set theory-first results*, Computers and Mathematics with Applications, vol. 37, no. 4-5, pp. 19-31, 1999.
- [2] SHABIR, M., NAZ, M., *On soft topological spaces*, Comput. Math. Appl. 61, pp. 17861799, 2011.
- [3] K. KANNAN, *Soft generalized closed sets in soft topological spaces*, journal of Theoretical and Appl. Inform. Technology, 37 (1) pp.17-21, 2012.
- [4] METIN AKDAG, ALKAN OZKAN, *Soft α -open sets and soft α -continuous functions*, Abstr. Anal. Appl. Art ID 891341, pp. 1-7, 2014.
- [5] METIN AKDAG, ALKAN OZKAN, *Soft b -open and soft b -continuous functions*, Math Sci 8:124, 2014.
- [6] I. ZORLUTUNA, M. AKDAG, W. K. MIN, AND S. ATMACA, *Remarks on soft topological spaces*, Annals of Fuzzy Mathematics and Informatics, vol. 3, no. 2, pp. 171-185, 2012.
- [7] J. R. PORTER AND R. G. WOODS, *Subspaces of connected spaces*, Topology and Its Applications, vol. 68, pp. 113-131, 1996. DOI: 10.1016/0166-8641(95)00057-7.
- [8] S. HUSSAIN, *A note on soft connectedness*, journal of the Egyptian Mathematical Society 23, pp. 6-11, 2015. DOI: 10.1016/j.joems.2014.02.003.
- [9] HAMEED, S. Z. AND HUSSEIN, A. K., *On soft bc -open sets in soft topological spaces*, Iraqi Journal of Science, pp. 238-242, 2020.
- [10] HAMEED, SAIF Z., FAYZA A. IBRAHEM, AND ESSAM A. EL-SEIDY, *On soft b^* -closed sets in soft topological space* International journal of Nonlinear Analysis and Applications, 12.1 pp. 1235-1242, 2021. <https://doi.org/10.24996/ijns.2020.SI.1.32>.
- [11] HAMEED, SAIF Z., ABDELAZIZ E. RADWAN, AND ESSAM A. EL-SEIDY, *On soft b^* -Continuous functions in soft topological space* Measurement: Sensors, vol. 27, pp. 1-5, 2023.
- [12] HAMEED, SAIF Z., ABDELAZIZ E. RADWAN, AND ESSAM A. EL-SEIDY, *On soft strongly b^* -closed sets and soft strongly b^* -continuous functions in soft topological space* Under the Publication, 2023.
- [13] MAJI, P.K., BISWAS, R., ROY, A.R., *Soft set theory*, Comput. Math. Appl. 45, pp. 555562, 2003.
- [14] I. AROCKIARANI AND A. AROKIALANCY, *Generalized soft $g\beta$ -closed sets and soft $gs\beta$ -closed sets in soft topological spaces* International journal of Mathematical Archive, vol. 4, no. 2, pp. 1-7, 2013.
- [15] B. CHEN, *Soft semi-open sets and related properties in soft topological spaces*, Applied Mathematics and Information Sciences, vol. 7, no. 1, pp. 287-294, 2013.

- [16] NIRMALA REBECCA PAUL, *Remarks on soft ω -closed sets in soft topological spaces*, Boletim da Sociedade Paranaense de Matemática, Vol. 33, No. 1, pp. 183-192, 2015. <https://doi.org/10.5269/bspm.v33i1.22719>.

Edited by: Mustafa M Matalgah

Special issue on: Synergies of Neural Networks, Neurorobotics, and Brain-Computer Interface Technology:
Advancements and Applications

Received: Jan 13, 2024

Accepted: Mar 12, 2024



SPORTS EVENT DATA MANAGEMENT SYSTEM AND ITS APPLICATION IN COMPETITION ORGANIZATION

ZHENYU LI*

Abstract. Traditional sports competition data management systems have poor security and reliability in management results. The author proposes a new sports competition data management system based on a network platform. Integrating internal and external services to handle system business, designing pages using JSP dynamic page technology. Utilize network platforms to transmit data, and verify the transmission of data on the network platform through hardware and software firewalls. Realize data transmission between the system and the client through the designed network platform. Analyzed data backup and recovery methods to ensure data security. Track important business and operations of the system through the log tracking layer, and provide UML modeling for competition data management. The experimental results show that the recovery speed of the system is significantly faster than SSH and MDA systems, indicating strong data recovery performance of the system. It has been proven that the competition data management results of the designed system are reliable and the system performance is strong.

Key words: Online platform, Sports competitions, Data management system, JSP dynamic page technology

1. Introduction. The organization of large-scale sports events is a huge systematic project, especially the management of competition data, which is the core link of event organization work. However, with the continuous increase in the scale of sports events, not only have the operating costs of sports events become increasingly high, but the data management process of sports competitions has also become more complex, leading to increasing risks in organizing sports events. With the vigorous development of computer technology, powerful solutions have been provided for the complex management process of sports competition results. Currently, the management of sports competition results cannot be separated from the support of computer software systems [1].

The development of competition data management software systems is generally based on the relevant competition management regulations and rules formulated by sports event organization and management institutions, and other management technical specifications. However, different sports events generally have different competition management regulations and rules, so different sports event organization and management institutions also use their own competition management regulations and rule management software systems, which leads to the current sports competition data management system's weak universality and poor adaptability [2].

The sports event data management system is an information system that integrates data collection, storage, analysis, and display functions, widely used in the organization and management process of various sports events. With the continuous improvement of electronic technology and information technology, the application of sports event data management systems has become an indispensable part of modern sports organizations [3]. Through comprehensive, accurate, and real-time management of competition data, this system can provide better competition experiences and services to event organizers, participants, and spectators. The core function of a sports event data management system is data collection. During the competition, the system can collect real-time game data, including game results, scores, goals scored, fouls committed, etc., through various sensors and devices such as timers, scoreboards, cameras, etc [4]. These data are transmitted to the central server through wireless transmission or wired connections, achieving real-time updates and storage of data. Through data collection and storage, the sports event data management system can perform data analysis. The system can perform various statistics and calculations based on competition data, generating competition reports, data charts, and event analysis reports. These reports can help event organizers understand the overall situation of

*College of Physical Education, Baicheng Normal University, Baicheng, Jilin, 137000, China (Corresponding author, lizhenyu8188@163.com)

the competition, identify the strengths and weaknesses of athletes, and develop more scientific and reasonable training and competition strategies.

The sports event data management system also has the function of data display. Through the internet and mobile terminals, the system can display real-time competition data to the audience and participating players. Viewers can watch live matches through devices such as television, computers, and mobile phones to understand the progress and results of the matches. Contestants can view their competition results and rankings, as well as the performance of other contestants, through their mobile phones or computers. This real-time display method makes the competition more open and transparent, improving the fairness and credibility of the competition [5]. The application of sports event data management system can greatly improve the organization and management efficiency of sports events.

Traditional manual recording and statistical methods suffer from human errors and inaccurate data, while sports event data management systems can automate data collection, storage, and analysis to ensure accuracy and timeliness. This not only saves labor costs, but also improves the efficiency and accuracy of data management.

In addition, the sports event data management system can also provide more commercial value for event organizers. By analyzing and mining competition data, the potential and market value of athletes can be discovered, providing more business cooperation opportunities and sponsorship resources for event organizers. At the same time, through the display and promotion of event data, it can attract more audiences and fans, improve the visibility and influence of the event [6]. However, the application of sports event data management systems also faces some challenges and problems. Firstly, data security and privacy protection are important considerations. The competition data includes personal information and performance data of athletes, and strict security measures need to be taken to prevent data leakage and abuse. Secondly, the stability and reliability of the system are also key issues. Once the system malfunctions or data is lost, it will have a serious impact on the organization of the event and the participating players. Finally, the popularization and promotion of sports event data management systems also need to overcome technical barriers and cost issues, so that more sports organizations and events can enjoy the convenience and advantages of information management. In summary, the application of sports event data management systems has become an important component of modern sports organizations [7]. By comprehensively, accurately, and in real-time managing competition data, the system provides better event experiences and services for event organizers, participants, and spectators. With the continuous development and innovation of technology, it is believed that the sports event data management system will play a more important role in the future, promoting the development and progress of the sports industry. For this purpose, a new sports competition data management system based on a network platform has been designed.

2. Sports Competition Data Management System on Network Platform.

2.1. Overall System Design. The designed sports competition data management system based on network platform is described in Figure 2.1.

Design a system that combines internal and external services to handle system business, and use JSP dynamic page technology to design pages, thereby reducing the complexity of page code writing.

In order to ensure the timeliness and security of the system, this section uses the network platform to transmit data, verifies the transmission requests of the network platform through hardware and software firewalls, intercepts abnormal requests, and only allows legitimate requests to access the system. Simultaneously utilizing different technological constraints to ensure data integrity [8].

2.2. Network Platform Design. The network platform is mainly used to achieve data transmission between the system and the client, consisting of server clusters, gateways, firewalls, and other network infrastructure devices. Due to the large amount of sports competition result data, in order to improve transmission efficiency, the system chooses a distributed processing system, adopts ASP technology and B/S architecture, and operates in a local area network and external network environment. The B-end is mainly used for inputting, managing, and outputting data, while the S-end is mainly used for saving, accessing, and processing data. The network platform structure is shown in Figure 2.2.

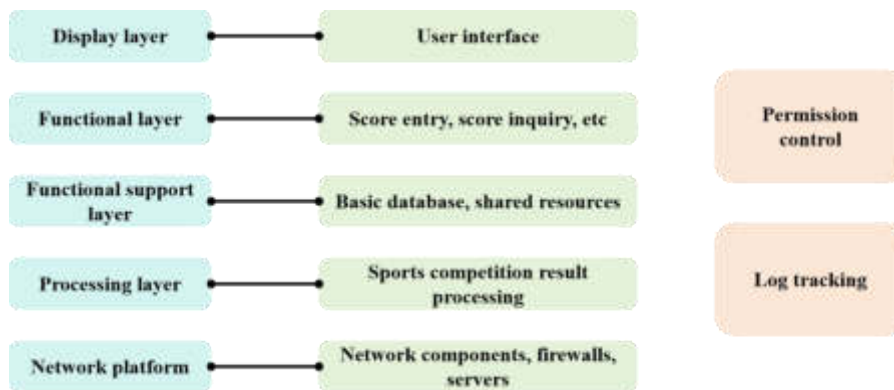


Fig. 2.1: Overall structure of the system

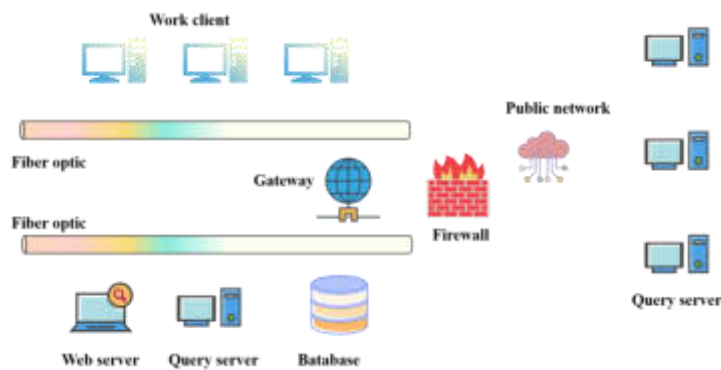


Fig. 2.2: Structure of the network platform

2.3. Functional Support Layer and Functional Layer Design. The functional support layer refers to the data center of the sports competition data management system and other network support, including components such as data storage and caching. Before athletes register online, the system administrator will publish the sports competition items and corresponding numbers, athlete grouping settings, and relevant information, making it easier for athletes to register independently [9].

During the competition, the system will promptly release competition information, and athletes can check their competition results through their own information. During the competition, the administrator inputs the competition results in real-time based on the competition status, and implements operations such as athlete grouping and competition result summary through the system. Users mainly include all sports competition staff, management personnel, athletes, etc., and can perform real-time queries on competition results[10]. The data flow diagram of the designed sports competition data management system is described in Figure 2.3.

During sports competitions, there are a large number of athletes, managers, and staff entering information, resulting in a large scale of data and heavy workload. Manual input can lead to data errors. In order to prevent these phenomena from occurring, data backup and recovery functions are provided in the design of the system to ensure the reliability of results. A detailed analysis will be conducted in the following text [11].

Before sports competitions, the administrator sets the competition items according to the requirements of the competition and processes the data based on the athlete's registration situation. The competition committee can check the athlete's competition results through online platforms, confirm them to be correct, and then output the results.

The functional layer contains several different types of functional modules, mainly responsible for imple-

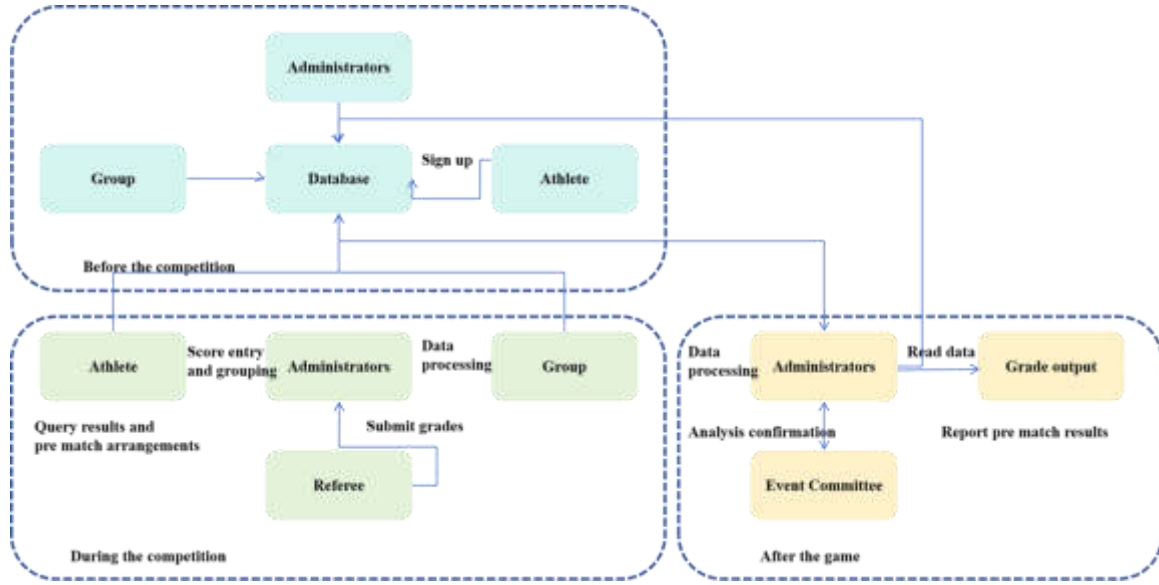


Fig. 2.3: Flow diagram of the system data

menting various business processes. By utilizing the interaction between the functional support layer and the data layer, corresponding business processing can be achieved, which not only displays information to users but also collects information from them[12].

2.4. Design of permission control layer. When designing the system, a firewall will be installed, with corresponding log records for all businesses in the system, mainly including operators and detailed operation information, for easy query.

In order to ensure the security of sports competition results, the system provides data backup and recovery functions. The system can autonomously backup data and recover it in the event of a failure, avoiding data loss [13].

When backing up sports competition data, encapsulate the request as a backup record and transfer it to the cache. The backup record mainly includes information such as the processing object, write length, and detailed data. The format is as follows:

$$R_{backup} = \{ \langle tid, offset, len, data \rangle \mid tid \in N, offset \in N, len \in N, data \in (0, 1)^t, t \in N \} \tag{2.1}$$

In the formula: tid is used to describe the excessive task identifier; $Offset$ is used to describe the offset address of the data to be processed; Len is used to describe the length of written data; $Data$ is used to describe writing data to a disk; N is used to describe a set of natural numbers. The process of recovering sports competition data is as follows:

1. Segmentation of sports competition data blocks and allocation of recovery tasks.
2. Calculate the summary values of different data and transfer the task record R_{task} to different remote backup servers for recovery. The task format is as follows:

$$R_{task} = \{ \langle tid, offset, len, LTH_i \rangle \mid tid \in N, offset \in N, len \in N, LTH_i \in N \} \tag{2.2}$$

In the formula: $LTH_i = H(L_i)$; H is used to describe the summary value function obtained; L_i is used to describe sports competition data to be restored.

3. After obtaining task records, read the object data block R_i from the remote backup server using a predetermined offset and data block size.



Fig. 2.4: Performance management: UML modeling

4. Find the summary values of different data blocks and compare them with the summary values transmitted by the local server. If they are inconsistent, it is considered that the data at both ends of the data block is different and needs to be restored; Otherwise, it is considered unnecessary to restore it[14].
5. Encapsulate the corresponding data blocks of different remote backup servers into recovery records and transmit them to the local server. The recovery record format is as follows:

$$R_{recovery} = \{ \langle tid, offset, len, data \rangle \mid tid \in N, offset \in N, len \in N, data \in (0, 1)^t, t \in N \} \quad (2.3)$$

6. After receiving data records, the local server transfers the data to the corresponding location on the disk through offset to achieve data recovery.

2.5. Log tracking design. The log tracking layer is mainly responsible for tracking important business and operations of the system, and recording relevant information in the form of logs, mainly including operation time, operator information, and specific operation information, in order to avoid losses caused by misoperations [15].

2.6. Processing layer design. The processing layer is mainly responsible for sports competition data management and is the core of the entire system. After the completion of different stages of the competition, the processing layer is responsible for inputting, processing, modifying, and printing the athlete's competition results, and grouping subsequent competitions based on the competition results. Detailed UML modeling is described using Figure 2.4.

3. Experiments and Result Analysis. The experiment tested the system from two aspects: The effectiveness of competition data management and system performance. In order to verify the effectiveness of the system, the SSH architecture system and MDA system were compared to manage the results of the men's 200m competition. The comparison results of the data management of the three systems are shown in Table 3.1.

Table 3.1: Management results of the three systems

Ranking	Actual grades	This article systematically announces the results	SSH system releases results	MDA system releases results
First place	2054	2054	2052	2054
Second place	2123	2123	2123	2123
Third place	2159	2159	2159	2146
Fourth place	2216	2216	2235	2223
Fifth place	2259	2259	2255	2259
Sixth place	2312	2312	2312	2312

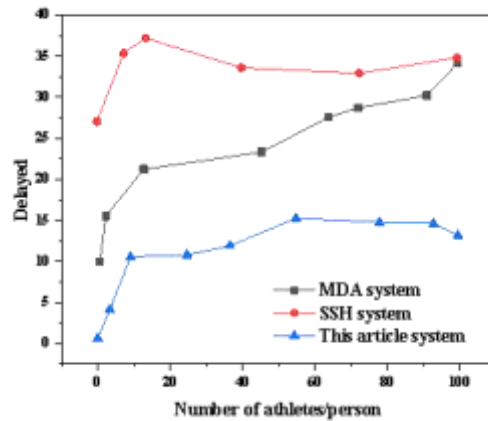


Fig. 3.1: Time delay comparison results of the three systems

According to Table 3.1, the results published by the system are completely consistent with the actual results of the athletes, while the SSH system has three results that do not match the actual results, and the MDA system has two results that do not match the actual results. This indicates that the data management results of the system are reliable and effective.

During the testing process, the memory usage of the system remained within the range of 1-2 GB, and the memory growth remained stable without any memory leakage, effectively ensuring that the server provided real-time services to users [16]. In order to verify the timeliness of the system, a comparison was made on the query latency of competition results for 100 athletes in different events under the system, SSH system, and MDA system. The results are shown in Figure 3.1.

Analyzing Figure 3.1, it can be seen that the query latency of the system is significantly lower than that of SSH and MDA systems, indicating that the system not only has high management reliability but also good query timeliness [17-20].

The data recovery performance is a security factor that affects the security of system data. The data recovery speeds of this system, SSH system, and MDA system are compared below, and the results are described in Figure 3.2. Analyzing Figure 3.2, it can be seen that the recovery speed of the system is significantly faster than that of SSH and MDA systems, indicating that the system has strong data recovery performance and high security.

4. Conclusion. The author has designed a new sports competition data management system based on a network platform, provided the overall structure of the designed system, and introduced the design process at each level. The sports event data management system is a software system that integrates multiple functions,

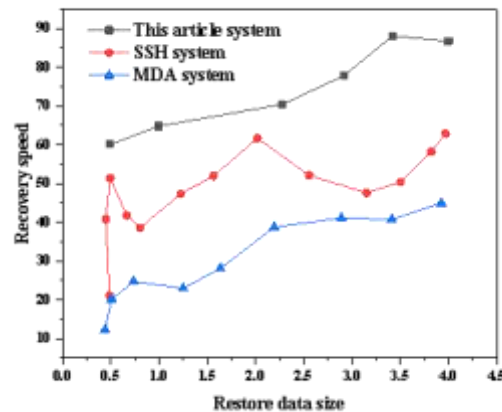


Fig. 3.2: Comparison of the data recovery speed of the three systems

which can help match organizers efficiently collect, store, and analyze match data. The basic principle of this system is to collect various data during the competition, such as competition time, scores, statistical data, etc., through various sensors and devices, and then store these data in a database for subsequent analysis. The experimental results show that the designed system has reliable competition data management results and strong system performance. This system has high practical value and can be applied to various sports events to improve the management efficiency and accuracy of competition results. With the continuous development of technology, there is still room for further development and improvement of the system in the future.

REFERENCES

- [1] Yao, L. , & Zou, J. . (2021). Research on sports competition information management based on computer database technology. *Journal of Physics Conference Series*, 1744(3), 032138.
- [2] Zhang, Y. , Zhao, X. , Shen, J. , Shi, K. , Yu, Y. , & Shi, G. . (2021). Optimization of sports event management system based on wireless sensor network. *Journal of Sensors(Pt.8)*, 41(2), 156-165.
- [3] Zhang, S. , & Liu, M. . (2021). Computer aided management system of sports horse registration based on distributed storage system and deep fusion learning - sciencedirect. *Microprocessors and Microsystems*, 40(3), 740-752.
- [4] Mu, X. , Yin, J. , Zhang, L. , & Jan, N. . (2022). Application and evaluation of sports event management method based on recurrent neural network. *Mathematical Problems in Engineering: Theory, Methods and Applications*, 33(1), 7-9.
- [5] Liu, J. , Zhao, Q. , Yao, Y. , Wang, J. , & Gu, Y. . (2021). Data organization method of distribution network management system based on geographic information system. *Proceedings of the 2021 5th International Conference on Electronic Information Technology and Computer Engineering*, 112(5), 651-653.
- [6] Babenko, V. , Baksalova, O. , Prokhorova, V. , Dykan, V. , & Chobitok, V. . (2021). Information and consulting service using in the organization of personnel management. *Studies of Applied Economics*, 38(4),74-78.
- [7] Yamashita, N. , & Hori, A. . (2023). Two-stage distributionally robust noncooperative games: existence of nash equilibrium and its application to cournotnash competition. *Journal of Industrial and Management Optimization*, 19(9), 6430-6450.
- [8] Saukh, I. , & Vikarchuk, O. . (2021). Creativity in management and creative management: meta-analysis. *Marketing and Management of Innovations*,96(1), 65-80.
- [9] Doan, T. , Manenti, F. M. , & Mariuzzo, F. . (2023). Reprint of: platform competition in the tablet pc market: the effect of application quality. *International Journal of Industrial Organization*, 25(1), 78-85.
- [10] Zinina, O. V. , Sharopatova, A. V. , & Olentsova, J. A. . (2021). Management of an agricultural organization based on building a quality management system. *IOP Conference Series: Earth and Environmental Science*, 677(2), 022029 (7pp).
- [11] Prasad, S. , & Tanase, S. . (2021). Competition, collaboration and organization design. *Journal of Economic Behavior & Organization*, 183(2), 1-18.
- [12] Liu, Y. . (2021). Characteristics of organization and management of private colleges and universities based on the big data analysis. *Journal of Physics Conference Series*, 1744(4), 042090.
- [13] Huang, H. , & Tan, X. . (2021). Application of reinforcement learning algorithm in delivery order system under supply chain environment. *Hindawi Limited*, 25(1), 85-116.

- [14] A, W. J. , & A, Z. S. . (2021). Design of sports course management system based on multi-core embedded system and motion capture - sciencedirect. *Microprocessors and Microsystems*, 82.
- [15] Fonti, F. , Ross, J. M. , & Aversa, P. . (2023). Using sports data to advance management research: a review and a guide for future studies. *Journal of management*, 18(2), 203-243.
- [16] Udroi, A. M. , Sandu, I. , & Dumitrache, M. . (2022). Integrated information system for the management of activities in the organization. *Studies in informatics and control*, 38(2), 255-266.
- [17] Kombate, B. , Emmanuel, M. , & Richard, K. K. . (2021). The implication of the strategic implementation style and middle management effort in public organization strategic management implementation and its organizational performance. *Journal of Public Administration and Governance*, 11(1), 1.
- [18] Song, C. . (2022). Analysis of the organization, management and operation of blantyre water board in malawi: future reforms and perspectives. *Environmental quality management*, 14(3), 494-523.
- [19] Parabakaran, D. , Bin, M. , & Lasi, A. . (2021). Human resource management practices and its impact on employee engagement and performance in an organization a study on labour force in malaysia. *Malaysian E Commerce Journal*, 4(1), 29-35.
- [20] Shust, O. , Grinchuk, Y. , Paska, I. , & Tkachenko, K. . (2021). Investment attractiveness in the system of management and business reputation of agricultural enterprises. *Ekonomika ta upravlinn APK*, 29(3), 23-42.

Edited by: Shaofei Wu

Special issue on: Deep Learning in Healthcare

Received: Dec 20, 2023

Accepted: Dec 30, 2023



THE INTEGRATION OF PERSONALIZED TRAINING PROGRAM DESIGN AND INFORMATION TECHNOLOGY FOR ATHLETES

PENGHUI HAO* AND KUN QIAN†

Abstract. In order to integrate the training of athletes with information technology, this paper proposes a method for evaluating the performance of athletes using training technology. HR, O₂, and hemoglobin were selected as input vectors of SVM, and the corresponding values were used as outputs to generate the training model. Adjust the support vector machine to minimize measurement error based on the learning objective. Support vector machines are used to study training patterns and develop models to evaluate athletes' performance. College athletes were taken as research subjects and the effects of training were examined in five sports: football, basketball, basketball, swimming, and running. The results show that the relative error of this type is less than 1 when measuring the performance of various sports subjects; Relative errors in measuring the academic performance of athletes in various sports using physical fitness standards and sport-specific skills 1. The proposed model proves that the athlete's training index is incorrect and has a useful application.

Key words: Machine learning algorithms, Training effectiveness, Support Vector Machine, High dimensional feature space, Indicator matrix

1. Introduction. With the rapid development of information technology, its application in various fields is becoming increasingly widespread. In the field of sports training, the design of personalized training programs is crucial for the growth and development of athletes. Traditional training programs are often based on general templates and cannot meet the individual differences and special needs of each athlete. Therefore, how to integrate information technology with the design of personalized training programs for athletes has become a focus of current research. The application of information technology in the field of sports training has achieved significant results [1]. The design of personalized training programs for athletes needs to fully consider factors such as their physical condition, technical level, and training objectives, in order to develop the most suitable training plan for them. Traditional training program design is usually formulated by coaches based on experience and routines, but this approach cannot meet the unique needs of each athlete.

The application of information technology provides new possibilities for the design of personalized training programs. By utilizing sensors, monitoring devices, and data analysis techniques, real-time physical data and performance of athletes can be obtained. These data can help coaches more accurately evaluate the training effectiveness of athletes and adjust training plans in a timely manner. For example, sports tracking devices can record the athlete's movement trajectory and speed, heart rate monitoring devices can understand the athlete's physical condition, and strength testing devices can evaluate the athlete's muscle strength.

The analysis and integration of these data can provide targeted training suggestions for coaches, helping them develop personalized training plans. In addition, information technology can also provide online teaching platforms, allowing athletes to train anytime, anywhere. Through video teaching, virtual reality and other technological means, athletes can receive guidance and guidance from professional coaches to improve the effectiveness of training. This approach not only saves time and costs, but also breaks geographical restrictions, allowing more athletes to benefit from professional guidance [2].

In addition, information technology can also provide a platform for coaches and athletes to communicate and share. Through social media, online forums, and other channels, athletes can communicate with other

*Police Skills and Tactical Training Department, Criminal Investigation Police University of China, Shenyang, 110035, China (Corresponding author, hph333666@163.com)

†College of Sports Science, Shenyang Normal University, Shenyang 110034, Liaoning, China (Corresponding author, qiankun20232@163.com)

athletes and coaches, share experiences and insights. This way of communication and sharing can stimulate innovation and cooperation, promote individual growth and team development of athletes. However, information technology also faces some challenges in the design of personalized training programs. Firstly, the accuracy and reliability of data are the core issues [3]. The body data of athletes needs to be collected through reliable sensors and devices, and the accuracy and stability of these devices have a significant impact on the credibility of the data. In addition, data analysis and processing require professional knowledge and technical support to ensure the extraction of useful information and guidance from massive amounts of data. Secondly, privacy and security issues also need to be taken seriously. The personal data involved in the design of personalized training programs needs to be legally and securely protected to avoid abuse or leakage. Relevant privacy policies and security measures need to be fully developed and implemented to safeguard the rights and interests of athletes and coaches.

In summary, the application of information technology in the field of sports training has brought new opportunities and challenges to the design of personalized training programs [4]. By using sensors, monitoring devices, and data analysis techniques, it is possible to more accurately evaluate the condition and training effectiveness of athletes, and develop personalized training plans for them. The establishment of online teaching platforms and communication sharing platforms also provides athletes with more learning and communication opportunities. However, data accuracy and privacy security issues need to be taken seriously and addressed. In the future, with the continuous development and innovation of information technology, the design of personalized training programs will be further improved, promoting the growth and development of athletes.

2. Analysis of the Application of Information Technology in Sports Training.

2.1. Information technology drives athlete selection through data-driven and evidence-based approaches. Scientific selection is based on the disciplines of life sciences such as biochemistry, biomechanics, and genetics, and utilizes modern scientific and technological means to comprehensively select athletes in terms of their physical form, physiological functions, physical fitness, psychological qualities, and sports skills according to their own characteristics and project characteristics [5]. It scientifically and reasonably selects outstanding athletes for high-level sports training. By combining modern information technology with biological data models, we aim to develop material selection criteria that are consistent with project characteristics, and strive to achieve rational, data-driven, and scientific material selection.

2.2. Information technology enables sports assistive devices to have intelligent brains. With the rapid development of modern science and technology such as electronic information technology, intelligent sensing technology, internet big data, cloud computing, artificial intelligence technology, etc., primitive traditional sports and fitness equipment that used to have no emotional color and no data feedback, such as barbells, treadmills, power extenders, etc., now have a "brain" that can interact and communicate with sports participants through the implantation of information technology, and automatic detection of physical fitness indicators functions such as automatic provision of training plans, real-time monitoring of exercise processes, and automatic feedback evaluation of exercise effects [6].

2.3. Information technology enhances the diversification of sports training forms and promotes scientific transformation. Information technology, as a product of technological development, has rapidly penetrated into people's daily work and life, and its impact is gradually expanding. The concept of the Technology Olympics not only increases people's expectations for modern sports events, but also makes them more visually appealing. The use of information technology for auxiliary training can make the training methods more scientific, diverse, and effective. In a sense, it is technology that makes sports more attractive, keeps sports up with the times, and makes sports more meaningful. Coaches use information processing technology to quickly and efficiently develop training plans, outlines, and lesson plans; Reasonably utilizing text processing technology, image editing technology, and video capture technology, the training content and methods are transmitted to the athlete's audio-visual system through technological means, improving the training experience in a new way, stimulating training motivation, and enhancing training effectiveness [7].

As one of the sports powerhouses, the United States has achieved impressive results in applying computer information processing technology to various fields of track and field training. For example, by loading a video into the system, the trajectory of the high jump athlete's movement before takeoff, the takeoff point, and

various posture changes, angles, speeds, etc. of the body after takeoff are recorded [8]. These data are then compared with various data from world-renowned high jump athletes in the database to identify gaps, identify problems, and provide solutions. Through repeated recording and playback of technical movements, combined with simulation teaching of information technology, athletes can intuitively, three-dimensional, actively, and comprehensively master standardized and correct technical movements, thereby improving the scientific and effective nature of sports training, reducing the training cycle of athletes, improving the success rate, and effectively increasing the scientific output of sports results. The application of information technology can not only improve training effectiveness, enhance competitive level, refresh sports performance, but also reduce sports injuries and extend sports lifespan. Information technology has broken the traditional mode of sports competition and training. Traditional sports training requires coaches to conduct planned, purposeful, and organized speech and actions through language or physical means. However, various data during training often better illustrate the actual competition ability of athletes.

3. Evaluation of Athlete Training Effectiveness Based on Machine Learning Algorithms.

3.1. Machine learning algorithms. Let $F(z)$ be a probability measure that exists in space z , with a set of functions $Q(z,a)$ and $a \in \Lambda$, which can achieve the goal of minimizing the risk functional or machine learning. The formula is as follows:

$$R(a) = \int Q(z, a)dF(z) \tag{3.1}$$

In the formula, the probability measure $F(z)$ is unknown, but there are fixed independent distribution samples. The formula for obtaining the loss function is as follows:

$$L(y, f(x, a)) = \begin{cases} 0, & y = f(x, a) \\ 1, & y \neq f(x, a) \end{cases} \tag{3.2}$$

Using the hazard function to justify the function $f(x, a)$ and the probability of exit from the trainer, it is necessary to obtain the function with the minimum distribution error based on the knowledge configuration.

The support vector machine method finds the final model result with the best learning ability, low complexity and high generalizability based on limited sample data. Support vector machine has a high degree of generality, suitable for large-scale, small-sample and non-data fields. Choosing support vector machines as a way to measure sports performance can fully understand the advantages of support vector machines for processing small data. By obtaining the best surface distribution from a small sample size, the surface distribution obtained was the best, reducing the cost of sports performance analysis.

Use (x_i, y_i) to represent training data, and satisfy $i = 1, 2, \dots, l, x \in R^d, y \in \{1, -1\}$, where l represents the number of samples. Assuming the existence of hyperplane H , there is the following formula:

$$w \cdot x + b = 0 \tag{3.3}$$

Using hyperplane H to separate positive and negative data, the hyperplane spacing formula is as follows:

$$\frac{2}{\|w\|} = d_1 + d_2 \tag{3.4}$$

In the formula, the euclidean norm of w and the two types of samples closest to H are represented by $\|w\|$ and d_1, d_2 , respectively. Using support vector machine method to search for hyperplanes with the maximum interval, the expression is: $\min_{w,b} \frac{1}{2}\|w\|^2, s.t. y_i(x_i \cdot w + b) - 1 \geq 0$, in the formula, $i = 1, 2, \dots, l$. The optimal classification surface problem is transformed into its dual problem through Lagrangian optimization method, and the formula obtained is as follows:

$$Q(\alpha) = \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) - \sum_{i=1}^l \alpha_i \tag{3.5}$$

Using $\sum_{i=1}^l y_i \alpha_i$ and $\alpha_i \geq 0, i = 1, 2, \dots, l$. as constraints, use the Lagrange multiplier α_i corresponding to the sample to solve for the minimum value of Equation 3.5. The sample corresponding to α_i that is not 0 in the obtained result is the support vector[9].

The final optimal classification function is as follows:

$$f(x) = \text{sgn}\{(w \cdot x) + b\} = \text{sgn}\left\{\sum_{i=1}^l y_i \alpha_i (x_i \cdot x) + b^*\right\} \tag{3.6}$$

In the formula, b^* represents the classification threshold.

When the experimental sample is linearly inseparable, add a relaxation term; The method for obtaining $\xi_i \geq 0$ in Equations 3.5 and 3.6 is as follows:

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \xi_i, \text{ s.t. } y_i(x_i \cdot w + b) - 1 + \xi_i \geq 0 \tag{3.7}$$

If the penalty factor $C > 0$ for the degree of punishment for correctly and wrongly divided samples is constant, then the dual problem condition is transformed into $0 \leq \alpha_i \leq C$.

The objective function obtained is as follows:

$$Q(\alpha) = \frac{1}{2} \sum_{i,j=1}^l \alpha_i \alpha_j y_i y_j K(x_i \cdot x_j) - \sum_{i=1}^l \alpha_i \tag{3.8}$$

At this point, the optimal classification function is as follows:

$$f(x) = \text{sgn}\left\{\sum_{i=1}^l y_i \alpha_i K(x_i \cdot x) + b^*\right\} \tag{3.9}$$

Different types of nonlinear decision surface support vector machines can be implemented through differential kernel functions. The radial basis kernel function is selected as the kernel function for evaluating the training effectiveness of athletes, and its formula is as follows:

$$K(x_i \cdot x) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right) \tag{3.10}$$

3.2. Evaluation of Athlete Training Effectiveness. The physiological information during athlete training can reflect the training effect and showcase more information that the previous athlete training evaluation system could not evaluate.

Use $X = \{x_1, x_2, \dots, x_n\}$ and $Y = \{y_1, y_2, \dots, y_n\}$ to represent the athlete training sample set and the test indicator set that can reflect the physiological indicators of athlete training effectiveness, such as heart rate, oxygen uptake, hemoglobin, creatine kinase, etc. The matrix $A = (a_{ij})_{n \times m}$ represents the indicator matrix of the athlete training sample set X for the measurement indicator set Y, and the indicator values of the athlete training sample x_i for the measurement indicator y_j are represented by $a_{ij} = y_j(x_i) (i = 1, 2, \dots, n; j = 1, 2, \dots, m)$.

The input unit for evaluating athlete training effectiveness is the athlete training sample x_i , and the measurement vector $(r_{i1}, r_{i2}, \dots, r_{im})$ under physiological measurement index y_j . If the evaluation result u_i of athlete training effectiveness x_i is used as the output unit, there exists a non-linear mapping between the normalization matrix R and the evaluation result, as shown in F, and its formula is as follows:

$$u_i = F(r_{ij}) \tag{3.11}$$

Select athlete training effect sample $x_i(r_{i1}, r_{i2}, \dots, r_{im})$ as the input vector of the support vector machine, select training effect sample evaluation value as the regression target value U of the support vector machine, establish a learning sample set, represented by $G = \{(x_i, u_i)\}_i^n$, and obtain the regression function as follows:

$$u = \sum_{k=1}^s (\alpha_k - \alpha_k^*) K(x, x^k) + b \tag{3.12}$$

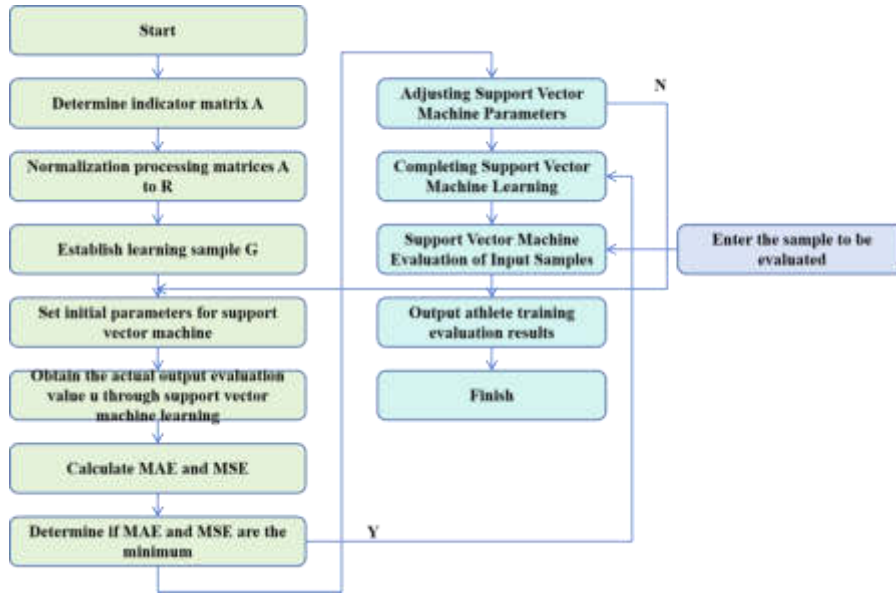


Fig. 3.1: Structural diagram of athlete training effect assessment

In the formula, X^k and s represent support vectors and the number of support vectors, respectively, $x^k = (r_{k1}, r_{k2}, \dots, r_{km}), k = 1, 2, \dots, s$.

Implement athlete training sample x_i through the above process; The non-linear mapping F between the measurement vector $(r_{i1}, r_{i2}, \dots, r_{im})$ of physiological indicators y_j and the evaluation value u_i of training effectiveness[10].

The structure diagram of evaluating athlete training effectiveness using the support vector machine method in machine learning algorithms is shown in Figure 3.1.

The process of evaluating the effectiveness of athlete training is as follows:

1. Determine the evaluation index matrix A for athlete training effectiveness based on physiological indicators that affect athlete training effectiveness;
2. Convert the indicator matrix A for evaluating the training effectiveness of athletes to a normalized matrix R ;
3. Establish a learning sample set $G = \{(x_i, u_i)\}$ using athlete training effect sample $x_i = (r_{i1}, r_{i2}, \dots, r_{im})$ and evaluation value u_i , and randomly select samples from the learning sample set to establish a training and validation set for support vector machine learning.
4. Select the radial basis kernel function as the support vector machine kernel function to obtain the regression function formula. The criteria for selecting parameters for evaluating the training effectiveness of athletes are as follows:

$$\begin{cases} MAE = \frac{1}{l} \sum_{i=1}^l |u_p^i - u_{SVM}^i| \\ MSE = \frac{1}{l} \sum_{i=1}^l (u_p^i - u_{SVM}^i)^2 \end{cases} \quad (3.13)$$

In the formula, MAE and MSE represent the average absolute error and mean square error of the validation samples, u_p^l and u_{SVM}^l represent the expert evaluation value and support vector machine calculation value of the validation samples, and l represents the total number of samples to be evaluated.

5. On this basis, SVM is used to evaluate the training results. By inputting the physiological testing indicator vector $(r_{i1}, r_{i2}, \dots, r_{im})$ of the training effect sample x_i of the athlete to be evaluated, the final support vector machine's athlete training effect evaluation result u_{SVM}^i can be obtained.

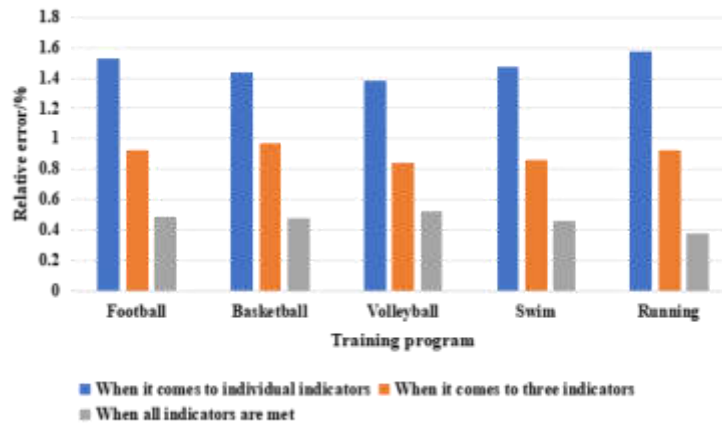


Fig. 4.1: Relative error in the number of different indicators

4. Example Analysis. Ten athletes majoring in sports training from a certain sports university were selected as the research subjects, and they were trained in five sports: football, basketball, volleyball, swimming, and running. This method was used to evaluate the training effects of the ten athletes. Heart rate and lung capacity were collected before and after each exercise training, and morning venous blood was collected for the training cycle.

Out of a total of 10,000 samples collected, 2,000 were used as research samples and the remaining 8,000 were used as test samples. Select radial root kernel function as support vector machine kernel function, use winSVM software to solve support vector machine problem, this software is optimized for Windows operating system, fully support vector machine classification and regression problem.

The training samples are divided into 10 groups of 200 each, and finally the parameters of the support vector machine are determined as $C=150$, $\sigma^2 = 0.016$, and the final decision is used to train the support vector to carry the vector and get the number of support vectors and the parameter b . 30 and -0.228 for the regression function [11]. To assess sports performance, heart rate, maximal oxygen uptake, hemoglobin, creatine kinase, and blood lactate are selected as physical parameters. This method was used to estimate the relative error of measuring the academic performance of athletes in various sports when only heart rate, heart rate, maximal oxygen uptake, and hemoglobin were used as measures of physical fitness five physical parameters used. From the experiments in Figure 4.1, it can be seen that as the number of physical parameters increases, the relative error of sports evaluation results decreases. This suggests that adding more physical parameters to the machine learning algorithm can improve the accuracy of sports analysis. The main reason is that many physical parameters improve the various results of sports analysis and improve the accuracy of sports analysis [12,13,14,15].

The output results of evaluating the training effects of 5 sports for 10 athletes using this method are shown in Figure 4.2.

Based on the evaluation of sports training results shown in Figure 4.2, the results of expert evaluation were selected as the correct decision criteria for sports performance evaluation, and five types of sports were analyzed in this type of performance analysis: football, basketball, and basketball. swimming and archery. To justify this type of performance assessment, physical activity and sport-specific skills were chosen as the model for comparison. Figure 4 shows a comparison of the relative errors in measuring the fitness index of athletes in five sports using different models[16,17,18,19]. From the comparison results in Figure 4.3(a)-(e), it can be seen that when selecting the expert evaluation the basis of performance evaluation, the relative error of this type is less than 1 when evaluating the training of athletes of various sports; Relative errors in measuring academic performance of athletes in various sports using fitness standards and sport-specific skills 1. Comparison of the results shows that this type has a good performance in sports evaluation. Analyzing the results of this type

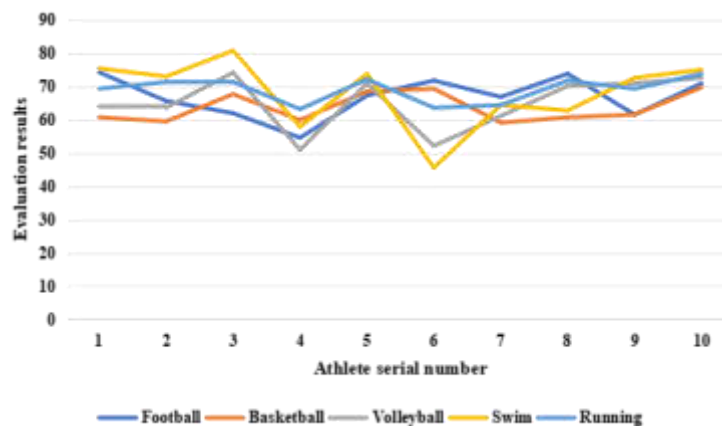


Fig. 4.2: Results of the training effect evaluation of the athletes

of measurement can help coaches develop training plans for soccer performance. This type can be used to objectively evaluate soccer performance.

5. Conclusion. The training effect of athletes determines their final competitive performance, which is influenced by factors such as training methods and personal qualities, resulting in a wide range of fluctuations in athlete performance. Evaluate the training effectiveness of athletes using physiological indicators, and achieve the evaluation of athlete training effectiveness through the support vector machine method with higher evaluation performance in machine learning algorithms. Support vector machines have high learning and generalization abilities, the effectiveness of using the proposed method to evaluate the athletic performance of athletes through sports such as football has high guiding significance for improving the training effectiveness of athletes.

REFERENCES

- [1] Li, X. (2022). Design of industrial logistics information integration method based on supply chain management. *International Journal of Manufacturing Technology and Management*, 39(3), 511-534.
- [2] Zhang, Y. (2023). Construction of accounting information system specialized creative integration course based on "financial sharing". *Curriculum and Teaching Methodology*, 78(4), 4767-4785.
- [3] Dong-Bo, W. (2021). Investigation and design of information science education system and talent training. *Information Studies: Theory & Application*, 44(1), 0-0.
- [4] Vandette, M. P., Jones, G., Gosselin, J., & Kogan, C. S. (2021). The role of the supervisory working alliance in experiential supervision-of-supervision training: a mixed design and multiple perspective study. *Journal of psychotherapy integration*, 85(4), 31.
- [5] Cheng, Y., Zhao, X., Wu, J., Liu, H., Zhao, Y., & Shurafa, M. A., et al. (2021). Research on the smart medical system based on nb-iot technology. *Mobile Information Systems*, 11(1), 116.
- [6] Zhao, H., Yang, Z., Chen, C., Liu, Z., & Qi, B. (2023). Optimal design of ladder-stress accelerated degradation test plan for motorized spindle in non-cube test area. *The International Journal of Advanced Manufacturing Technology*, 10(3), 1-34.
- [7] Zhang, Z., & Jia, L. (2023). Optimal multiobjective design of guidance information systems in underground spaces: model development and a transportation hub case study. *Tunnelling and Underground Space Technology*, 134(7), 105007.
- [8] Joshy, C. G., Elavarasan, K., & Zynudheen, A. A. (2021). Design and development of web based information system for value added fish by-products. *Fishery Technology*, 85(1), 58.
- [9] Wang, Z. (2021). Evaluation model of parking equipment planning and design based on object-oriented technology. *Applied Sciences*, 11(3), 74-78.
- [10] Shin, Y. J. (2021). The improvement plan for personal information protection for artificial intelligence(ai) service in south korea. *Journal of Convergence Information Technology*, 11(2), 20-33.
- [11] Tariku, N., & Lessa, L. (2021). Towards a conceptual framework for information technology disaster recovery plan for banks: based on cases from ethiopia. *International Journal of Computer Applications*, 49(4), 314-325.
- [12] Tang, Y., Song, M., Xiao, S., Liu, X., & Liang, G. (2021). Cax integration and its design application based on feature extension of sensor components. *Hindawi Limited*.

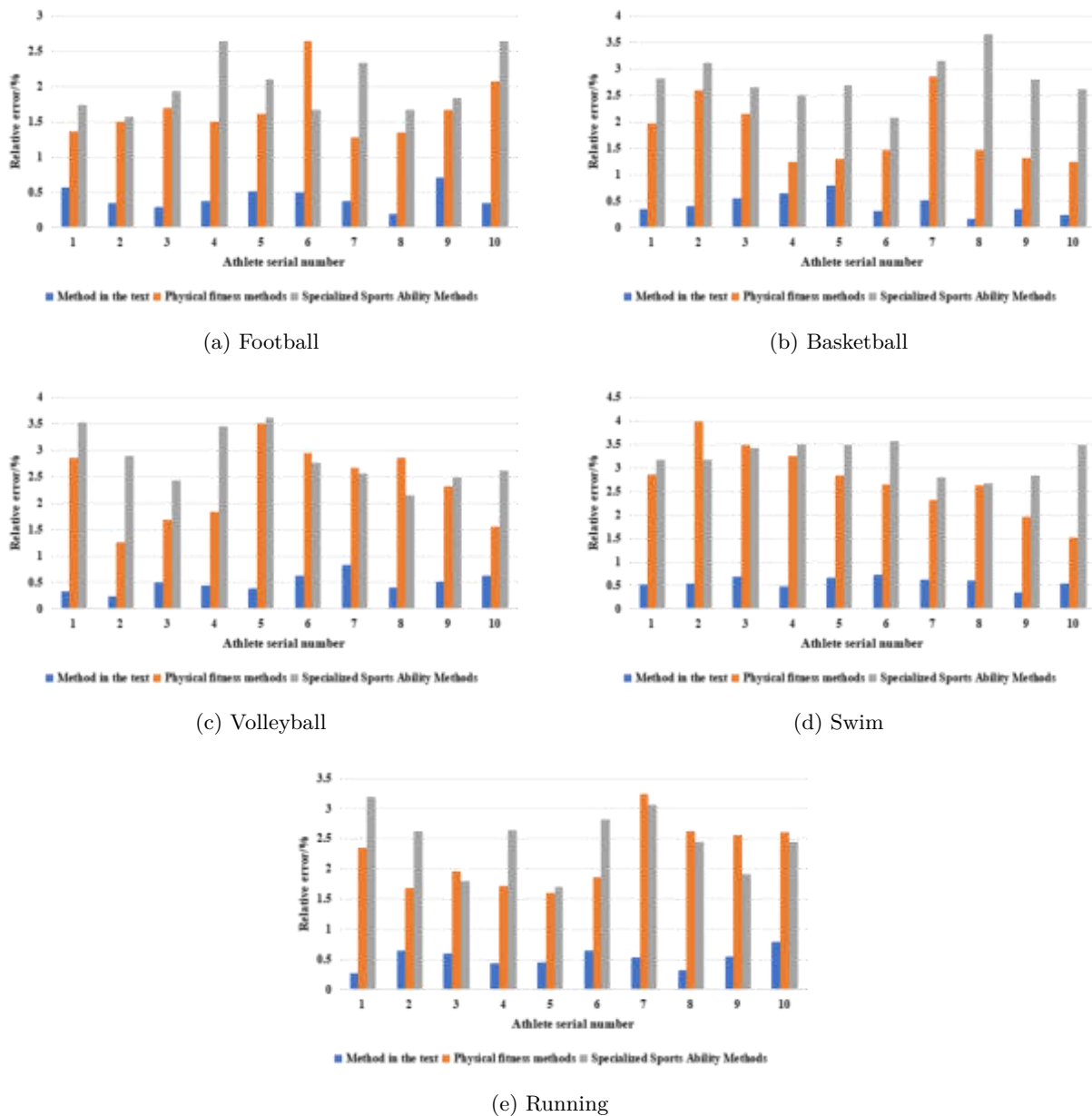


Fig. 4.3: Relative errors were assessed for different exercise items

[13] Embuldeniya, R., & Merabet, A. (2023). Design, modelling and simulation of hybrid vibration energy harvesting system using integration of piezoelectricity and electromagnetism. *Renewable Energy and Power Quality Journal*.

[14] Sasatake, H., Tasaki, R., Yamashita, T., & Uchiyama, N. (2021). Imitation learning system design with small training data for flexible tool manipulation. *International journal of automation technology*(5), 15.

[15] Liu, Y., Wu, H., Lv, H., & Wu, X. (2021). Strategic integration of moo2 onto mn0.5cd0.5s/cu2o p-n junction: rational design with efficient charge transfer for boosting photocatalytic hydrogen production. *Powder Technology: An International Journal on the Science and Technology of Wet and Dry Particulate Systems*,74(394), 394.

[16] Xu, Z., & Zhou, W. (2021). A data technology oriented to information fusion to build an intelligent accounting computerized model. *Scientific programming*,857(Pt.12), 2021.

[17] Abdurahimovna, U. F. (2021). Methodology of training students in design and modeling of clothes using information com-

- munication technologies. *Revista Gesto Inovao e Tecnologias*, 19(5), 416-424.
- [18] Zhigunova, A. (2021). Integration of students' academic and extracurricular activities by creating conditions for business-oriented design and practical activities at the university. *Standards and Monitoring in Education*, 32(7), 21-25.
- [19] Samandari, H., Khave, L. J., Janani, M., Farazmand, S., Motafavi, S. S., & Gholami, J., et al. (2022). Evaluation of a training program in general practitioners' attitude toward the integration of substance use disorders services in primary health care. *Iranian Journal of Psychiatry and Behavioral Sciences*, 37(5), 813-830.

Edited by: Hailong Li

Special issue on: Deep Learning in Healthcare

Received: Dec 27, 2023

Accepted: Jan 14, 2024



DIAGNOSIS AND TREATMENT SYSTEM BASED ON ARTIFICIAL INTELLIGENCE AND DEEP LEARNING

XIAOXI ZHENG*, QILI FAN† AND GENG WANG‡

Abstract. This paper designs an assisted diagnosis and treatment system based on deep learning algorithms and medical knowledge to solve the problem of poor use efficiency of massive electronic medical information. First, the disease data in the medical database is segmented to get the reverse order search table. Secondly, the similarity between the obtained clinical manifestation data and the corresponding diseases is analyzed and classified to obtain the clinical diagnosis. Then, the feedback-query method is used to analyze the weighted ratio of the original and feedback data, and the optimal fault diagnosis is carried out. The method of implicit semantic modeling is used to give the diagnosis scheme of the disease. The search method based on inference rules is introduced to realize personalized diagnosis and treatment resource recommendations to users. In this way, the specific attributes of medical resources based on individual information are effectively combined. Experiments show that the initial diagnosis recognition rate of the proposed method is 95%, the correct rate is 85%, and the recognition rate is 95% after optimization.

Key words: Deep learning; Medical knowledge base; Artificial intelligence; Electronic diagnosis and treatment; Lingo model

1. Introduction. Many patients with complicated diseases will come to first-tier cities to seek better treatment. There are many online information platforms to make the masses better understand the relevant medical information. In the past, access to medical information was mainly based on keyword-based information retrieval. However, due to the rapid development of medical data, this approach rarely meets the needs of patients. Many intelligent recommendation algorithms have recently been applied to medical data retrieval. Literature [1] provides a collaborative screening algorithm for patients with a particular disease, which can effectively solve the problem of personalized diagnosis and treatment of patients. Literature [2] proposes a medical knowledge recommendation method based on cooperation, which generates a trusted factor through the evaluation of medical services and introduces the trusted factor into the recommendation algorithm of joint filtering to realize personalized recommendations for physicians. This method not only overcomes the "information overload" of doctors but also improves the recommendation effect. Literature [3] extracts user data from user text, constructs a medical data knowledge map, and then carries out medical service resource recommendations based on collaborative filtering. Hidden semantic models such as matrix decomposition, probabilistic cryptology analysis, and implicit Dirichlet distribution effectively find hidden data. Reference [4] applies matrix decomposition to the prescription problem to discover the combination mode. However, the biggest drawback is that its parameter scale depends on the sample size, which can easily cause overfitting. Reference [5] provides the implicit semantic analysis method. They use the hidden tree method to find the hidden structure of the disease and construct the objective diagnostic criteria. The study used only disease data, not drug data, which is integral to the dialectics. Previous studies focused on the analysis of clinical manifestations. Yet, each case includes a set of symptoms and a set of medications. This project intends to use implicit-semantic modeling and case analysis methods to realize the association analysis of drug - pathogenesis. Then, the network platform provides personalized medication recommendations for patients.

2. Diagnosis and treatment based on the medical knowledge base. Reference [6] gives an architecture for assisting disease diagnosis and treatment. The system mainly includes the following parts: 1) Analysis of disease information through the characteristics of the disease segmentation operation to analyze the disease

*School of Architecture and Environmental Engineering, Zhengzhou Technical College, ZhengZhou, 450121, China (Corresponding author, zxxdtg@163.com)

†School of Automation and Internet of things, Zhengzhou Technical College, ZhengZhou, 450121, China

‡The sixth design and Research Institute of Mechanical Industry Co., Ltd, Zhengzhou 450100, Henan, China

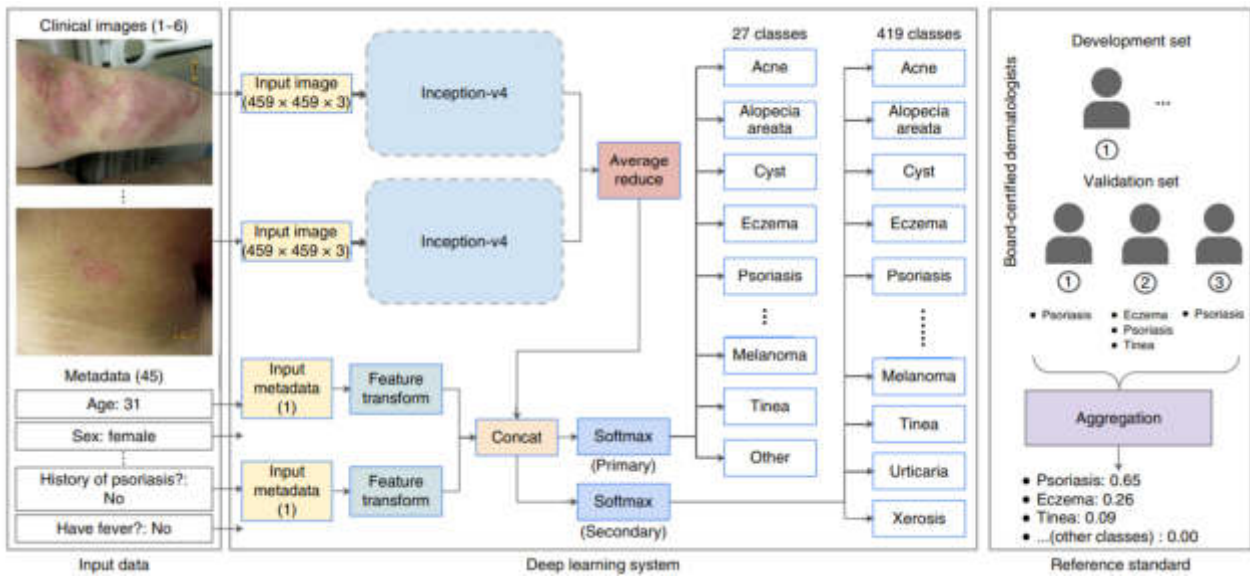


Fig. 2.1: Structure of diagnosis and treatment system based on deep learning.

information; 2) Disease index table an inverted index table can be constructed based on the analysis of disease information. 3) The condition assistance diagnosis module can input patient information into the system and judge the condition. 4) Diagnostic results and information feedback module displays diagnostic results. The optimal diagnosis of the disease is combined with the user’s opinion. 5) The case data analysis module can obtain the detection methods related to the disease. 6) Check the suggestion form to develop the diagnosis method according to the case analysis. 7) Diagnosis The diagnostic suggestion module displays the detection patterns and diagnostic possibilities. 8) The disease association information module displays disease-related information, such as etiology, diagnosis, and treatment mode. The system analyzes the relevant information on the disease in the medical database and generates the disease search form and diagnosis mode suggestion form in the background. Secondly, based on the clinical manifestation data of the patients, the suspected disease was diagnosed, and the correlation analysis was carried out. Then, if the user has feedback on their condition, the optimal diagnosis will be made, and the details of the condition and further tests can be seen.

3. Auxiliary diagnostic technology. This section discusses specific embodiments of the assisted diagnostic techniques described in Figure 3.1.

1. Perform the same segmentation on the clinical manifestations of the input patients and compare the segmentation results with the search table. All eligible cases are entered into the potential outcome set, and the total number of these cases is called N.
2. A symptom index table was used to calculate the degree of correlation of each condition in the potential outcome set for each case and ranked according to the degree of correlation.

Use $S = \frac{1}{t} \sum (q * d)$ to express the correlation function. A similarity measurement method based on the vector space model is proposed starting from the internal product method of the vector space model. Here S represents the degree to which the patient’s condition is related to the input. t is the number of types of diseases occurring in the disease, which is used to balance the matching probability of diseases with more significant symptoms in the disease database [7]. This solves the problem that the weight of the disease is too large due to the wide range of the disease. Where q is the proportion of each condition in different keywords. Treat the condition as the first possible condition. d represents the weight of each condition, again increasing the weight of the likely primary condition [8]. The correlation degree obtained by this function does not have a specific upper limit. For the convenience of calculation, the maximum and minimum values are standardized, so the correlation degree presented in the end is a relative value.

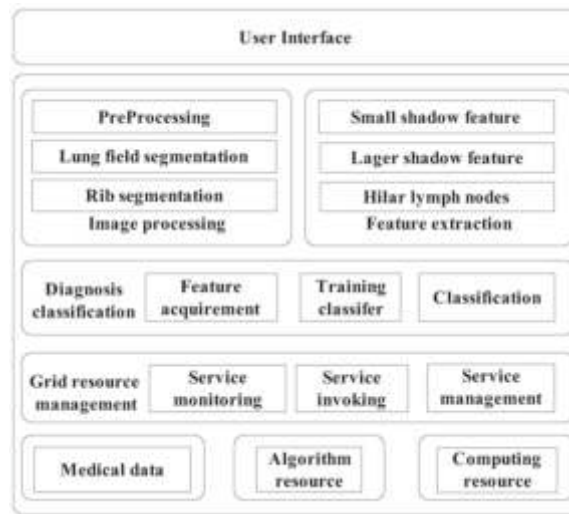


Fig. 3.1: Deep learning-based assisted diagnosis and treatment system flow.

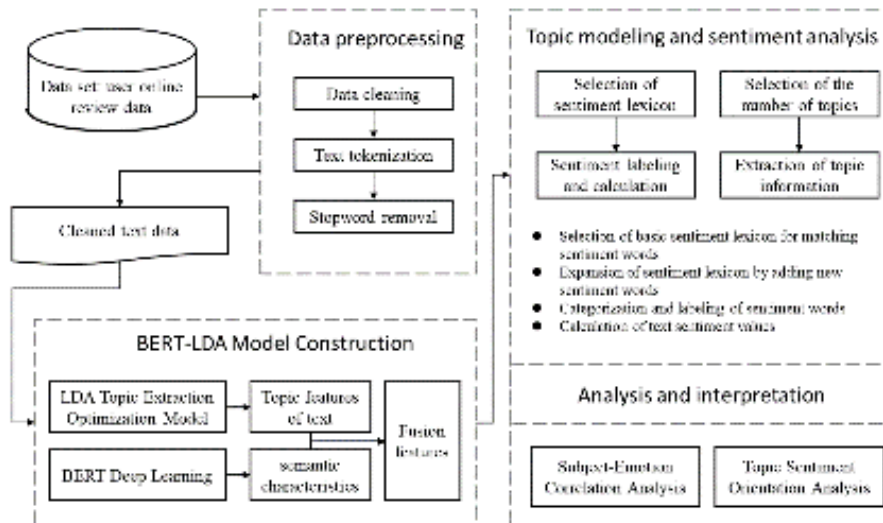


Fig. 4.1: Linear discriminant model flow.

4. The linear discriminant model was used to analyze medical case samples.

4.1. Introduction of Linear Authentication Mode. The implied Dirichlet distribution was first established in the multi-level Bayesian model by Blei et al. in 2003. The linear authentication pattern flow (Healthcare. MDPI, 2023, 11 (15):2142) is shown in Figure 4.1. This method was initially used in word processing and has been widely used in many data mining applications. The linear discrimination model is a typical topic model, which is to build θ and ϕ and discover.

A topic is a set of interrelated words that can be used to illustrate the topic. The words in the document are expressed in terms of conditional possibilities related to the topic [9]. Each topic contains words that have a high probability and are highly relevant to that topic. One word can have many different topics at the same time. Therefore, this paper proposes the "document-topic-vocabulary" correlation based on the generation

pattern. The method produces a topic with a certain probability and then a word with a certain probability. The probability of each word γ appearing in A document s can be calculated by the following formula:

$$| f(\gamma | s) = \sum_t f(\gamma | t)f(t | s)$$

t is the topic of this passage. In linear discrimination mode, the text contains the assignment of topics, and the topic contains the assignment of keywords [10]. Think of each prescription as a document that records the patient's condition and the medication the doctor prescribes. If symptoms and drugs are treated as "words" without distinction, they can be clustered through a linear identification pattern. Each group in the cluster contains several words, which can be a disease or a drug. Theoretically, explaining the rationality of such a cluster model is also problematic. Etiology is the subject of the model in the diagnosis problem related to conditions and both types of literature. If the two documents can be classified, establishing a link between the disease and the cause can better describe the problem.

4.2. Multiple linear discrimination model. Because of the existing methods in the aspect of data analysis of the connotation of the project plans to build a more linear differential model (figure 4.2) and reveal the whole process of production (PeerJ Computer Science, 2023, 9: e1016), the case will be regarded as files, including disease vocabulary r and γ two different types of vocabulary [11]. Given the pathogenesis, the distribution of these two types of words is independent and determined by the subject of the prescription, that is, the pathogenesis. c is a twovariable observation. There are only two values for the font: SYMPTOM or HERB. If it is $c = SYMPTOM$, the resulting word is the symptom word r . If it is $c = HERB$, the resulting word is the medical word γ . This project proposes a linear identification method based on multiple connotations. Case and drug words have the same topic, and there is a certain correlation between the two words under a specific topic. The spatial distribution of the problem c is obtained by using polynomial distribution characteristics. The post-disease word can be obtained according to the polynomial distribution lattice consistent with the problem. The medical word γ is derived from a polynomial distribution that agrees with the topic. It is the polynomial distribution lattice ζ and δ is the Dirichlet priori of χ .

After the distribution parameter $\varepsilon, \varphi, \delta$ is known, the joint probability of the "prescription - disease" distribution β , "disease - disease" distribution ζ , and "disease - drug" distribution χ is obtained as follows:

$$\begin{aligned} f(r, \gamma, c, \varepsilon, \beta, \zeta, \chi | \varepsilon, \varphi, \delta) = \\ f(\zeta | \varphi)f(\chi | \delta)f(\beta | \varepsilon)f(r, \gamma, c, \varepsilon | \zeta, \chi, \beta) = \\ f(\zeta | \varphi)f(\chi | \delta)f(\beta | \varepsilon)f(r, \gamma, c, \varepsilon, \zeta, \chi) \\ f(c | \varepsilon, \beta)f(c) \end{aligned}$$

The Gibbs sampling method is used to identify the parameters in the system [12]. In the multi-connotation linear identification model, the possibility of disease $f(r_i | r'_i)$ and drug $f(\gamma_i | \gamma'_i)$ belonging to topic $c = \{1, 2, L, K\}$ was calculated for the drugs for disease $r = (r_1, r_2, L, r_S)$ and disease $\lambda = (\lambda_1, \lambda_2, L, \lambda_W)$ in the observed samples respectively, and iteration was carried out for each stage.

5. Data online assisted diagnosis and treatment algorithm. A multi-connotation linear identification method obtained the corresponding relationship between disease and drug. A reasonable medication plan can be obtained using the established mathematical model for clinical diagnosis and then taking the clinical manifestations of patients as input. First, the "pathogenesis" is extracted based on the known disease characteristics, and then the drug is administered according to the cause [13]. The second method establishes a "disease-drug" matrix to obtain the corresponding relationship between each disease and each drug. In order of the conditions provided, get the corresponding medication recommendations. Therefore, this project intends to adopt a hybrid model to merge the two algorithms and improve the recommendation accuracy.

5.1. Pathology-based drug recommendation methods. Assuming that the clinical manifestation of the patient is $r = (r_1, r_2, L)$, the "disease-topic" distribution ζ , "drug-topic" distribution χ and prior parameters obtained by the multi-connotation linear identification model, then the etiological drug recommendation method can be expressed as:

Enter: A list of the patient's symptoms.

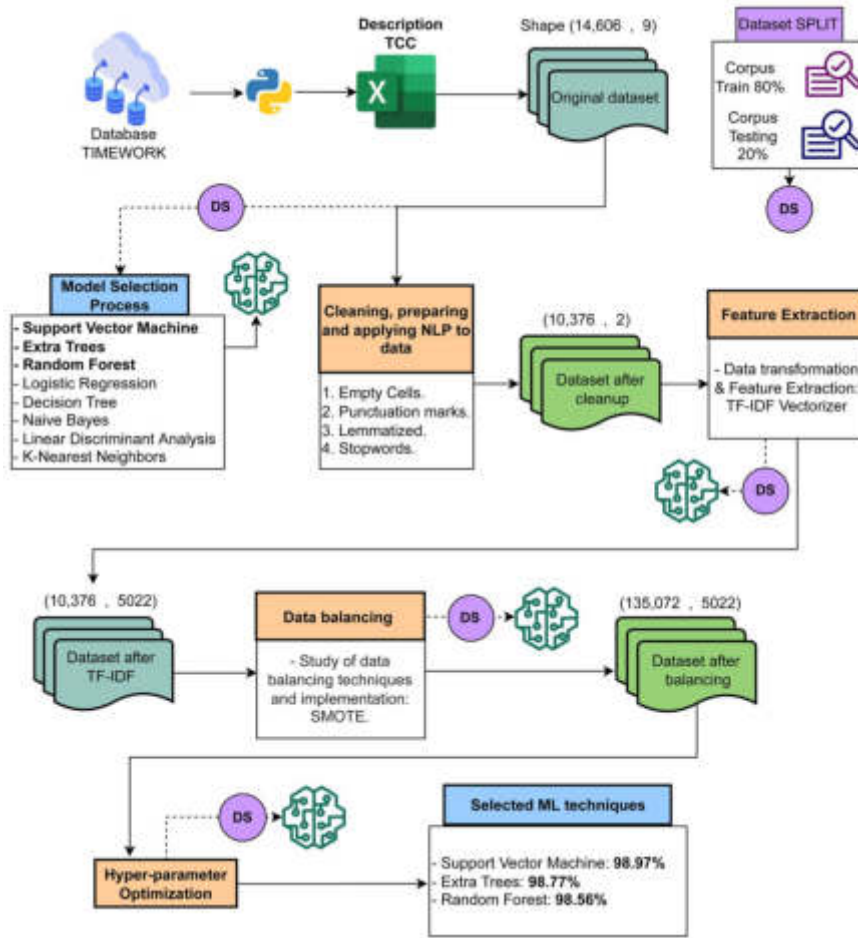


Fig. 4.2: Multi-content linear discriminant probability graph model.

Output: According to the patient W disease, medication is recommended.

1. Random initialization: Each current disease r will be randomly assigned a topic number c .
2. According to the sampling equation, the current file is tested again. For each disease r , the topic is sampled again.
3. The above treatment is repeated until the sampling convergence is achieved.
4. This paper determines the pathogenesis of the disease based on the distribution of the obtained topics. Top N drugs related to this disease are listed as recommendations [14]. First, the clinical manifestations of patients are regarded as the new document topic, then the disease information in the new literature is obtained by inference algorithm, and then the drug related to this disease is selected according to probability. The sampling equation to be used in the second step of the method is:

$$f(c_{i_r} = t \mid c = SYMPTOM, c_{-i_r}, r, \gamma) \propto \frac{W_{r_m}^{(t), -i_r} + W_{\gamma_m}^{(t)} + \varepsilon_t}{\sum_{t=1}^K (W_{r_m}^{(t), -i_r} + W_{\gamma_m}^{(t)} + \varepsilon_t) \sum_{f=1}^S (W_{r_t}^{(f), -i_r} + \varphi_f)}$$

$W_{r_m}^{(t)}$ and $W_{\gamma_m}^{(t)}$ represent the symptoms section in document m . The number of diseases listed in the

title t and the number of drugs used. $W_{r_t}^{(f)}$ is the number of disease words with an average f number in the topic t .

5.2. Drug recommendation methods based on clinical manifestations. The process of drug recommendation based on disease is shown in Figure 5.1, with a set of conditions and weights of patients as their inputs. It can be expressed as an n -dimensional vector. The default value is 1 ; otherwise, it is 0 . When the medical record rights are entered, the corresponding value is the input value [15]. This method expresses the correspondence between each disease and each drug by establishing a "disease-drug" matrix. Finally, according to the clinical manifestations of the patients, the most closely related to the clinical manifestations of the drug regimen.

Input: A set of symptoms for the patient and the proportion of each symptom.

Output: According to the patient W disease, medication is recommended.

1. Establish an "etiology - drug use" model. The information in column j , row i , of the matrix is $f(\gamma_j | r_i)$, which is the likelihood of taking the drug j when disease i occurs.
2. The rank value of each drug was calculated according to the clinical manifestations of patients and corresponding weights.
3. The drugs were ranked according to the rank, and the top W ranked drugs were recommended to patients. The formula used to construct the first step of the matrix is:

$$f(\gamma_j | r_i) = \sum_{c=1}^K f(\gamma_j | c) \cdot f(c | r_i)$$

c represents the topic in multinomial linear discrimination mode. K is the number of topics. The expression of $f(\gamma_j | c)$ as the parameter χ , $f(c | r_i)$ in the multi-connotation linear discrimination model is as follows:

$$f(c | r_i) = \frac{f(r_i, c)}{f(r_i)} = \frac{f(r_i | c) \cdot f(c)}{f(r_i)} \propto f(r_i | c) \cdot f(c)$$

$f(r_i | c)$ is the parameter ζ in the multi-content | model. $f(c)$ is the subject of the disease, which is obtained by the model derivation process. The grade value of each drug γ_j is calculated using the following formula:

$$\text{rank}(\gamma_j) = \sum_{i=1}^S f(\gamma_j | r_i) \cdot \text{weight}(r_i)$$

$\text{weight}(r_i)$ is a weighting of symptoms r_i provided by the user.

By default, conditions that have been typed have a weight of 1, and conditions that have not been typed weight of 0.

The first medication method analyzes the pathogenesis based on the disease and prescribes reasonable medication according to the cause. The latter is to give the corresponding treatment plan for the patient's condition [16]. The focus of the two types of treatment is different. In treatment, the doctor will start with the cause and then prescribe the corresponding drug to the patient. This paper intends to design a hybrid recommendation algorithm. Take the intersection of the two as the final recommendation.

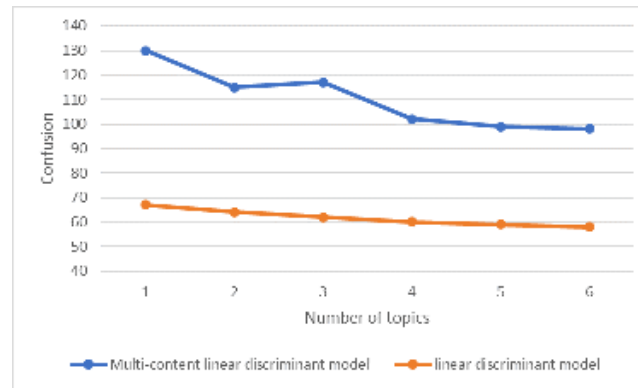
6. Experimental results and analysis.

6.1. Data Sets. 900 cases were obtained by eliminating invalid samples based on the information of 1000 patients in the modern pharmacological database [17]. Instead of measuring each case, the paper classifies each case into a class or category of diseases.

6.2. Baseline Method. The benchmark method chosen in this article is the most popular drug recommendation method. Recommend the most popular products to customers [18]. The drug recommendation questions were all for cases of lung cancer patients, so several commonly used drugs were found based on the analysis of these cases.

Table 6.1: *Statistics of modern medical records.*

Project name	Quantity
Caseload	900
Number of disease species	85
Total number of medicinal materials	123

Fig. 6.1: *Results of confusion degree experiment.*

6.3. Degree of confusion of multiple linear discrimination modes. The results of multi-content linear discrimination are compared with those of traditional linear discrimination. It can be seen from Figure 6.1 that the effect of the multi-content linear discrimination model is significantly better than that of traditional linear discrimination [19]. In addition, the degree of ambiguity decreases with the increase of the number of categories Z , which indicates that the degree of ambiguity tends to converge, which is consistent with the theoretical argument. Increasing Z will not improve the efficacy after Z is large enough to cover all the hidden causes.

7. Conclusion. The improved linear identification model was used to analyze clinical cases and find hidden causes. Find out the internal relationship between recessive cause, syndrome and medication. Two methods of drug recommendation based on symptom were designed using the correlation between symptom, pathogenesis and drug. The algorithm's effectiveness is verified by testing the modern medical case base.

REFERENCES

- [1] Tran, K. A., Kondrashova, O., Bradley, A., Williams, E. D., Pearson, J. V., & Waddell, N. (2021). Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Medicine*, 13(1), 1-17.
- [2] Goecks, J., Jalili, V., Heiser, L. M., & Gray, J. W. (2020). How machine learning will transform biomedicine. *Cell*, 181(1), 92-101.
- [3] Yu, K., Tan, L., Lin, L., Cheng, X., Yi, Z., & Sato, T. (2021). Deep-learning-empowered breast cancer auxiliary diagnosis for 5GB remote E-health. *IEEE Wireless Communications*, 28(3), 54-61.
- [4] Ellahham, S. (2020). Artificial intelligence: the future for diabetes care. *The American journal of medicine*, 133(8), 895-900.
- [5] Rahman, A., Hossain, M. S., Alrajeh, N. A., & Alsolami, F. (2020). Adversarial examples Security threats to COVID-19 deep learning systems in medical IoT devices. *IEEE Internet of Things Journal*, 8(12), 9603-9610.
- [6] Yuan Liu, Ayush Jain, Clara Eng, David H. Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, Vishakha Gupta, Nalini Singh, Vivek Natarajan, Rainer Hofmann-Wellenhof, Greg S. Corrado, Lily H. Peng, Dale R. Webster, Dennis Ai, Susan J. Huang, Yun Liu, R. Carter Dunn & David Coz. A deep learning system for differential diagnosis of skin diseases. *Nature medicine*, 2020, 26(6): 900-908.
- [7] Chang, V., Bailey, J., Xu, Q. A., & Sun, Z. (2023). Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Computing and Applications*, 35(22), 16157-16173.

- [8] Char, D. S., Abràmoff, M. D., & Feudtner, C. (2020). Identifying ethical considerations for machine learning healthcare applications. *The American Journal of Bioethics*, 20(11), 7-17.
- [9] Quer, G., Arnaout, R., Henne, M., & Arnaout, R. (2021). Machine learning and the future of cardiovascular care: JACC state-of-the-art review. *Journal of the American College of Cardiology*, 77(3), 300-313.
- [10] Alimadadi, A., Aryal, S., Manandhar, I., Munroe, P. B., Joe, B., & Cheng, X. (2020). Artificial intelligence and machine learning to fight COVID-19. *Physiological genomics*, 52(4), 200-202.
- [11] Echle, A., Rindtorff, N. T., Brinker, T. J., Luedde, T., Pearson, A. T., & Kather, J. N. (2021). Deep learning in cancer pathology: a new generation of clinical biomarkers. *British journal of cancer*, 124(4), 686-696.
- [12] Karar, M. E., Alsunaydi, F., Albusaymi, S., & Alotaibi, S. (2021). A new mobile application of agricultural pests recognition using deep learning in cloud computing system. *Alexandria Engineering Journal*, 60(5), 4423-4432.
- [13] Khan, S., Barve, K. H., & Kumar, M. S. (2020). Recent advancements in pathogenesis, diagnostics and treatment of Alzheimer's disease. *Current neuropharmacology*, 18(11), 1106-1125.
- [14] Chen, C., Liu, B., Wan, S., Qiao, P., & Pei, Q. (2020). An edge traffic flow detection scheme based on deep learning in an intelligent transportation system. *IEEE Transactions on Intelligent Transportati on Systems*, 22(3), 1840-1852.
- [15] Muhammad, K., Khan, S., Del Ser, J., & De Albuquerque, V. H. C. (2020). Deep learning for multigrade brain tumor classification in smart healthcare systems: A prospective survey. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2), 507-522.
- [16] Oh, Y., Park, S., & Ye, J. C. (2020). Deep learning COVID-19 features on CXR using limited training data sets. *IEEE transactions on medical imaging*, 39(8), 2688-2700.
- [17] Saood, A., & Hatem, I. (2021). COVID-19 lung CT image segmentation using deep learning methods: U-Net versus SegNet. *BMC Medical Imaging*, 21(1), 1-10.
- [18] Alsharif, W., & Qurashi, A. (2021). Effectiveness of COVID-19 diagnosis and management tools: A review. *Radiography*, 27(2), 682-687.
- [19] Li, T., Zhao, Z., Sun, C., Cheng, L., Chen, X., Yan, R., & Gao, R. X. (2021). WaveletKernelNet: An interpretable deep neural network for industrial intelligent diagnosis. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 52(4), 2302-2312.

Edited by: Hailong Li

Special issue on: Deep Learning in Healthcare

Received: Dec 27, 2023

Accepted: Jan 14, 2024



THE INTEGRATION AND INNOVATION OF SPORTS SOCIAL PLATFORMS AND INFORMATION TECHNOLOGY

YONGJUN CHEN*

Abstract. In order to better achieve the integration and innovation of sports social platforms and information technology, the author proposes an SFD (Sport Friend Discover) sports friend recommendation model based on physical testing big data. The core idea of this model is to use physical measurement data to match the similarity between athletes and recommend suitable exercise partners. Specifically, we collected a large amount of physical measurement data, including height, weight, body fat percentage, muscle mass, etc. Then, through data mining algorithms, these data are transformed into feature vectors of the movers. Next, we use a similarity algorithm to calculate the similarity between different athletes and find the most matching motion partner with the user. The results show that the SFD method outperforms the other two traditional recommendation methods on the dataset, with P @ 10, P @ 20, P @ 30, and P @ 40 of SFD reaching 0.099, 0.095, 0.085, and 0.591, respectively. SFD not only utilizes more neighboring information than FOAF based on local graph structure, but also compared to TRW based on global graph structure method, at the same time, the importance of the node itself is also considered, resulting in higher accuracy. It has been proven that the SFD sports friend recommendation model based on physical testing big data has achieved good results in recommending sports partners. Users can quickly find sports partners with similar body types and health conditions, improving the fun and effectiveness of exercise.

Key words: Recommendation algorithm, Sports and social interaction, Integrated innovation

1. Introduction. With the continuous development and popularization of information technology, sports social platforms are playing an increasingly important role in today's society. Sports social platform refers to a platform that combines sports and socializing through the internet and mobile applications. They provide a convenient way for people to share their sports experiences, challenges, and achievements, and interact and communicate with other sports enthusiasts. In the past few years, the number of users on sports social platforms has grown rapidly, attracting more and more people to join [1]. These platforms provide a virtual community where people can find like-minded partners and share their sports experiences and insights. By posting their own exercise records and achievements, users can receive praise and encouragement from others, which is crucial for improving their motivation to exercise and persevere [2]. In addition, sports social platforms also provide many useful features, such as sports data analysis, training plan development, and health advice, to help users better manage and improve their exercise status.

However, there are currently some issues and limitations with sports social platforms in the market. Firstly, for users, existing platforms often lack personalized and customized functions, which cannot meet the needs of different users. Everyone has different sports hobbies and goals. Some people like running, some like cycling, and some like exercising. Existing platforms often only provide some basic functions and cannot meet the personalized needs of users. Users hope to customize their exercise plans and training content based on their interests and goals [3]. Therefore, future sports social platforms need to strengthen the development of personalized and customized functions to meet the diverse needs of users. Secondly, for platform operators, there are difficult to solve privacy protection and data security issues. Sports social platforms involve users' personal information and exercise data, which are very important to users. However, due to the lack of effective privacy protection mechanisms, users' personal information and exercise data may be abused or leaked. In addition, data security is also an important issue.

Sports social platforms store a large amount of user data, which may be exploited by hackers or criminals without appropriate security measures. For platform operators, protecting user privacy and data security is an

*School of Physical Education, Chaohu University, Hefei, 238024, Anhui, China (chenyongjun@chu.edu.cn)

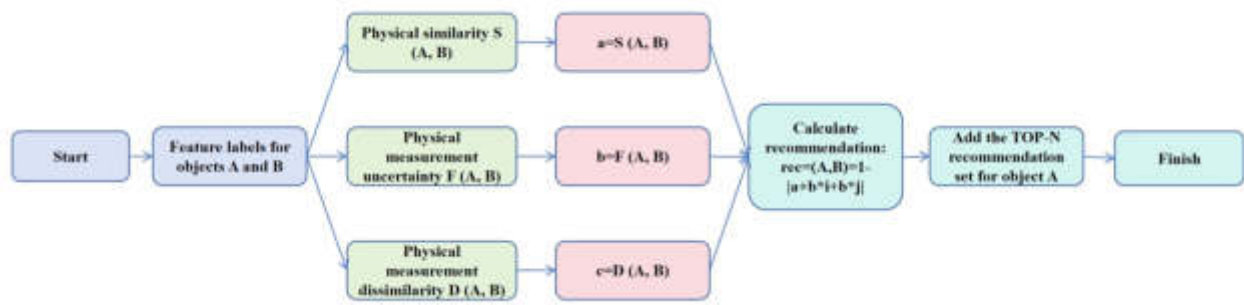


Fig. 2.1: Algorithm flow of college student physical examination

important responsibility, and they need to strengthen technical and management measures to ensure that user information is fully protected.

In addition, due to technological limitations and operational strategies, the functionality and experience of sports social platforms still need further improvement and innovation. At present, sports social platforms mainly focus on user interaction and sharing, but often overlook the needs of users for professional knowledge and guidance. Many users hope to receive professional sports advice and guidance to help them better engage in sports training. Future sports social platforms can strengthen cooperation with professional sports coaches and health experts, providing users with more comprehensive and professional services. In addition, sports social platforms can also combine virtual reality and augmented reality technology to provide a richer and more immersive sports experience [4].

In summary, sports social platforms play an important role in today's society, providing people with a convenient way to share sports experiences, challenges, and achievements, and interact and communicate with other sports enthusiasts. However, there are currently some problems and limitations with sports social platforms in the market, such as a lack of personalized and customized functions, privacy protection and data security issues, as well as limitations in functionality and experience [5].

Future sports social platforms need to strengthen the development of personalized and customized functions, as well as measures to protect privacy and data security, simultaneously enhancing functionality and experience to meet the diverse needs of users. Only in this way can sports social platforms better play their role in promoting health and social interaction.

2. SFD Sports Friend Recommendation Algorithm.

2.1. Overall Algorithm Design. The college student physical testing recommendation algorithm is specifically designed based on the big data of college student physical testing, and the algorithm process is shown in Figure 2.1. The concept of set pair analysis was introduced in this study to transform the traditional similarity. In the recommendation process, the first step is to calculate the similarity, uncertainty, and dissimilarity between two objects. Then, based on the theory of set pair analysis, the set pair recommendation degree $rec(A, B)$ needs to be calculated. Finally, the set pair recommendation degree is used to select suitable recommended objects and make recommendations [6].

2.2. Data preprocessing. Before designing the recommendation algorithm, in order to calculate the physical fitness recommendation between two objects, it is necessary to rate the physical fitness test items of college students according to the National Physical Fitness Standards, standardize the data of student physical test items into percentages, and classify the physical condition of students: The scores for explosive, endurance, flexibility, and strength categories are calculated based on the scores of student physical testing items with different weights [7].

Then, through threshold grading, different groups of strong, medium, and weak students are determined. Finally, it is necessary to fit the textual sports characteristics and give definitions: {"Strong": 1, "Medium": 2, "Weak": 3}.

Assuming the existence of objects A and B, their motion features are represented as feature vectors, that is $A = \langle a_1, a_2, a_3, a_4 \rangle, B = \langle b_1, b_2, b_3, b_4 \rangle$, the difference between the features of A and B can be represented by $|a_k - b_k|$, where $k \in \{1, 2, 3, 4\}$,

$|a_k - b_k| = 0$, indicating that A and B have similarity

$|a_k - b_k| = 1$, indicating that A and B have uncertainty

$|a_k - b_k| = 2$, indicating that A and B have a degree of dissimilarity.

This study will design the similarity, dissimilarity, and uncertainty of physical measurements in the context described above[8].

2.3. Set pair recommendation . Unlike traditional similarity calculation recommendation algorithms, it is necessary to consider the similarity between users, the differences between users, and the uncertain factors between users. Therefore, the concept of set pairs is introduced in the similarity calculation of recommendation algorithms, and the following definitions exist:

$$rel(A, B) = a + b \times i + c \times j \quad (2.1)$$

Among them, $rel(A, B)$ represents the correlation between A and B, $rel \in [-1, 1]$, the larger the rel , the higher the similarity, and vice versa, the lower the dissimilarity. a represents the physical similarity between A and B, that is $a = S(A, B)$; b represents the measurement uncertainty between A and B, that is $b = D(A, B)$; c represents the physical measurement dissimilarity between A and B, that is $c = F(A, B)$, and satisfies $a + b + c = 1$. i is the uncertainty marker and j is the dissimilarity marker. During the operation, i and j are both coefficients involved in the operation, and a constant value of -1 is specified for j , the value of i in the range of $[-1, 1]$ depends on the situation[9]. Set pair recommendation is a transformation based on the correlation degree rel , which comprehensively considers the factors of similarity, difference, and uncertainty to avoid errors caused by high similarity or difference, and is defined as follows:

$$rec(A, B) = 1 - |a + b \times i + c \times j| \quad (2.2)$$

$rec(A, B)$ represents the set pair recommendation degree between A and B, $rec \in [0, 1]$, the closer rec approaches 0, the less likely it is to be recommended; The closer it approaches 1, the easier it is to be recommended[10].

(1) *Physical similarity.* Similarity is a numerical measure of the degree of similarity between two objects, and is an important reference indicator in personalized recommendation systems. Traditional similarity is calculated based on user ratings of items and recommendations using relevant formulas. The calculation method of physical similarity in this study is different from traditional methods[11]. It refers to obtaining student physical measurement data, analyzing and processing the data to extract and describe the movement characteristics of students, calculate similarity by comparing the eigenvalues of different students through certain methods. If there are objects A and B in the recommendation system, the recommended objects A and B have the following similarity:

$$S(A, B) = \frac{N(a_k = b_k)}{n} \quad (2.3)$$

Among them, $S(A, B)$ represents the physical similarity between objects A and B, n is the total number of feature attribute types, and a_k and b_k represent the feature values of the two objects, respectively. By using certain rules, students' grades are divided into different levels. When two students have equal levels of physical education grades, it is considered that they have a certain degree of similarity, that is, $N(a_k = b_k)$ is the number of features that two objects have the same characteristic value[12].

(2) *Physical measurement dissimilarity.* Dissimilarity is a numerical measure that describes the degree of difference between two objects. In the recommendation problem based on physical fitness test scores, it is often unreasonable to only recommend based on similarity, if the two recommended parties only have similarity, it means that there is very little knowledge that both parties can learn from each other, so there needs to be a certain degree of difference between them. Only when there are two different parties can there be the possibility of learning from each other[13]. If there are two objects A and B with dissimilarity in the recommendation

system, then the recommended objects A and B have the following dissimilarity:

$$D(A, B) = \frac{\sum_{k=1}^n ||a_k - b_k| - 1| - N(a_k = b_k)}{n} \tag{2.4}$$

In the formula, D (A, B) represents the degree of dissimilarity between objects A and B, n is the total number of feature attribute types, and ak and bk represent the feature values of the two objects, respectively. It $\sum_{k=1}^n ||a_k - b_k| - 1| - N(a_k = b_k)$ is the number of distinct features that two objects have, among them, $\sum_{k=1}^n ||a_k - b_k| - 1|$ is the sum of the number of similar and different features of two objects, $N(a_k = b_k)$ is the sum of similar features of two objects, and n is the type of feature attribute[14].

(3) *Physical measurement uncertainty.* If there are objects A and B in the recommendation system, when one party's sports performance is average and the other party's performance is strong or weak, it cannot be used to determine whether the gap between the two parties is really large enough to teach the other party, thus there is uncertainty. In set pair theory, the sum of similarity, dissimilarity, and uncertainty is 1, that is, a+b+c=1. Therefore, there are the following uncertainties for recommended objects A and B:

$$F(A, B) = 1 - \frac{\sum_{k=1}^n ||a_k - b_k| - 1|}{n} \tag{2.5}$$

(4) *Determination of connectivity i.* The value of the difference uncertainty coefficient i corresponding to the set pair recommendation degree of objects A and B in the recommendation system is a key point that needs to be determined. As the value of i approaches 1, the similarity between the two objects increases. Therefore, using cosine similarity, a method for determining the value of i using computational value method is proposed, which has the following definitions:

$$i = \frac{1}{1 + d/s} \tag{2.6}$$

where s is the cosine similarity of objects A and B, the formula is as follows:

$$S = \frac{\sum_{k=1}^n (a_k \times b_k)}{\sqrt{a_k} \sqrt{b_k}} \tag{2.7}$$

d is the cosine dissimilarity of objects A and B. Under certain conditions, cosine dissimilarity can be transformed from cosine similarity. Use the following formula to transform similarity:

$$d = e^{-s} \tag{2.8}$$

d represents the degree of dissimilarity between two objects, s represents the degree of similarity between two objects. The larger the value of s, the smaller the value of d, and the closer i approaches 1; On the contrary, the further i moves away from 1 [15].

(5) *Calculation of Top-N Recommendation Set and Friend Recommendation.* The Top-N recommendation algorithm sorts data according to certain rules and selects the largest or smallest N data from the sorting list for recommendation. Different rules can be formulated for different social environments to filter the data in the recommendation set. Based on the study of university student groups, multiple factors need to be considered when making friend recommendations. Based on the obtained user set for the recommendation set, the average recommendation degree between the user and the recommended user is calculated as the threshold r. Recommendations greater than the threshold are stored in the user's Top-N recommendation set, as shown in Figure 2.2.

Due to the fact that in the process of friend recommendation, not only do we need to consider the recommendation level between users, but there are also some practical issues that need to be considered, such as the user's class, gender, and distance between living areas. When recommending, we filter based on the user's needs [16].



Fig. 2.2: Determination of Top-N recommendation set and friend recommendation

Table 3.1: Partial Physical Examination Data for Male Students

number	height /cm	weight /kg	vital capacity /ml	long jump /cm	sit-and-stand reach /cm	50 m/s	1 km/s	Pull up /piece
Male 1	173.5	55.2	3235	252	8.6	7.3	213	7
Male 2	177.3	74.4	3902	210	19	8.7	285	0
Male 3	174.4	68	4464	250	10.1	6.8	201	7
Male 4	174.2	64.2	4111	224	3.7	8.1	360	15
Male 5	175.5	62.7	3176	240	13.2	7.8	255	17

Table 3.2: Partial Physical Examination Data for Female Students

number	height /cm	weight /kg	vital capacity /ml	long jump /cm	sit-and-stand reach /cm	50 m/s	800 m/s	Sit ups /piece
Female 1	159.4	69.7	3012	154	18.5	10.2	270	33
Female 2	156.8	45.8	2462	176	14.8	9.5	246	26
Female 3	151.6	41.7	2063	160	19.9	10.2	300	32
Female 4	157.5	50.7	3035	194	26.8	8.6	227	44
Female 5	152.4	45.6	2323	186	23	8.4	237	45

3. Experimental results. The selected data is taken from the physical examination results of college students in a certain university, which are true and reliable, and only the part of the physical examination results is retained, the identification information such as name and student ID have been deleted. The student physical examination results are shown in Tables 3.1 and 3.2.

According to the National Physical Fitness Standards, calculate the physical test scores of students in Tables 3.1 and 3.2, as shown in Tables 3.3 and 3.4.

Based on the actual situation, provide the weight coefficients of the physical testing project for the features, and calculate the four major feature scores of students according to the weights, as shown in Figure 3.1.

In order to evaluate the effectiveness of the newly proposed SFD friend recommendation algorithm, the author compared SFDH with some typical local and global methods: FOAF: If two vertices have more com-

Table 3.3: Partial Physical Examination Results for Male Students/score

number	BMI	vital capacity	long jump	sit-and-reach	50m	1km	pull-up
Male 1	18.34	60.58	81.25	66.29	77	84.38	20
Male 2	23.67	71.7	60	82.35	63	52.5	0
Male 3	22.36	81.28	80	68.43	90	94	20
Male 4	21.16	75.18	67	55.69	15	76	
Male 5	20.36	58.59	75	72.86	72	66	85

Table 3.4: Female Partial Physical Examination Results/score

number	BMI	vital capacity	long jump	sit-and-reach	50m	800m	Sit ups
Female 1	27.43	79.24	64	78.46	60	60.8	66
Female 2	18.63	68.24	108	72.77	67	70.4	55
Female 3	18.14	60.26	76	81.33	60	32	65
Female 4	20.44	79.7	88.57	100	76	78	77
Female 5	19.63	65.46	82.86	91.5	78	74	78

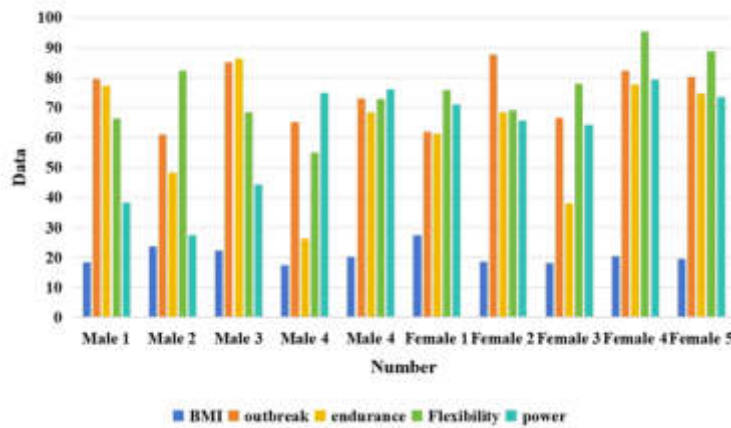


Fig. 3.1: Male and female partial feature scores/score

mon friends, they are more likely to become friends. Global Walkthrough Algorithm (TRW): Preserves all path structures in the network, investigates all structural information, and calculates node similarity. The performance of the three algorithms on the mathematical data of ScienceNet is shown in Figure 3.2.

As shown in Figure 3.2, the SFD method outperforms the other two traditional recommendation methods on the dataset, with P @ 10, P @ 20, P @ 30, and P @ 40 of SFD reaching 0.099, 0.095, 0.085, and 0.591, respectively [17,18,19].

SFD not only utilizes more neighboring information than FOAF based on local graph structure, but also considers the importance of nodes themselves, resulting in higher accuracy compared to TRW based on global graph structure method.

4. Conclusion. The author proposes a friend recommendation model based on set pair theory, which innovates extensively in the formulas of similarity, dissimilarity, and uncertainty. The cosine similarity calculation method and its transformation are used to determine the value of connectivity i , ultimately obtaining the rec-

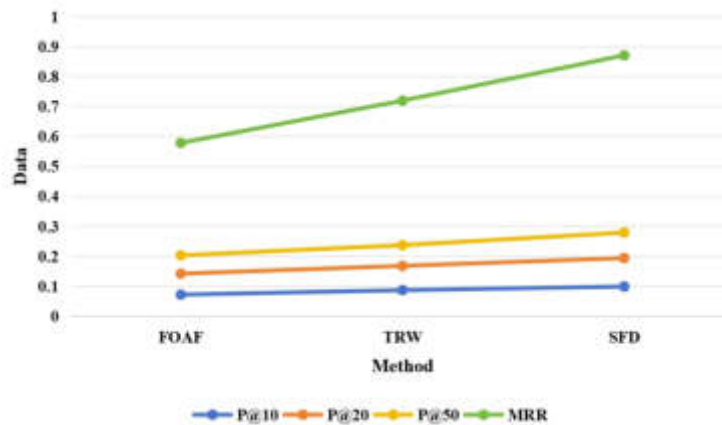


Fig. 3.2: Performance comparison of methods on the Science Network dataset

ommendation degree between users. The experimental results demonstrate that the designed recommendation method is more targeted and complementary, and the quality of recommended users is improved. Due to the relatively small number of feature attribute types and feature grading levels, the degree of difference between users is not significant, which has a better effect on users with larger differences. Therefore, in future research, more feature attributes will be added, such as user gender, user grade, and other personal information, and the formula will be further improved to further improve recommendation accuracy and rationality.

REFERENCES

- [1] Papalexandris, A., Mammassis, C. S., & Kostopoulos, K. (2021). Absorptive capacity and innovation outcomes in teams: the role of ceo servant leadership and tmt information exchange frequency on team social integration, task conflict and knowledge breadth. *SSRN Electronic Journal*, 1651(2), 012005.
- [2] Liu, Y., & Wang, Y. (2021). Case study on the deep integration of information technology and english curriculum. *Proceedings of the 2020 International Conference on Modern Education Management, Innovation and Entrepreneurship and Social Science (MEMIESS 2020)*, 16(3), 17-19.
- [3] Chen, H., & Greitens, S. C. (2021). Information capacity and social order: the local politics of information integration in china. *Governance*, 36(1), 47.
- [4] Sweis, R., Abed, S., Zu', N. A., Alzu', B. M. F., Bi, N. A., & Suifan, T., et al. (2022). The relation between information technology adoption and the pharmacists' job satisfaction in the chain community pharmacy in amman. *International Journal of Business Innovation and Research*, 27(3), 297.
- [5] Lei, Y., Guo, Y., Zhang, Y., & Cheung, W. (2021). Information technology and service diversification: a cross-level study in different innovation environments. *Information & Management*, 58(4), 103432.
- [6] Balanchuk, I. S., & Mykhalchenkova, O. Y. (2021). Technological platforms in the field of innovation trends in urope and ukraine. *Science Technologies Innovation*, 2(18), 14-24.
- [7] Tabares, S. (2023). Corporate social responsibility or corporate social innovation? two approaches towards the labour integration of disabled employees in colombia. *Social Responsibility Journal*, 19(4), 626-640.
- [8] Effendi, M. I., Widjanarko, H., & Sugandini, D. (2021). Green supply chain integration and technology innovation performance in smes: a case study in indonesia. *Korea Distribution Science Association*, 6(6), 38.
- [9] Li, J. (2021). Research on the reform and innovation of preschool education informatization under the background of wireless communication and virtual reality. *Wireless Communications and Mobile Computing*, 56(2), 17.
- [10] Timothy, W., Shannon, F., & Alanna, K. (2021). Technologies and the effects on social engagement in long-term care facilities during covid-19: a scoping review. *Innovation in Aging (Supplement_1)*, Supplement_1, 64(7), 101-103.
- [11] Huang, M., Jiang, Q., Qu, Q., Chen, L., & Chen, H. (2022). Information fusion oriented heterogeneous social network for friend recommendation via community detection. *Applied Soft Computing*, 9(2), 33.
- [12] Behl, R., & Kashyap, I. (2021). A unified probabilistic factor model with social regularization for point of interest recommendation. *Materials Today: Proceedings*, 21(6), 979-984.
- [13] Flugum, R. (2021). The trend is an analyst's friend: analyst recommendations and market technicals. *The Financial Review*, 17(5), 69-72.
- [14] Chang, L., Chen, W., Huang, J., Bin, C., & Wang, W. (2021). Exploiting multi-attention network with contextual influence

- for point-of-interest recommendation. *Applied Intelligence: The International Journal of Artificial Intelligence, Neural Networks, and Complex Problem-Solving Technologies*,74(4), 51.
- [15] Wang, S., Zhang, L., Yu, M., Wang, Y., Ma, Z.,& Zhao, Y. (2021). Attribute-aware multi-task recommendation. *Journal of supercomputing*,47(5), 77.
- [16] Papneja, S., Sharma, K.,& Khilwani, N. (2021). Movie recommendation to friends using whale optimization algorithm. *Recent advances in computer science and communications*,96(5), 14.
- [17] Alguacil, M.,J.Núez-Pomar, Calabuig, F., Escamilla-Fajardo, P.,& Staskeviciute-Butiene, I. (2021). Creation of a brand model through sem to predict users' loyalty and recommendations regarding a public sports service. *Heliyon*, 7(6), e07163.
- [18] Liao, S. H., Widowati, R.,& Yang, K. C. (2021). Investigating sports behaviors and market in taiwan for sports leisure and entertainment marketing online recommendations. *Entertainment Computing*, 39(4), 100442.
- [19] Mayorga-Vega, D., Casado-Robles, C., Lopez-Fernandez, I.,& Viciano, J. (2022). Activity wristband-based physical activity recommendations in young people. *Science & sports*,65(7),68-74.

Edited by: Hailong Li

Special issue on: Deep Learning in Healthcare

Received: Dec 27, 2023

Accepted: Jan 14, 2024



THE APPLICATION OF INFORMATION TECHNOLOGY FOR ATHLETE DATA ANALYSIS AND AUTOMATIC GENERATION OF TRAINING PLANS

SHULI YUAN*

Abstract. In response to the demand for scientific training of sports athletes, the author combined data mining technology to study an improved sports training mode decision support evaluation system. In this regard, the author analyzed the characteristics of association rule algorithms and elaborated on their functions in data preprocessing, data mining, and pattern evaluation. Based on the software design of decision support systems, the characteristics of system operation were analyzed. At the same time, the author focused on explaining the data fusion processing of association rules in sports evaluation decision support systems, and proposed an improved Apriori algorithm output mode to improve the effectiveness of system evaluation. Compared with other algorithms such as Apriori, DC Apriori and Apriori, this algorithm has higher reliability. When the minimum confidence is increased, the advantage of prior information will gradually disappear, and the final result will be obtained. Experimental results show that this method can effectively provide support for sports training decision-making.

Key words: Data mining technology, Association rules, Sports training evaluation, Data fusion processing

1. Introduction. In modern sports competition, the analysis of athlete data and the formulation of training plans are crucial for improving competitive level and achieving excellent results. With the development of technology and continuous innovation in data collection technology, more and more athletes and coaches are using information technology to collect, analyze, and apply sports data. The training plan for athletes has gradually shifted from subjective experience in the past to objective decision-making based on data, making training more scientific and efficient. The application of data analysis in sports competitions has become a trend. By analyzing athlete data, coaches can understand their performance during training and competition, identify their strengths and weaknesses, and then develop targeted training plans [11]. For example, in basketball games, coaches can evaluate a player's performance in offense and defense by analyzing their shooting percentage, rebounds, and assist data, and provide targeted technical and tactical training. Through data analysis, coaches can discover the potential abilities of athletes and assist them in personalized training to improve their competitive level.

Data analysis can not only help coaches develop training plans, but also help athletes understand their performance and improve their space. By analyzing their own sports data, athletes can gain a deeper understanding of their strengths and weaknesses, identify their problems in the competition, and find ways to improve. For example, in track and field competitions, athletes can analyze their speed, endurance, and technical data to identify their weaknesses in training and conduct targeted training to improve their competitive level. Data analysis can enable athletes to have a more comprehensive understanding of their performance and potential, thereby formulating more scientific and effective training plans. In addition to data analysis, the application of information technology in sports training also includes data collection and application [4]. With the continuous advancement of technology, athletes can use various sensors and devices to collect exercise data, such as heart rate, step frequency, exercise trajectory, etc. These data can help athletes and coaches have a more comprehensive understanding of their sports status and performance, enabling more precise training and adjustments. For example, in football training, athletes can monitor their heart rate and movement track by wearing smart bracelets or chest bands, so as to adjust their intensity and rhythm in training. Through data collection and application, athletes and coaches can more scientifically manage training and competitive processes, and improve training effectiveness [1]. The application of data analysis and information technology has not only changed the training methods of athletes, but also put forward new requirements for the role of coaches. Traditionally,

*Zhengzhou University of Technology, Zhengzhou, 450044, China (mylifeyuan@163.com).



Fig. 2.1: Data mining process.

coaches relied mainly on their own experience and intuition to develop training plans, but now they need to have certain abilities in data analysis and information technology application. Coaches need to learn to collect, process, and analyze sports data, obtain valuable information from it, and apply it to training programs. This poses new requirements for the comprehensive quality of coaches, who need to constantly learn and update their knowledge to adapt to the needs of technological development and data analysis.

In summary, the application of data analysis and information technology in modern sports competitions has become a trend. Through data analysis, coaches can develop targeted training programs to help athletes improve their competitive skills. Athletes can also analyze their sports data to identify their problems and engage in targeted training. Data collection and application can help athletes and coaches have a more comprehensive understanding of sports status and performance, and improve the scientificity and effectiveness of training. The use of data analysis and information technology not only changes the teaching process, but also places new demands on the effectiveness of teachers. With the continuous development of techniques and technology, the use of statistical data and information technology in sports competitions is becoming more and more popular, challenging many methods and approaches to athletes and coaches.

The author aims to explore the application of information technology in athlete data analysis and automatic generation of training plans, and evaluate its impact on athlete training and competitive performance. By collecting and analyzing physiological, technical, and competitive data of athletes, combined with advanced data mining and machine learning algorithms, we hope to automatically generate personalized training plans to help athletes better tap into their potential, improve training effectiveness, and achieve better results in competitions. Through this study, we will be able to gain a deeper understanding of the current status and potential of the application of information technology in athlete training and competition, providing scientific training guidance and decision support for athletes and coaches [15].

2. Decision support system for sports training mode.

2.1. Overview of Association Rule Mining Algorithms. The most crucial aspect of data mining technology is the association rule algorithm, and the Apriori algorithm is a classic algorithm in association rule algorithms. At present, there are various ways to classify association rules, and the most common one is to classify them according to the dimensions of the data types in the association rules [7]. It can be classified into one-dimension and multiple-attribute. There are a lot of influential factors in practice for athletes' physical training, and the data types obtained are much bigger than that of 3D ones. Therefore, the multi-dimension association rules must be taken into account in the design of the DSS of sports training model. Multi-dimensional association rules are more complicated than single-dimension association rules. Usually, multidimensional association rules include data preprocessing, data mining, and model assessment. The concrete data mining process is illustrated in Figure 2.1.

The preprocessing stage in the data mining process mainly involves collecting, processing, and transforming data, which takes the longest time throughout the entire data mining process; The data mining stage mainly analyzes the data in the preprocessing stage through selected association rules, neural network techniques, etc [12]. The evaluation stage of the pattern mainly involves presenting the information obtained from data mining to users, or providing a visual program for real-time viewing and analysis.

2.2. Data Fusion Processing of Sports Training Evaluation Decision Support System. In the evaluation of sports training model based on large data mining, the related data should be integrated into the

DSS. This paper focuses on the classification of the extracted data by neural networks.

2.2.1. Fusion and Clustering of Sports Evaluation Decision Information. The data feature identification function of the sports evaluation decision support system is:

$$P_c = \sum_{i=0}^n \sum_{j=0}^n \alpha(i, j) P(i, j) \tag{2.1}$$

Assuming that the starting symbol for each of the above attributes is: $C_0 = C_{N/2} = 0, C_{N-n} = C_n^*, n = 0, 1, 2, \dots, N/2 - 1$, the model relationship between sports training evaluation decision data and cluster center distribution is:

$$P_r = \frac{P_t}{(4\pi)^2 \left(\frac{d}{\lambda}\right) r} \left[1 + \alpha^2 + 2\varepsilon \cos\left(\frac{4\pi h^2}{d\lambda}\right) \right] \tag{2.2}$$

Based on the association criteria of sports training modes, feature recognition of sports decision support systems is carried out based on the different types of data obtained. Among them, the attribute categories of association criteria in the system are:

$$R_\beta X = U\{E \in R \mid c(E, X) \leq \beta\} \tag{2.3}$$

$$R_\beta X = U\{E \in R \mid c(E, X) \leq 1 - \beta\} \tag{2.4}$$

For different data block types m_i and m_j combined with the association criteria of sports training modes, the iterative process of sports training decisions obtained by using the fuzzy mean method in the data types is as follows:

$$S_b = \sum_{i=1}^e p(\omega_i) (u_i - u) (u_i - u)^T \tag{2.5}$$

$$S_\omega = \sum_{i=1}^e p(\omega_i) E \left[\frac{(u_i - u) (u_i - u)^T}{\omega_i} \right] \tag{2.6}$$

$$S_i = S_b + S_\omega \tag{2.7}$$

In the formula, $p(\omega_i)$ is the set of association rule vectors for the sports training decision system. Based on the above calculation formula, the fusion of information in the sports training mode decision support system is achieved [14].

2.2.2. Improve Apriori algorithm output. It is necessary to create a policy organization that presents the information only during the request for information that will be used to make design decisions for sports training models. The author created a collection of sports training standards organization standards using a modified weight system, and the last important rule is the only product that accepts weight products.

$$\omega_{sij} (n_0 + 1) = \omega_{sij} (n_0) - \eta_{sij} \frac{\partial J}{\partial \omega_{sij}} \tag{2.8}$$

Through the similarity analysis of relevant data in the DSS, the adaptive learning process can be obtained as follows:

$$\alpha_{desira}^i = \alpha_1 \cdot \frac{Density_i}{\sum_i Density_i} + \alpha_2 \frac{AP_i}{AP_{init}} \tag{2.9}$$

Rearrange the relevant data stored in the system in quintuples to obtain the probability density function used for data mining:

$$P_s = P_{2D}^k (1 - P_{2D})^{N-1-k} \sum_{i=1}^{\infty} \lambda_s^i = \frac{\lambda_s}{1 - \lambda_s} \tag{2.10}$$

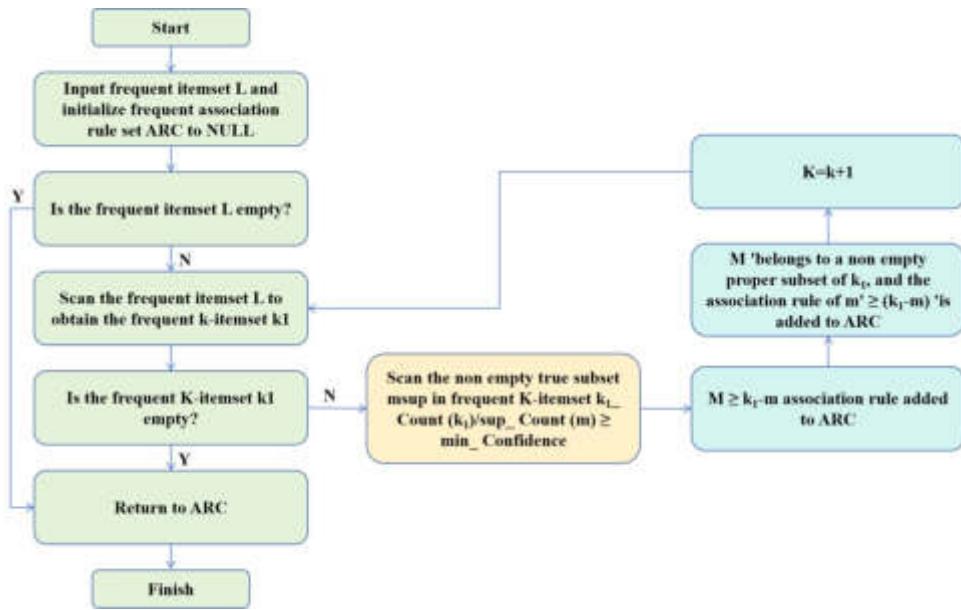


Fig. 2.2: Improved Apriori algorithm for frequent association rule process.

In the formula, λ_s and P_{2D} represent the correlation dimension of the data and the probability that the data can be effectively detected. The motion training model determines that the data clusters in the support system differ in the following ways:

$$\text{DisSim}(A, B) = 1 - \left| \frac{\text{SameDis}(A) - \text{SameDis}(B)}{\text{Dis}(A) + \text{Dis}(B)} \right| \quad (2.11)$$

In summary, the author can establish a system database model by mining the data of the sports training mode decision support system and extracting its association rules, and design the corresponding system in combination with software development [5].

2.3. Software Design. The decision support system for sports training mode is based on modern computers and uses computer and programming languages to simulate the sports training effects of athletes through human-computer interaction. It mainly addresses decision-making issues for managers during the implementation of plans. Decision systems can provide athletes with the convenience of obtaining real-time sports data, while also assisting in guiding and tracking the effectiveness of sports.

The author chooses the improved Apriori algorithm with frequent association rules as shown in Figure 2.2 to design the human-computer interaction system. Its main functions include:

1. The human-computer interaction system can provide a more convenient computer environment for decision-makers in sports training. It can check the physical fitness indicators of each athlete based on the front-end display settings, and use computers for tracking and processing.
2. Visually display the operational status of the decision support system, allowing users to fully understand the data changes during the system's operation and make timely adjustments.
3. Based on the output results of the system, targeted adjustments can be made to the training plan, and the simulation can be calculated to form the optimal training plan.
4. The human-computer interaction system can also correct erroneous information and perform preliminary verification and judgment on input data [6].

On the basis of analyzing the decision support system algorithm, the author integrated data mining technology with the system program through algorithm compilation to design a sports training mode decision support system [2]. The main functional modules of the system include communication module, program module,

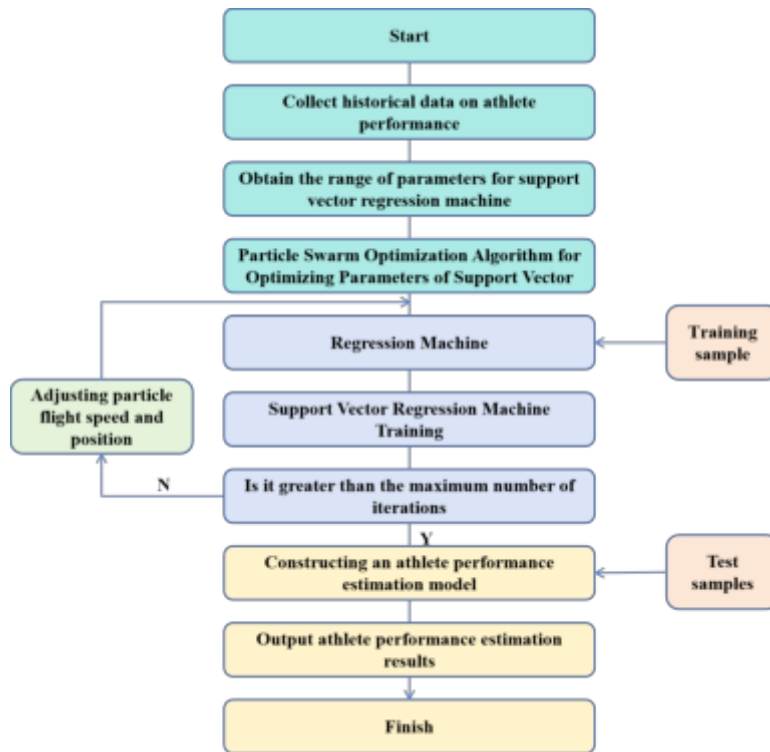


Fig. 2.3: Flow chart of athlete performance modeling and estimation.

database module, and data output module. Using Conti-ki bus technology to transmit and coordinate the data types of the sports training mode decision support system, meanwhile, combining VIX integrated control technology can achieve integrated control of the system. The data perception of decision support systems is built based on the 6LoWPAN protocol stack. The design of the wireless sensor network system adopts Atmel1284P as the main chip to control the overall IoT address allocation and mobilization of the sports training mode decision-making system. After the taskbar address is determined, the system’s human-computer interaction is achieved through the TaskBasic interface program.

2.4. Athlete Performance Modeling and Estimation. The workflow of athlete performance modeling and estimation based on big data analysis technology is shown in Figure 2.3.

The steps for modeling and estimating athlete performance using big data analysis technology are as follows:

1. Collect historical data of athlete performance, process the historical data of athletes to obtain the range of athlete performance:

$$x_i'' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \tag{2.12}$$

In the formula, the maximum value of the athlete’s score is x_{\max} , and the minimum value of the athlete’s score is x_{\min} .

2. On this basis, a support vector machine model based on PSO is proposed [10].
3. Train athletes according to each set of parameters and implement learning through support vector regression machine.
4. If the number of iterations exceeds the set maximum value, the algorithm ends; If the number of iterations is less than the set maximum value, adjust the flight speed and position of the particle swarm.

Table 3.1: Data related to the badminton team training in a certain place

Type	Tid	Item	Avg
Data	5401	114	38

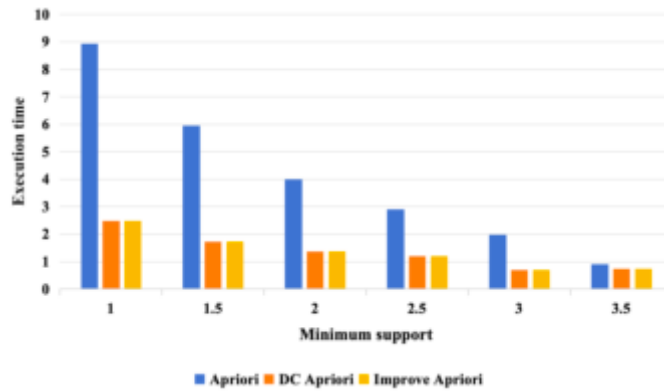


Fig. 3.1: Comparison of execution times at minimum support.

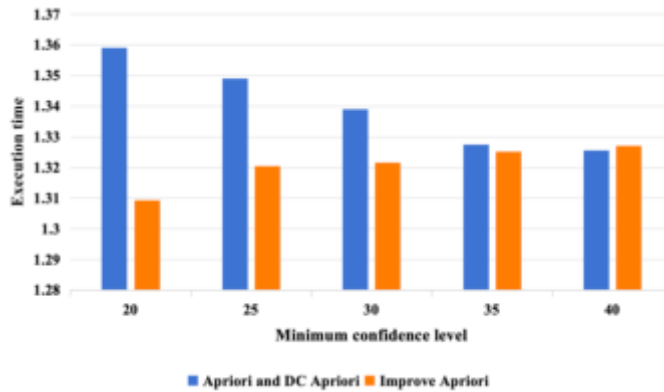


Fig. 3.2: Comparison of execution times at minimum confidence.

5. Increase the number of iterations of particle swarm optimization algorithm.
6. Retrain athlete performance through the optimal parameters of the support vector regression machine, and obtain the optimal parameters of the support vector regression machine through the optimal solutions pbest and gbest. Construct an athlete performance estimation model based on support vector regression machine.
7. Test and analyze the performance of the athlete performance estimation model through athlete performance test samples, and output the final athlete performance estimation results [3].

3. Simulation experiment analysis. On this basis, Apriori algorithm, DC Apriori algorithm and modified Apriori algorithm are compared, the results show that the algorithm is feasible. Most of the time during the experiment, programming was done in the Java language. In this paper, a badminton team in the relevant training data as the research object.

The “Tid”, “item”, and “quantities” in Table 3.1 represent the specific types of training items, the total

number of data items, and the average level per training.

In Figures 3.1 and 3.2, the change in system runtime is shown as minimum support and minimum confidence increase.

Figure 3.1 shows the reaction speed of the modified Apriori algorithm proposed in this article under the minimum support, which shows that it is a more efficient method [9, 13, 8]. As you can see in Figure 3.2, the improved Apriori algorithm has better performance when the minimum confidence is low. As the minimum confidence level increased, the advantages of Apriori faded and the same effect was achieved.

4. Conclusion. In order to improve the competitive level of athletes, we must constantly improve the competitive level. Based on information technology, this paper studies the problem of decision support in sports teaching mode. On this basis, a method of physical education based on network is proposed. Secondly, using Apriori, DC Apriori, Apriori and other classical algorithms to test and verify the Apriori algorithm, the Apriori algorithm can better support training decisions, with high practical value.

REFERENCES

- [1] N. E. BARCZAK-SCARBORO, E. KROSHUS, B. PEXA, J. K. R. MIHALIK, AND J. DEFREUSE, *Athlete resilience trajectories across competitive training: The influence of physical and psychological stress*, Journal of Clinical Sport Psychology, 1 (2022), pp. 1–19.
- [2] H. CARRERA, J. ROMERO, I. BERGANZA, J. DUNAVANT, G. A. ABASCAL-PONCIANO, J. D. STARKEY, C. STARKEY, AND S. CHO, *192 evaluation of textural characteristics (texture profile analysis and shear force) on commercial training pet treats*, Journal of Animal Science, 101 (2023), pp. 98–99.
- [3] Y. CHEN AND K. YE, *A wireless network based technical and tactical analysis of volleyball game based on data mining techniques*, Wireless Networks, 29 (2023), pp. 161–172.
- [4] J. G. CONN, J. W. CARTER, J. J. CONN, V. SUBRAMANIAN, A. BAXTER, O. ENGVIST, A. LLINAS, E. L. RATKOVA, S. D. PICKETT, J. L. McDONAGH, ET AL., *Blinded predictions and post hoc analysis of the second solubility challenge data: Exploring training data and feature set selection for machine and deep learning models*, Journal of Chemical Information and Modeling, 63 (2023), pp. 1099–1113.
- [5] ———, *Blinded predictions and post hoc analysis of the second solubility challenge data: Exploring training data and feature set selection for machine and deep learning models*, Journal of Chemical Information and Modeling, 63 (2023), pp. 1099–1113.
- [6] F. S. DE JESUS AND L. M. L. FAJARDO, *The benefits of training and development programs for lending organization personnel: Basis for development of training program*, IRA-International Journal of Management & Social Sciences (ISSN 2455-2267), 18 (2022), pp. 13–32.
- [7] B. DESSIRIER, K. M. SHARMA, J. PEDERSEN, C.-F. TSANG, AND A. NIEMI, *Channel network modeling of flow and transport in fractured rock at the äspö hrl: Data-worth analysis for model development, calibration and prediction*, Water resources research, (2023), p. e2022WR033816.
- [8] Y. GE ET AL., *Intelligent analysis and evaluation method of athletics running data based on big data statistical model*, Mathematical Problems in Engineering, 2022 (2022).
- [9] D. P. GOLDEN AND J. N. HERTEL, *Clinician impact on athlete recovery and readiness in a 24-hour training cycle*, International Journal of Athletic Therapy and Training, 1 (2022), pp. 1–6.
- [10] A. R. M. T. ISLAM, S. C. PAL, R. CHAKRABORTTY, A. M. IDRIS, R. SALAM, M. S. ISLAM, A. ZAHID, S. SHAHID, AND Z. B. ISMAIL, *A coupled novel framework for assessing vulnerability of water resources using hydrochemical analysis and data-driven models*, Journal of Cleaner Production, 336 (2022), p. 130407.
- [11] T. W. JONES, H. P. LINDBLOM, M. S. LAAKSONEN, AND K. MCGAWLEY, *Using multivariate data analysis to project performance in biathletes and cross-country skiers*, International Journal of Sports Physiology and Performance, 1 (2023), pp. 1–12.
- [12] L. P. KOZIRIS, *Ncaa student-athlete training during covid-19 stay-at-home restrictions*, Strength & Conditioning Journal, 44 (2022), pp. 128–130.
- [13] Z. LI, Y. ZHOU, C. ZHAO, Y. GUO, S. LYU, J. CHEN, W. WEN, AND Y. HUANG, *Design of a cargo-carrying analysis system for mountain orchard transporters based on rgb-d data*, Applied Sciences, 13 (2023), p. 6059.
- [14] Z. MEI, *3d image analysis of sports technical features and sports training methods based on artificial intelligence*, Journal of Testing and Evaluation, 51 (2023), pp. 189–200.
- [15] L. YIN, W. ZHENG, H. SHI, AND D. DING, *Ecosystem services assessment and sensitivity analysis based on ann model and spatial data: A case study in miaodao archipelago*, Ecological Indicators, 135 (2022), p. 108511.

Edited by: Hailong Li

Special issue on: Deep Learning in Healthcare

Received: Jan 4, 2024

Accepted: Jan 25, 2024



DEEP LEARNING MODEL CONSTRUCTION OF URBAN PLANNING IMAGE DATA PROCESSING AND HEALTH INTELLIGENCE SYSTEM

CAN XU*

Abstract. In order to study the deep learning model of urban planning image data processing and health intelligent system, based on existing remote sensing image change detection methods, the author introduces and proposes the use of deep belief networks in deep learning to classify high-resolution remote sensing images and analyze urban expansion change detection. Compared with traditional methods, deep learning has the highest overall accuracy and Kappa coefficient. Deep learning has the highest producer accuracy and relatively low misjudgment rate, making it the most suitable for studying the trend of urban built-up areas. By calculating the information entropy of the image to predict the number of hidden layer nodes, the time for deep learning is greatly reduced. Under the same experimental conditions, the training time for each image can be shortened by 12 525 seconds has improved classification efficiency and made a significant contribution to research on urban expansion applications. Finally, the improved deep belief network was applied to classify and detect changes in the three phase remote sensing images of Beijing, and the urban expansion trend and characteristics of Beijing were analyzed. Provide technical reference and inspiration for urban planning and land use protection.

Key words: Deep learning, Deep belief network, Remote sensing images, Urban planning, Image data processing

1. Introduction. In the evaluation process of urban master planning methods, how to use practical and feasible evaluation methods to accurately and efficiently evaluate a large number of planning schemes is a practical problem faced by every urban planner [1]. As one of the deep machine learning methods, deep learning technology can extract features within samples and transform them. It has the outstanding advantage of strong learning ability and is widely used in fields such as image classification, object recognition, and object evaluation.

The commonly used deep learning techniques include automatic encoding, sparse encoding, deep belief networks, etc. These deep learning techniques utilize layer by layer feature transformation to transform meta spatial features into another space, facilitating feature classification and evaluation. Urban planning is a comprehensive work that not only considers the construction of tangible entities such as urban space, but also considers multiple aspects such as economy, society, environment, culture, etc. Throughout history, the formulation of urban planning has relied heavily on qualitative analysis and empirical judgment. Planners often fail to provide objective explanations for phenomena, and the scientific nature of the planning discipline has long been questioned, one important reason for this is the non collectability or non quantifiability of data. Nowadays, the emergence of intelligent planning has brought immeasurable vitality and change to the development of urban planning discipline.

Intelligent planning and design, with the assistance of computers, is particularly effective in saving manpower and computational costs, and the calculation and analysis results can also ensure absolute objectivity.

Remote sensing change detection is a technology that quantitatively analyzes the characteristics and information of land surface changes based on remote sensing images obtained at different times in the same region. It has become an important direction in current remote sensing image processing and analysis, and is widely used in many fields such as land use, vegetation cover, urban planning, crop growth monitoring, and disaster assessment and prediction.

At present, the change detection methods for remote sensing images mainly include algebraic operations, image classification, feature description, and other methods. In the field of urban expansion, due to the need to determine how urban land is transformed from other land uses, the research in this direction adopts the

*School of Economics and Management, Hubei Polytechnic University, Hubei, 435000, China (xu1982can@163.com)

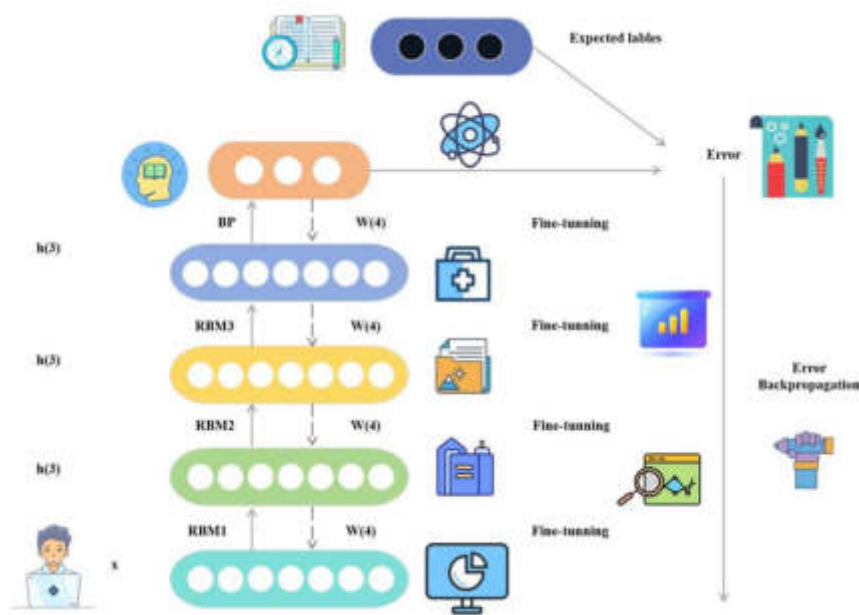


Fig. 2.1: DBN schematic diagram

method of image classification and change detection, which is more intuitive.

Image classification based methods can provide the types of surface changes and reduce the impact of external factors such as lighting and atmosphere on detection accuracy. However, in practical operations, a large number of learning samples need to be obtained, and the training time is relatively long. Therefore, how to obtain learning samples and reduce the training time of classification is an urgent problem that needs to be solved in the field of urban expansion.

The author takes urban expansion as an example and introduces a classification method of deep learning. After image classification, multi temporal change detection is performed. The accuracy index is compared with existing change detection methods, and information entropy is used to improve the efficiency of deep learning, achieving fast, efficient, and accurate change detection [2].

2. Methods. The deep learning method utilizes Deep Belief Network (DBN) for data classification and has made breakthrough progress. Subsequently, various research and engineering fields have adopted deep learning methods for application experiments [3]. In recent years, deep learning methods have also been continuously applied in the classification and recognition fields of videos, images, speech, etc. The essence of deep learning is a multi-level neural network, which improves the accuracy of results by extracting features from each layer to form final features suitable for classification.

Deep learning is applied in the field of remote sensing, utilizing deep belief network models for road target recognition in airborne images. But so far, there is still a lot of research space to apply this method to the classification of remote sensing images in large regions.

A deep belief network is composed of a multi-layer unsupervised Restricted Boltzmann Machine (RBM) network and a layer of supervised Backpropagation (BP) network, as shown in Figure 2.1 [4]. The experimental process of DBN includes two steps. Firstly, the input data is pre trained, and the output of the lower layer Boltzmann machine is used as the input of the higher layer, which is trained layer by layer.

In the fine-tuning stage, supervised learning is used to train the neural network layers, and the obtained errors are passed down to fine tune the weights of the deep belief network. The pre training stage actually initializes the weights of the neural network, thereby avoiding the drawbacks of local optima caused by random initialization.

Unlike traditional neural networks and shallow learning, deep learning methods generally have multiple layers, ranging from 4-7 layers to more than 10 layers. Moreover, through layer by layer feature extraction, they make classification and prediction results more accurate. However, the difficulty of using deep learning methods for classification lies in determining the depth of the network and the number of hidden layer nodes. Therefore, how to improve computational efficiency is an urgent problem that needs to be solved.

3. Experiments and Analysis.

3.1. Research Area and Data Sources . The author chose City A as the research area. The remote sensing image data used in the study includes LandsatTM and ETM+data with imaging times in 2009, 2015, and 2022, as well as auxiliary data such as A city topographic map, A city center administrative area map, and A city yearbook statistical data.

3.2. Land use classification standards and training sample selection. Appropriate classification standards and the number of training samples are the basis for accurate classification. Generally, a hierarchical classification system is adopted. The national standard GB/T21010-2007 "Classification of Land Use Status" stipulates that land use is divided into three categories: agricultural land, unused land, and construction land, each of which is further divided into several primary and secondary categories. Based on the research purpose and in combination with the provisions of national standards, the author categorizes land use consolidation into 5 categories [5]. Farmland, forests, and grasslands can also be merged into vegetation. A sufficient number of training samples and their representativeness are key to image classification. The selection method of training samples will affect the accuracy of classification, such as using pixel method, polygon method, etc. The mixed pixels in medium and low resolution remote sensing images contain complex information. Considering the complexity of land use in the study area, high-resolution remote sensing images should be selected for training sample selection.

The author used high spatial resolution images as training samples, randomly selected a training area with a sample size of 200, of which 100 samples were used for training the model and 100 samples were used for detecting model accuracy, each sample contains 10 pixels, accumulating 1000 pixels. For each class of samples, 213 are selected for unsupervised training, and the remaining 1/3 is used for fine-tuning in the network. As the author is studying the changes in the built-up areas of Beijing, it is advisable to select as many construction land as possible when selecting training samples to improve the classification accuracy of construction land.

3.3. Parameter Setting and Experimental Process . The restricted Boltzmann machine used in DBN only allows connections between hidden layer neurons and visible neurons, and there is no connection between two visible neurons or between two hidden layer neurons. In RBM, the energy equation is shown in Equation 3.1:

$$Energy(v, h) = h'Wv + b'v + c'h \tag{3.1}$$

In the formula, W represents the weight matrix of the neuron connections between the hidden layer and the visible layer, and b and c are the bias vectors on the visible and hidden neurons, respectively. When training Boltzmann machines, we input to the network through visible layer neurons, with the goal of updating and adjusting weights and biases, so that when training data is used as input, the configuration energy is minimized. Training RBM first inputs training vectors to the visible layer, and then compares and disperses them by alternately sampling hidden layer units and visible layer units. When using RBM, we do not need to calculate the joint probability and it is easy to sample. After only one Gibbs sampling iteration, we can reset (update) the weights and biases of RBM, as shown in Equation 3.2:

$$\begin{cases} W_{kj} = W_{kj} - \alpha(\langle v_{k0}h_{j0} \rangle - \langle v_{k1}h_{j1} \rangle) \\ b_k = b_k - (\langle v_{k0} \rangle - \langle v_{k1} \rangle) \\ c_j = c_j - (\langle h_{j0} \rangle - \langle h_{j1} \rangle) \end{cases} \tag{3.2}$$

In the formula, α is the learning rate, v_0 is sampled from the training sample, h_0 is sampled from $P(h|v_0)$, v_1 is sampled from $P(v|h_0)$, and h_1 is sampled from $P(h|v_1)$. We repeat this update operation on several

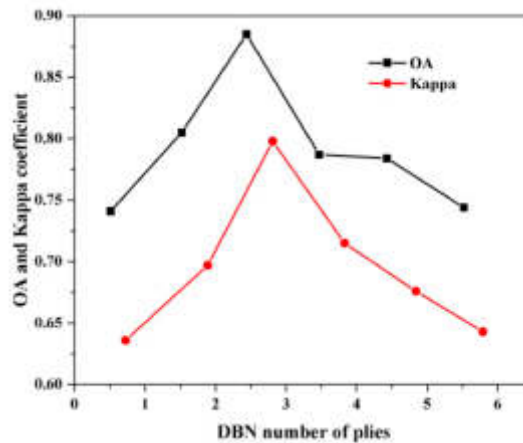


Fig. 3.1: Classification accuracy of different DBN layers

samples of the training data, and then iteratively train each next layer by using the activity of the hidden units in the previous layer as input data/visible units for the first layer.

For each pixel to be classified on an image, it is necessary to consider a region that includes its surrounding neighboring pixels. Assuming the neighborhood window size is $winsize$, it can be expanded into a one-dimensional vector with a $winsize \times winsize$ dimension. For DBN, the input data consists of three processed Pauli parameters, namely the diagonal elements of the correlation matrix ($0.5 | HH+VV12$, $0.5 | HH-VV12$, and $2 | HV12$), which can be assembled into a data vector for the first phase, therefore, for the data of period m , the dimension of the input vector is $winsize \times winsize \times 3 \times Em$. Calculate the spectral and texture feature vectors of the three bands synthesized by pseudocolor separately, and combine the three feature vectors into one feature vector as the input. The input dimension is 147. The experimental parameters are set as follows: The learning rate is initially set to 0.01, W is all random numbers from a normal distribution, and the hidden layer bias is initialized to 0. Due to the fact that the number of input and output nodes in the experiment is 147 and 5 (to be classified), the hidden layer nodes of the Boltzmann machine are set to take values of 5 to 147, respectively, when the error is minimized, it is the number of nodes in the first hidden layer, and so on for the remaining hidden layer nodes. The depth of the deep belief network is set to 1-6 layers, and the misclassification error, omission error, producer accuracy, user accuracy, overall accuracy, and Kappa coefficient are calculated to evaluate the classification accuracy. The results show that the highest accuracy is achieved when the network depth is 3, as shown in Figure 3.1 [6].

The experimental method is to preprocess remote sensing images; Based on the analysis of existing data and the establishment of interpretation criteria, the main land types in Beijing are extracted using computer classification methods; For smaller land classes, set an area threshold for neighboring merging; Extract boundaries of various land uses after merging, and correct incorrect boundaries through visual interpretation; Overlay the classification results of each image for later analysis and evaluation.

3.4. Classification Results and Analysis. The experiment focuses on the detection of changes after classification, using the ISODATA method in unsupervised classification, the maximum likelihood classifier in supervised classification, and deep learning methods to classify and analyze remote sensing images. The accuracy of the three methods is evaluated using measurement indicators such as classification accuracy, overall accuracy, and Kappa coefficient. The results are shown in Tables 3.1 to 3.4 [7,8].

The results showed that the overall accuracy and Kappa coefficient of deep learning were the highest, followed by the maximum likelihood classifier, and ISODATA was the worst. Deep learning has the highest producer accuracy and Kappa coefficient, with a relatively low misjudgment rate, and is most suitable for studying the trend of changes in built-up areas. The possible reason is that deep belief networks avoid ran-

Table 3.1: Precision Evaluation of ISODATA Classification Results

Category of features	Misclassification error	Omission error	Producers accuracy	User Accuracy
land used for building	56.09	15.87	84.15	46.93
Forest and grassland	21.82	71.54	28.48	78.2
Water bodies	5.13	0.00	100.00	94.89
Naked ground	86.86	76.2	23.82	13.16
cultivated land	67.02	15.93	84.09	32.99

Table 3.2: Maximum likelihood classifier result accuracy evaluation Table

Category of features	Misclassification error	Omission error	Producers accuracy	User Accuracy
land used for building	33.37	1.59	98.43	66.65
Forest and grassland	9.65	0	100.00	91.37
Water bodies	0.00	0.6	99.42	100.00
Naked ground	1.68	37.46	62.56	98.34
cultivated land	0.56	1.89	98.13	99.46

Table 3.3: Precision Evaluation of Deep Learning Classification Results

Category of features	Misclassification error	Omission error	Producers accuracy	User Accuracy
land used for building	11.99	9.72	90.3	88.03
Forest and grassland	25.2	12.29	87.73	74.82
Water bodies	21.63	25.01	75.01	78.39
Naked ground	18.8	0.00	100.00	81.22
cultivated land	1.41	21.65	78.37	98.61

Table 3.4: Overall accuracy and Kappa coefficient accuracy evaluation Table

classification method	Overall accuracy	kappa coefficient
ISODATA	54.3569	0.4499
Maximum likelihood classifier	87.5641	0.8934
Deep learning	93.4144	0.9141

domly assigning initial values to neural networks and better overcome the problem of local optima through pre training methods. Therefore, deep belief networks combine the advantages of unsupervised neural networks and supervised classification, which can improve classification accuracy and efficiency.

3.5. Determining the number of hidden layer nodes by calculating information entropy. Deep learning methods have high classification accuracy, but it is difficult to determine the number of hidden layer nodes and the number of hidden layer layers, which requires continuous attempts to determine [9,10]. While ensuring that the feature dimensions are maintained within a small range, it is also necessary to ensure that

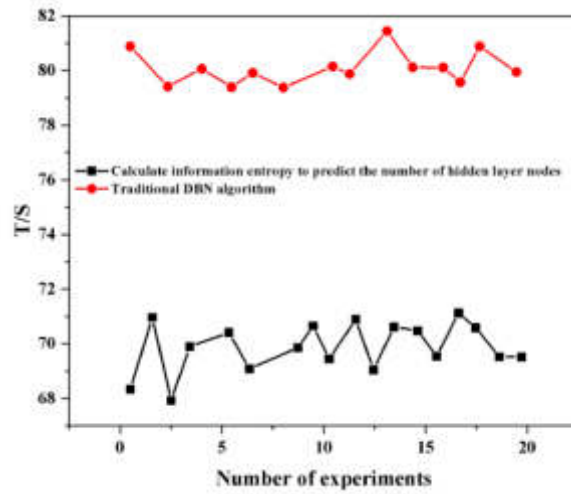


Fig. 3.2: Comparison of Calculation Time between Information Entropy Estimation of Hidden Layer Nodes and Traditional DBN Algorithm

there is sufficient classification information in the features. Therefore, we need to estimate the number of hidden layer nodes and the number of hidden layer layers to improve classification efficiency, save time and labor costs. During the calculation process, we found that for images with richer content, their information entropy increases, on the contrary, images with more uniform terrain types have relatively lower information entropy. Therefore, we attempt to apply information entropy to determine the number of hidden layer nodes used for classification, which greatly shortens the calculation time and improves classification efficiency. In order to verify the above proposed ideas, we conducted experiments using the following preset training parameters and compared the calculation time between traditional methods and estimation methods. The experiments were conducted more than 20 times. The running time results are shown in Figure 3.2.

The experiment compared the calculation of information entropy to estimate the number of hidden layer nodes and the training time of traditional DBN algorithms. When the maximum training period is set to 100, the method of calculating information entropy to estimate the number of hidden layer nodes has an average training time of 68.427 seconds, while the traditional DBN algorithm network training time is 80.952 seconds. Compared with traditional methods, the training time is shortened by 12.525 seconds, greatly reducing the training time and improving classification efficiency.

4. Conclusion. The author applied the DBN model in deep learning to conduct change detection research on City A and compared it with traditional classification methods. The experiment shows that deep learning has the highest producer accuracy, overall accuracy, and Kappa coefficient, with a relatively low misjudgment rate, and is most suitable for studying the trend of urban expansion and change. By calculating the information entropy of the image to predict the number of hidden layer nodes, the time of deep learning is greatly reduced. When the maximum training cycle is set to 100, the method of calculating the information entropy to predict the number of hidden layer nodes reduces the training time by 12.525 seconds compared to traditional methods, improving the efficiency of classification. The experiment proves that this method is suitable for the classification and change detection of urban expansion. Accurately characterizing the characteristics of urban expansion changes is of great significance for identifying its expansion regularity, further explaining the coordination between built-up area expansion and socio-economic development, analyzing the spatiotemporal evolution process, and predicting future evolution trends. On the basis of change detection, further analysis of the characteristics and mechanisms of urban expansion, analysis of influencing factors, and prediction of change trends can be carried out in the future.

5. Acknowledgement. Supported by: 1. the Public Culture Research Center of the Key Research Base of Humanities and Social Sciences of Universities in Hubei Province, Grant No. 2020GKY03Y; 2. Project source: Hubei Polytechnic University school level Horizontal Research Project, Project Name: Intelligent Information Management System of Construction Enterprise, Project No.: ky2022110.

REFERENCES

- [1] Mukai, N., Mori, K., & Takei, Y. (2022). Tongue model construction based on ultrasound images with image processing and deep learning method. *Journal of Medical Ultrasonics*, 49(2), 153-161.
- [2] He, L., Guo, C., Su, R., Tiwari, P., Pandey, H. M., & Dang, W. (2022). Depnet: an automated industrial intelligent system using deep learning for videobased depression analysis. *International Journal of Intelligent Systems*, 37(7), 3815-3835.
- [3] Dong, X., Mao, X., & Yao, J. (2023). A novel cardiac image segmentation method using an optimized 3d u-net model. *Journal of Mechanics in Medicine and Biology*, 23(09),111.
- [4] Chen, R., & Yang, B. (2022). Construction of an intelligent analysis model for website information based on big data and cloud computing technology. *Discrete Dynamics in Nature and Society*, 2022(35),4367.
- [5] Yan, J., He, Z., & He, S. (2022). A deep learning framework for sensor-equipped machine health indicator construction and remaining useful life prediction. *Comput. Ind. Eng.*, 172(42), 108559.
- [6] Yu, X., Li, W., Zhou, X., Tang, L., & Sharma, R. (2023). Deep learning personalized recommendation-based construction method of hybrid blockchain model. *Scientific Reports*, 13(1)67.
- [7] Zhang, X., Kong, J., Zhao, Y., Qian, W., & Xu, X. (2022). A deep-learning model with improved capsule networks and lstm filters for bearing fault diagnosis. *Signal, Image and Video Processing*, 17(4), 1325-1333.
- [8] Mehr, A. D., Ghiasi, A. R., Yaseen, Z. M., Sorman, A. U., & Abualigah, L. (2022). A novel intelligent deep learning predictive model for meteorological drought forecasting. *Journal of Ambient Intelligence and Humanized Computing*, 14(8), 10441-10455.
- [9] Zheng, Y., Lin, Y., Zhao, L., Wu, T., Jin, D., & Li, Y. (2023). Spatial planning of urban communities via deep reinforcement learning. *Nature Computational Science*, 3(9), 748-762.
- [10] Cheng, Y., Hu, X., Chen, K., Yu, X., & Luo, Y. (2023). Online longitudinal trajectory planning for connected and autonomous vehicles in mixed traffic flow with deep reinforcement learning approach. *Journal of Intelligent Transportation Systems*, 27(3), 396-410.

Edited by: Hailong Li

Special issue on: Deep Learning in Healthcare

Received: Jan 4, 2024

Accepted: Feb 13, 2024



SPORTS DATA PRIVACY PROTECTION AND INFORMATION SECURITY MANAGEMENT

BIAO JIN*

Abstract. In order to achieve the protection of personal privacy, the author proposes research on sports data privacy protection and information security management. The author used the two-dimensional fractional Fourier transform (2D-FRFT) method to encrypt the detected human body parts, which can be decrypted when needed for viewing. Compared to traditional Fourier transform, Fractional Fourier Transform (FRFT) can better express the time-frequency characteristics of signals and is very sensitive to the order of the transform. It is widely used in image encryption systems. 2D-FRFT increases the range of keys, further enhancing the security of the system. The author achieved encryption by extracting the detected human body parts and then performing a certain order of FRFT in the x and y directions respectively; When decrypting, use the same order of encryption to perform inverse fractional Fourier transform. Finally, based on research on pedestrian detection and encryption technology, the author designed a human-machine interaction interface that integrates the functions of detection and encryption interfaces, making the entire operation more intuitive and concise.

Key words: Pedestrian detection, Fractional Fourier transform, Interactive interface, Information Security Management

1. Introduction. Human body detection refers to the use of computers to detect human targets in image files, which has been studied and developed for many years. However, it is affected by factors such as shooting angle, human posture, background and lighting intensity, occlusion, and pedestrian gathering, its detection algorithm still needs continuous improvement [1]. At present, human body detection technology has been widely applied in intelligent video surveillance, robotics, virtual reality, and safe driving of vehicles. With the rapid development of society, more and more unstable factors have emerged. For safety reasons, intelligent video surveillance has been successfully applied in many public places, such as security checks, highways, elevators, banks, shopping malls, etc. When the system detects abnormal emergencies, it can understand, analyze and judge the behavior of the detection target, timely alarm and respond to corresponding measures, in order to ensure the safety of people's lives and property, and then minimize losses. Intelligent robot technology has been widely applied in various industries in society. Currently, intelligent robots are mainly used to complete high difficulty and high-risk actions that are not easy to manually complete. For example, in the event of an earthquake disaster, intelligent robots can be used to participate in search and rescue work, which to some extent improves search and rescue efficiency [2-3]. And the positioning of these disaster victims can be achieved through human detection technology. With the increasing number of private cars year by year, the time and space distance between people has been shortened, but the negative impact cannot be ignored. For example, traffic safety accidents that may occur at any time can threaten people's life and property safety. Therefore, research on safe driving of automobiles is particularly important. If a camera is installed on the vehicle that can collect information in front of the vehicle's line of sight, it can detect and recognize whether there are pedestrians in front, whether they are within a safe range, understand and analyze pedestrian behavior, predict the possibility of collision, and provide relevant information to the driver in a timely manner, so that the driver can make timely and correct responses and prevent frequent traffic accidents. Virtual reality is a high-tech that has developed in recent years. Its main principle is to use computer technology to construct a three-dimensional simulation space, providing users with simulations of visual, auditory, and tactile senses, giving them a sense of firsthand experience. There is enormous potential research value in medical surgery, entertainment games, military aerospace, and other fields. Among them, relevant positioning requires pedestrian detection technology to support. Human beings have always been in a dominant position in social activities, and they are also a

*Beijing University of Technology, Beijing, 100000, China (ab2023000@163.com)

key object of object detection. With the development of human detection technology, it has provided practical convenience for many fields and has broad development and application prospects.

Human detection is a research topic with strong practical application value. Many scholars at home and abroad have conducted in-depth research and achieved certain research results [4-5]. Since 1997, the US Defense Advanced Research Projects Agency has funded a major research project on visual surveillance technology and video understanding, with the participation of many universities. The European Union has also increased funding for pedestrian detection related technology research since 2000. Compared to foreign research in this field, China's development research is later. However, in recent years, the discovery of potential application value has attracted the attention of many domestic universities and scientific research institutions, and in-depth research has been carried out in this field. The State Key Laboratory of Pattern Recognition of the Chinese Academy of Sciences has made many outstanding achievements in this field. The effectiveness of pedestrian detection has a significant impact on the stability of video surveillance systems. So far, there are many methods for pedestrian detection, but there is no universal algorithm, and each method has its own characteristics. These methods mainly include frame difference method, background difference method, template matching method, optical flow method, and machine learning based detection method. The frame difference method mainly determines whether a pixel belongs to a foreground point by subtracting the corresponding grayscale values of the two frames in the video sequence, in order to determine the motion target that appears in the video sequence. But this method is limited to only pedestrians in the moving target. The background subtraction method is to subtract the current image frame from a known background model, in order to determine whether the pixels in the current image frame are foreground points. This algorithm has a simple principle and good real-time performance, but if there is no movement of pedestrians in the image, the detection will fail.

The template matching method requires prior knowledge of the characteristics of the detected object, such as contours, edges, etc. The principle is to compare these feature templates with the video image frames to be detected. If the template's features are met, it can be determined as the detection target. However, due to the fact that humans are non rigid bodies and have different postures, the actions presented at different times are not uniform, making it impossible to form an accurate and effective template, resulting in low accuracy in pedestrian detection. The optical flow method can detect independent moving targets without prior knowledge of any background prior. However, the optical flow method requires high hardware requirements for the equipment and has certain limitations in areas with high real-time requirements. Machine learning based methods can overcome many unfavorable conditions and have good stability. Due to the differences in height, posture, movement, clothing, skin color, lighting conditions, and background, pedestrian detection poses great difficulties. It is difficult to find a unified detection algorithm that meets real-time requirements and has a high recognition rate. In order to overcome these difficulties, pedestrian detection methods based on statistical learning are often used more frequently. This method mainly consists of three steps. Firstly, it extracts the features of pedestrians. Currently, the commonly used features include grayscale directional histogram features, hall features, and edge contour features, each with its own characteristics and suitable application scenarios; Then, the classifier is selected to train the extracted features and obtain file data that can be used for final classification; Finally, send the image to be detected to the classifier for detection. Many domestic scholars have conducted extensive research on statistical learning based pedestrian detection methods, which have the advantage of adapting to the variability of non rigid detection objects, and the accuracy of detected pedestrian targets is high and relatively stable. The disadvantage is that a large number of training samples are required. Sometimes, due to the high dimensionality of sample features, it takes a long time to train the classifier and use it for detection, and the real-time performance is not ideal.

With the deepening of research, the technology of pedestrian detection has achieved many achievements so far, but it still lacks a wider application and more stable performance. This is mainly due to the particularity of the pedestrian object, and its main difficulties are as follows: pedestrians are located with different backgrounds, which will also have a greater impact on the detection results. Some images have a single background, such as in the grassland, desert, open vision, etc., relatively pedestrians are relatively easy to detect. Some images have a complex background, such as in shopping malls, squares, tourist attractions and crowded roads and streets, the flow of people gathers, so there will inevitably be pedestrian blocking phenomenon, so that only a part of the human body can be seen in the image, so that enough information can not be obtained from the

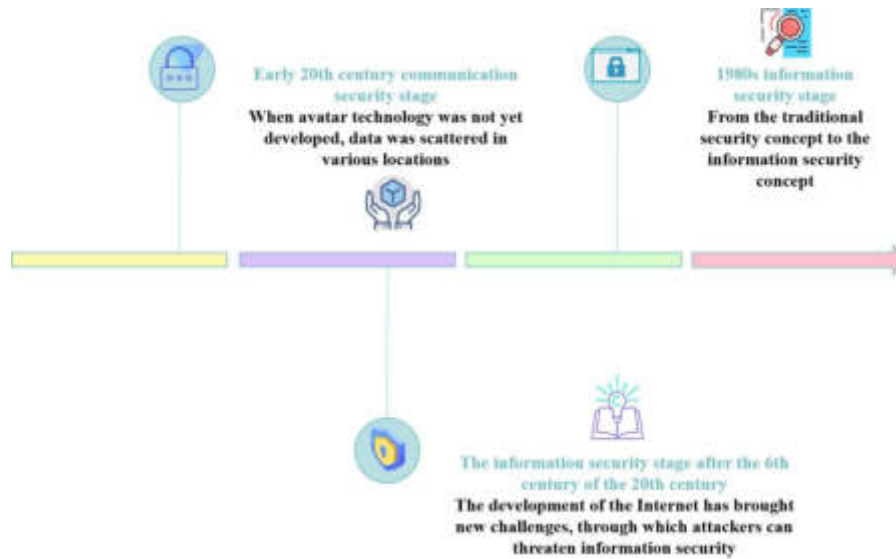


Fig. 1.1: Information Security Management and Privacy Protection Diagram

image, which affects the detector for analysis and judgment. At the same time, in some special backgrounds, some backgrounds are similar to the human body, such as tree trunk, telephone poles, etc., these disturbing background objects can easily be mistaken for the human body by the detector, thus reducing the detection accuracy. Camera can from the front, side, the back of the pedestrian, the results have the very big difference, and the influence of the distance between the camera and people, showing the pedestrian size is very different, and body parts also have difference, and in large traffic, sometimes in order to make the view more broad, need to improve the height of the camera, lead to the clarity of the photo is not quite the same. Because people are not rigid body, to a certain extent, both a certain rigidity, and a certain flexibility. Lead to the human body can show a rich change of posture, such as: upright, squatting, etc., even if the same person shows different movements, the test results-like also has a relatively large impact. At the same time, there are also differences in height, weight, clothing, etc., these characteristics will also bring some difficulties to detection. Even in the same place and under the same background, due to the different moments of photo collection, there can also be some differences in the light intensity. Some photos are brighter and some are darker. The differences in these rays will directly lead to the different image information extracted, which will ultimately affect the classification results. Therefore, the effect of reducing light intensity is also an aspect worth studying. Commonly used to describe the characteristics of the human body have edge, contour, texture features, gradient direction histogram features and the features back, these features can be used for pedestrian detection, but the results of different characteristics are often different, some feature detection results in addition to the algorithm used, will also be affected by the detection of image properties. So choosing a suitable feature is very important for the accuracy of the detection. In practical application, the system is often required to have good real-time, can quickly detect the human body part, and the corresponding understanding and analysis. However, the calculation amount of the algorithm directly restricts the real-time performance of the system, so the accuracy of detection is not affected as much as possible. It is very necessary to select a suitable algorithm above.

The author designed a concise human-computer interaction interface, which mainly includes image reading, detection of pedestrian parts, extraction of detected human parts, and encryption and decryption. The interface integrates all steps, making the operation of the entire privacy protection system more concise, intuitive, and user-friendly. Finally, the time complexity of all process modules in the system, including pedestrian detection, encryption, and decryption, was analyzed. Figure 1.1 shows information security management and privacy protection [6].

2. Methods. With the rapid development of computer and internet technology, network-based information exchange platforms have provided great convenience for the dissemination of digital works. However, the protection of property rights of digital works is also an issue that cannot be ignored [7]. Due to the intuitive and informative nature of image information, as well as the confidentiality requirements in certain specific fields, encryption and protection technologies for the growing demand for image information transmission are receiving increasing attention.

2.1. Introduction to the Development of FRFT. The reason why FRFT can be quickly and widely applied in the field of optics is because it is relatively easy to achieve this transformation based on optical devices [8,9]. Although FRFT has strong practical value in signal processing, this transformation lacks effective physical explanations and has low algorithm efficiency, resulting in it not receiving enough attention in signal processing. Nowadays, with the continuous in-depth research and development of FRFT, a large number of related research results have emerged.

2.2. Main Applications of FRFT. The unique properties of FRFT have attracted the attention of many learners and researchers. Nowadays, FRFT has been widely applied in many fields of scientific theory research and engineering application technology, such as wavelet transform, quantum mechanics, optical signal processing, artificial neural networks, video analysis, etc. The following mainly introduces some typical applications of FRFT in different fields.

(1) *Chirp class signal detection and parameter estimation.* The traditional Fourier transform is a transformation within the overall range that obtains the entire spectrum of a signal [10]. And FRFT can be seen as the decomposition of signals on orthogonal chirp basis, so FRFT is particularly suitable for the analysis and processing of chirp signals. In the detection of moving targets by radar, most of the received echo signals are chirp signals. The convenient processing of chirp signals by FRFT greatly improves the performance of signal processing systems in detecting moving targets and estimating their parameters.

(2) *Filtering.* The FRFT of a signal refers to the rotation of the signal at a certain angle in the time-frequency plane, which is very beneficial for the processing of non-stationary signals [11]. Traditional filtering methods mainly perform windowing operations in the frequency domain, but when the time-frequency coupling between the signal and noise is strong, it is difficult for traditional filtering methods to completely separate them. At this point, if the signal is rotated at a specific angle in the time-frequency plane, causing the signal and noise to lose coupling in the new Fourier domain, they can be separated well.

(3) *Neural networks.* Compared to traditional Fourier transform, FRFT has an additional transformation order p and can be freely selected from 0 to 1, making the transformation more flexible. In this case, if the neural network is in this domain, it can have a more stable effect [12].

(4) *Digital watermark.* The so-called digital watermark refers to embedding relevant marks in certain digital works in a specific way to verify copyright ownership. At the same time, this information should be ensured to be invisible and not perceived by anyone, and can only be detected by the copyright owner through special technical means. Digital watermarking technology is mainly aimed at protecting digital media. When property disputes arise, relevant information in the watermark can be extracted to verify copyright ownership. Due to the sensitivity of FRFT to the order of transformations, watermarking techniques based on FRFT correspond to different transformations at different orders. By adding key parameters, the security of the watermarking system is improved without knowing the order of transformations.

2.3. Definition of FRFT. The traditional Fourier transform is well-known to everyone, and its theoretical research and development are relatively mature, and it has been well applied [13]. And FRFT is based on the theoretical foundation of traditional Fourier transform, which is a refinement and supplement to it. The traditional Fourier transform can be regarded as a linear operator that rotates the angle $/2$ counterclockwise from the time axis to the frequency axis, while FRFT can be regarded as an operator that rotates any angle a . Compared to traditional Fourier transform, FRFT has greater advantages in application. It not only possesses some properties of traditional Fourier transform, but also adds some new characteristics related to its own properties. There are various definitions of FRFT based on different perspectives, but each definition has a certain inherent connection. Defining FRFT from different perspectives can help us have a more comprehensive understanding of it.

2.4. FRFT Properties and Characteristics.

(1) *The properties of FRFT.* FT represents the differential operator acting on the function, and F^P is the p-th power of FT, which means that if the function is rotated by an angle p, the p-th order FRFT can be understood as treating FP as an operator to generate FRFT [14]. The main properties of FRFT are as follows: Interchangeability: Perform FRFT on a function with order P1 first, after performing FRFT with order p2, the result is the same as performing FRFT with order P2 first and then FRFT with order p1, that is: $F^{P1}F^{P2} = F^{P2}F^{P1}$.

Order additivity: For different p1 and p2, there is always $F^{P1}F^{P2} = F^{P1}F^{P2}$.

Linear transformation: Satisfies the superposition principle, that is $F^P|\sum c_n f_n(u)| = \sum c_n|F^p f_n(u)|$.

Periodicity: According to $a=pr/2$, the period of order p is 4, which is $F^{p+4} = F^p$.

Reversibility: After performing p-order FRFT on a function, followed by-p order FRFT, the original function can be obtained. There are: $(F^p)^{-1} = F^{-p}$ [15].

(2) *The main characteristics of FRFT.* From the definition form of FRFT, it can be seen that FRFT reflects a time-frequency information of the signal, which is an extension of traditional Fourier transform and particularly suitable for processing non-stationary signals. Moreover, it has an additional transformation order p, making the transformation angle more flexible. Moreover, the discrete algorithm of FRFT has high efficiency and fast speed. Compared to traditional Fourier transform, FRFT mainly has the following characteristics:

The transformation order of FRFT can continuously increase from 0 to 1, demonstrating the continuous time-frequency variation characteristics of the signal, providing a better platform for time-frequency analysis of the signal [16].

FRFT can be seen as a decomposition of chirp groups. In radar signal processing, most of the echo signals for detecting moving targets are chirp signals, which is beneficial for improving the performance of signal processing systems in detecting moving targets and estimating their parameters.

FRFT has an additional transformation order p compared to traditional Fourier transform, which increases the value space of the key and improves the security performance of the system in image watermarking and encryption systems.

FRFT is a linear transformation without cross interference, which has advantages in the presence of additive noise.

FRFT is easy to achieve through optical transposition and has a wide range of applications in the field of optics.

The development of FRFT is relatively mature, with fast and efficient discrete algorithms, and it provides fast discrete algorithms for fractional convolution and other applications. At the same time, it also makes it easier to promote in practical applications.

2.5. 2D-FRFT. Perform one FRFT on f (x, y) in the x and y directions respectively to obtain 2D-FRFT. 2D-FRFT is implemented based on FRFT, and its kernel function can be expressed as:

$$K_{p1,p2}(x, y, u, v) = \frac{\sqrt{1 - j\cot\alpha}\sqrt{1 - j\cot\beta}}{2\pi} \times \exp\left[\left(\frac{x^2 + u^2}{2\tan\alpha} - \frac{xu}{\sin\alpha}\right)j\right]\exp\left[\left(\frac{y^2 + v^2}{2\tan\beta} - \frac{yu}{\sin\beta}\right)j\right] \tag{2.1}$$

In the above equation, $\alpha=p1n/2$, $\beta= P2 /2$ represents the two rotation angles of 2D-FRFT, respectively. If the corresponding transformation orders p1 and p2 are given, 2D-FRFT can be represented in the following form:

$$F^{p1p2}(u, v) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} K_{p1p2}(x, y, u, v) f(x, y) dx dy \tag{2.2}$$

Similar to FRFT, the inverse transformation of 2D-FRFT only requires taking the opposite order of the

corresponding order. As shown in the following equation:

$$f(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} K_{-p_1-p_2}(x, y, u, v) F^{p_1 p_2}(u, v) dudv \quad (2.3)$$

From equation 2.1, it can be seen that the transformation kernel of 2D-FRFT can be separated, which is:

$$K_{p_1, p_2}(x, y, u, v) = K_{p_1}(x, u) \times K_{p_2}(y, v) \quad (2.4)$$

The two-dimensional discrete FRFT can be seen as the result of two one-dimensional discrete FRFTs. The transformation process is as follows:

1. Firstly, perform discrete FRFT on the two-dimensional discrete signal in the x-direction to obtain F1;
2. Then perform discrete FRFT on the two-dimensional discrete signal in the y-direction to obtain F2;
3. Finally, by transposing F2, a two-dimensional discrete FRFT can be obtained.

When $\alpha = \beta$ When, it is symmetric 2D-FRFT; If $\alpha = \beta = /2$ is the traditional 2D-FRFT; When $\alpha \neq \beta$ When, it is an asymmetric FRFT [17].

3. Results and Analysis. The author designed and implemented a demonstration system that can integrate human body detection and privacy protection functions. This system can perform human body detection on input images and extract the detection part for encryption and decryption. The input image can be a real-time image captured by the camera or a pre saved image in the hardware device.

3.1. Introduction to System Interface Function Modules. As a demonstration system, it is required to be able to interact with users. Based on basic functional requirements, the interface of this demonstration system includes four parts: input area, functional area, display area, and operation instruction area. Input area: By inputting in the x and y directions, the detected pedestrian part can be encrypted and decrypted with any order of fractional Fourier transform. Functional area: Mainly realizes the input of images (supports real-time shooting or opening of saved images from the camera), detects pedestrians in the images, encrypts and decrypts the detected pedestrian parts, and has the function of exiting the system. The included function buttons include: Turn on the camera, turn off the camera, take and save photos, open pictures, save pictures, pedestrian detection, encryption, decryption, and exit. Display area: Mainly displays the input image, as well as the results of pedestrian detection, encryption, and decryption of the input image. Operation Instruction Area: This area mainly provides an operation instruction for the entire demonstration system and provides a brief summary of the system principles.

3.2. Demonstration System Design Principles. This interface is designed based on the GUI (Graphical User Interface) module of MATLAB. Human body detection is implemented on the open-source platform OpenCV. MATLAB is currently one of the most widely used mathematical software, but due to its use of line interpretation to execute code, it to some extent limits the execution speed of the code. The open-source code of the OpenCV class library is written in C and C++, and the code execution efficiency is high [18]. Combining the advantages of these two languages and leveraging their respective strengths often yields better results. When calling a cpp file in MATLAB, it is necessary to first convert the cpp file into a mex file in a certain format. The mex file is developed in C/C++ language, after being compiled in a certain way, it can be called by the m language interpreter in MATLAB.

Compared to interface software packages such as MFC, the GUI functions in MATLAB are simple, and the message mapping mechanism is concise, making it particularly suitable for system interface design that is not particularly demanding. MATLAB integrates an efficient interface development environment called Guide, which includes various MATLAB supported control objects, and users can choose different interface appearances and set different action response methods for controls. A complete GUI creation, firstly, it involves the selection of controls and spatial layout in interface design. Then, in order to respond to certain operations, it is necessary to write callback functions. Through callback functions, the specified functions of the controls can be achieved.

Table 3.1: Human detection time

Image size (pixels)	648*384	463*636	301*488	364*556
mean time to detect (ms)	2139.42	2865.59	1629.19	1927.04

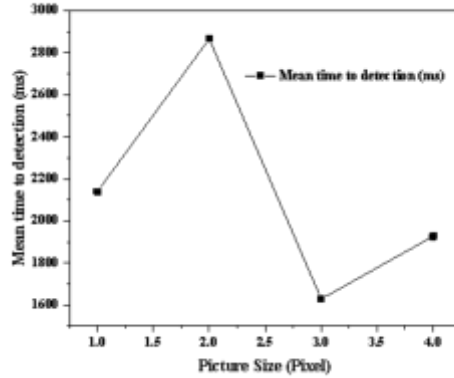


Fig. 3.1: Human detection time chart

Table 3.2: Detected encryption time of human body parts

Image size (pixels)	191*383	210*420	142*283	116*232
mean time to detect (ms)	834.15	1211.41	311.49	195.65

3.3. Experimental section. Firstly, load the saved image from the hardware device using the "Open Image" button, and use the "Pedestrian Detection" button to detect the human body part of the input image. After detecting pedestrians, the system extracts the human body part. Then, by setting the order in the x and y directions, the corresponding order of encryption can be achieved for the human body part.

A time complexity analysis was conducted on the human body detection, encryption, and decryption parts of the entire system, as shown in Tables 3.1, 3.2, and 3.3. The computer system used in the experiment was Windows 7 flagship version, and the software versions used were vs2010 flagship version, OpenCV2.44, and MATLAB 2013a. The main hardware environment of the computer is: Intel Pentium dual core T4300@2.10CHz Processor, memory 3.00GB [19].

The following table lists the image size and corresponding operation time. From Table 3.1 and Figure 3.1, it can be seen that the detection speed is related to the image size, and the larger the image, the longer the required time [20]. Detecting a $648 * 384$ image requires 2139.41 milliseconds, which takes a long time and has low real-time performance, this is because the selected HOG features have a dimensionality of 3781 and require complex computation. In order to ensure both detection rate and real-time performance, this is a major challenge that needs to be overcome in the current field of pedestrian detection. Tables 3.2, 3.3, Figure 3.2, and 3.3 respectively encrypt and decrypt the pedestrian parts detected in the images in Table 3.1. It can be seen from the tables that real-time performance is still a challenge that cannot be ignored. So, there are still many areas that need to be improved and enhanced for pedestrian detection and encryption in real-time videos.

4. Conclusion. The author mainly implemented it based on 2D-FRFT, not only analyzing the principle of FRFT for image encryption, but also providing the encryption and decryption effects of different orders in the x and y directions through experiments, as well as the slight deviation of the decryption order relative to the encryption order, which affects the decryption effect of the image. Enable readers to fully understand

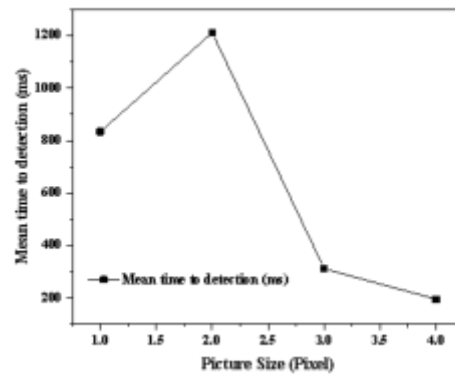


Fig. 3.2: Encrypted time graph of detected human body parts

Table 3.3: Decryption time for encrypted parts

Image size (pixels)	191*383	210*420	142*283	116*232
mean time to detect (ms)	831.56	1017.9	374.82	200.20

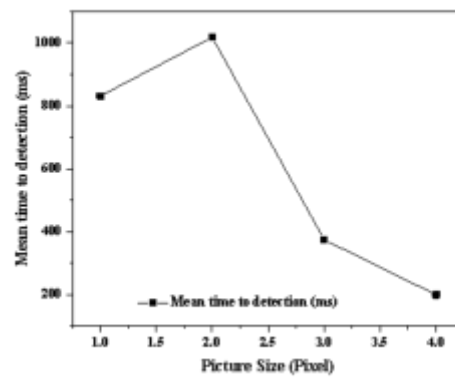


Fig. 3.3: Decryption time graph for the encrypted part

that different orders of FRFT have significant differences in image encryption and decryption effects. In the system design section, the author utilized the GraphicalUserInterface module in MATLAB, fully considering functional requirements and interface aesthetics, in order to design an interface that can take real-time photos and save, as well as both pedestrian detection and encryption and decryption of pedestrian parts. It has the characteristics of simple operation, complete functions, and beautiful interface.

REFERENCES

- [1] Guoquan, Z., Rongyuan, G., & Miao, H. (2023). Research on the mechanism of privacy lie flat on privacy protection behavior. *Library and Information Service*, 67(8), 129-140.
- [2] Wu, Q., Yin, J., & Ge, S. (2023). Research on the influence of social media emotion on information behaviour in information

- security events. *Journal of Information & Knowledge Management*, 22(05),22.
- [3] Fu, J. R. (2022). Research on the impact of big data tax collection and management on enterprise innovation. 2022 International Conference on Advanced Enterprise Information System 809(AEIS), 23-29.
 - [4] Jin, M. J. (2023). Computer network information security and protection strategy based on big data environment. *Int. J. Inf. Technol. Syst. Approach*, 16(78), 1-14.
 - [5] Hertzog, L., Chen-Charles, J., Wittesaele, C., Graaf, K. D., Titus, R., & Kelly, J. F., et al. (2023). Data management instruments to protect the personal information of children and adolescents in sub-saharan africa. *IASSIST Quarterly*.346(56),2114
 - [6] Si, Z. (2022). Research on security protection path of information infrastructure: how to identify the critical information infrastructure of the communication industry.68((234),8790
 - [7] Odutola, G. O., Ogbonyomi, M. A., & Umoru, C. O. (2023). Information resources management through ict application for e-governance: issues and prospects. *Library And Information Perspectives And Research*.6(99),24
 - [8] Qiu, Q. (2022). Research progress and trend of information technology methods for the protection of historical and cultural cities. *Proceedings of the 1st International Conference on Public Management, Digital Economy and Internet Technology*.55(99),976
 - [9] Miao, D., Lv, Y., Yu, K., Liu, L., & Jiang, J. (2023). Research on coal mine hidden danger analysis and risk early warning technology based on data mining in china. *Process Safety and Environmental Protection*, 171(357), 1-17.
 - [10] Hao, Y., Yasin, M. A. I., & Sim, N. B. (2023). A study on the influence of media use on college students' environmental protection behaviors. *Management of Environmental Quality: An International Journal*, 34(1), 177-191.
 - [11] Jinjin, G. (2022). Research on cloud computing and big data mining technology in the analysis of classified management data. 2022 IEEE 2nd International Conference on Power, Electronics and Computer Applications 79(ICPECA), 883-886.
 - [12] Oboudi, B. A., Elahi, A., Yazdi, H. A., & Pyun, D. Y. (2023). Impacts of game attractiveness and color of message on sport viewers' attention to prosocial message: an?eye-tracking study. *Sport, Business and Management: An International Journal*, 13(2), 213-227.
 - [13] Lina, W., Tianfang, D., Hong, Z., & Yang, L. (2023). Research on data quality control in electronic resource management: taking the practice of tsinghua university library based on alma system as an example. *Library and Information Service*, 67(10), 63-71.
 - [14] Hjort, J., Tukiainen, H., Salminen, H., Kemppinen, J., Kiilunen, P., & Snre, H., et al. (2022). A methodological guide to observe localscale geodiversity for biodiversity research and management. *The Journal of Applied Ecology*.25(66),88
 - [15] Riemenschneider, C. K., Burney, L. L., & Bina, S. (2023). The influence of organizational values on employee attitude and information security behavior: the mediating role of psychological capital. *Information & computer security*.4566(568),78
 - [16] Testa, M., Luciano, E. M., & Freitas, H. (2022). Management information systems and technologies: analysing research topics in france and brazil.45(57),23
 - [17] Oliveira, M. R. D., Sousa, T. B. D., Silva, C. V. D., Silva, F. A. D., & Costa, P. H. K. (2022). Supply chain management 4.0: perspectives and insights from a bibliometric analysis and literature review. *World review of intermodal transportation research: WRITR*,97890(1), 11.
 - [18] Salim, S., Turnbull, B., & Moustafa, N. (2022). Data analytics of social media 3.0: privacy protection perspectives for integrating social media and internet of things (sm-iot) systems. *Ad hoc networks*68(Apr.), 128.
 - [19] Haddas, R., Pipkin, W., Hellman, D., Voronov, L., Kwon, Y. H., & Guyer, R. (2022). Is golf a contact sport? protection of the spine and return to play after lumbar surgery:. *Global Spine Journal*, 12(2), 298-307.
 - [20] Yang, Z. C., Kuang, H., & Liu, J. X. (2023). Privacy protection model considering privacy-utility trade-off for data publishing of weighted social networks based on mst-clustering and sub-graph generalization. *International Journal of Modeling, Simulation, and Scientific Computing*, 14(04),768.

Edited by: Hailong Li

Special issue on: Deep Learning in Healthcare

Received: Jan 14, 2024

Accepted: Feb 22, 2024



DATA COLLECTION AND ANALYSIS BASED ON SENSOR TECHNOLOGY IN SPORTS TRAINING

XIANBIN SHI* AND HUAGANG ZOU†

Abstract. In order to realize the automatic monitoring of physical fitness index in athletic training, a new method of automatic monitoring and identification is put forward in this paper. In this paper, a network structure model is designed to automatically monitor the physical fitness index in sports training based on IOT and WSN. The collected a number of physical parameters in sports training, including heart rate, maximal oxygen absorption, respiration entropy, and load parameters. Then, by monitoring heart and lung function data, Cooper's method was used to measure the maximal oxygen intake, and EQO₂ was used as the training index. Then, the dynamic parameters of physical fitness index were extracted, and the change of gas metabolism and critical threshold were analyzed. This paper builds a system structure model to monitor the physical fitness index of physical training, and realizes the modular design of the system. The experiment results indicate that the system is not very different from the real one, and it can be used to automatically monitor the physical performance index in sports training. Practice has proved that the system has good stability and reliability, and is suitable for physical monitoring in the process of sports training.

Key words: Sensors, Sports training, Physical fitness indicators, automatic monitoring

1. Introduction. With the continuous progress of technology, the application of sensor technology in sports training is becoming increasingly common. Sensors, as an advanced technological tool, can collect and analyze athlete data, provide precise motion details and real-time feedback, thereby helping coaches and athletes improve training methods and technical levels. The development of sensor technology has brought tremendous changes and innovations to sports training [1]. The application range of sensor technology is very wide, which can be applied to various sports, including football, basketball, athletics, swimming, etc. Sensors can be embedded into the equipment of athletes, such as shoes, jerseys, protective gear, etc., or directly fixed on the sports field. Through real-time monitoring and data collection of sensors, coaches and athletes can obtain a large amount of information about the athlete's body posture, strength output, speed, acceleration, heart rate, and other aspects. These data can be used to analyze whether the athlete's technical movements are correct, whether the training intensity is appropriate, and whether their physical condition is good.

The application of sensor technology has revolutionized traditional observation and recording methods. Sensors can provide more objective and accurate data, reducing subjective interference. Sensors can monitor the movements of athletes in real-time, transmit data to computers or smart devices, analyze and process them through software, and generate visual results and reports [2]. Coaches can use this data to accurately analyze and evaluate athletes, develop targeted training plans, help athletes improve their technical skills, reduce injury risks, and optimize training outcomes. The application of sensor technology can also provide real-time feedback, helping athletes adjust their movements and postures in a timely manner. Sensors can transmit information to athletes through sound, light, vibration, and other means, guiding them to perform correct actions [3]. For example, in football training, sensors can remind athletes of the appropriate kicking force through vibration, indicate the correct angle when passing the ball through sound, and display the accuracy of the athlete's movement through light. This real-time feedback can help athletes correct mistakes faster, improve the accuracy and efficiency of technical movements.

The application of sensor technology can also promote communication and confrontation between athletes. Sensors can compare and analyze data from multiple athletes, helping coaches understand the differences and advantages among different athletes and develop corresponding training plans. At the same time, sensors can

*Anhui Business College, Wuhu, Anhui, 241002, China (Corresponding author's e-mail: shixianbin2021@abc.edu.cn)

†School of Physical Education, Anhui Normal University, Wuhu, Anhui, 241002, China

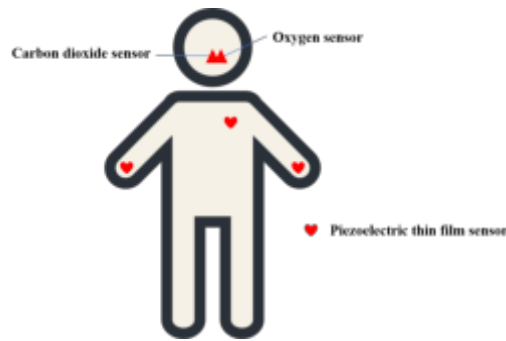


Fig. 2.1: Sensor Layout

also share data, allowing athletes to engage in real-time confrontation and competition [4].

For example, in swimming competitions, sensors can monitor the speed and posture of athletes in real-time, display data on screens next to the swimming pool, and allow the audience and coaches to clearly see the performance of each athlete, and compare and evaluate them. However, the application of sensor technology also faces some challenges and limitations. Firstly, the cost of sensors is relatively high, which may be difficult for some economically disadvantaged sports and athletes to afford. Secondly, the reliability and accuracy of sensors also need to be continuously improved and enhanced [5]. In the complex environment of sports fields, sensors may be subject to interference, resulting in inaccurate data. In addition, for some special sports such as gymnastics, judo, etc., the application of sensors may face technical difficulties. Due to the complex and varied actions of these projects, sensors may not be able to accurately capture and analyze relevant data.

Overall, the application of sensor technology in sports training has brought about significant changes and innovations. Sensors can collect and analyze athlete data, provide precise motion details and real-time feedback, thereby helping coaches and athletes improve training methods and technical levels. However, the application of sensor technology still faces some challenges and limitations, which require continuous improvement and refinement. With the further development of technology, it is believed that sensor technology will play a more important role in sports training, providing better support and guidance for the growth and progress of athletes [6]. The purpose of this study is to explore the application of sensor based data collection and analysis in sports training. By collecting sports data of athletes, such as posture, speed, strength, and other indicators, combined with real-time feedback provided by sensor technology, we can understand the performance of athletes and conduct scientific analysis. By collecting and analyzing data, more comprehensive and accurate training evaluations and guidance can be provided for coaches and athletes, thereby improving training effectiveness and competitive performance.

2. Overall System Architecture. In order to realize the automatic monitoring of the fitness index in the multi-sensor sports training, the overall structure of the system is based on the combination of the heart rate, the maximal oxygen absorption, the breathing entropy, and the collecting and analyzing of the load parameters. The ZigBee Network Sensor Monitoring System is used to build an auto-extracting model of Physical Fitness Indicator Parameters in Physical Training [7]. Combining with Physiology Parameter Recognition and Information Monitoring, a Hardware Device System for Monitoring and Detecting Physical Fitness Index in Sports Training with Wearable Device, Establishes Information Exchange Model and Command Transmission Control Model for Sports Training in XML and Web Middleware.

Wearable sensors such as oxygen, carbon dioxide and piezoelectric thin film sensors are used to collect heart rate, maximum oxygen uptake, respiratory entropy and load parameters, and video sensors are used to collect video data on exercise training as shown in Figure 2.1.

The maximum amount of oxygen is achieved by running or cycling with a mask on, depending on the difference between the data from the oxygen sensor and the CO₂ sensor; Calculate respiratory entropy based on the ratio of carbon dioxide production and oxygen consumption at the same time; When hemodynamic changes

occur, light enters the human body and undergoes predictable scattering. Using this principle, piezoelectric thin film sensors generate PPG waveforms for measuring heart rate and use heart rate data as basic biological characteristic values to ensure the accuracy of heart rate, load parameters, and other physiological parameter measurement results [8]. Arrange video sensors in the athlete training venue to collect athlete training video data, laying a solid data foundation for automatic monitoring of physical fitness indicators in subsequent sports training.

Among them, VO_2 , CO_2 emission (VCO_2), HR, etc., are the most important parameters in the automatic monitoring of ECG. The monitored the physical performance index of the sports training, which is divided into the primary and the secondary. Based on the change of gas metabolism and critical threshold, it can be used to automatically monitor the performance index in different sports programs and situations. Based on the general structure of the Physical Fitness Index Automatic Monitoring System in Sports Training, the SOA Framework Protocol is used. The automatic monitoring system of physical fitness index in physical training is composed of the main circuit control module, the data processing terminal module, the human-machine interaction module, and the bus output control module. Using JMS and HTTP protocols, we set up a system of service structure to monitor the performance index of sports training [9]. This article introduces an automatic monitoring system based on XML and Web middleware, which can monitor physical fitness index based on the M flag in the RTP header during exercise training. In sports training, we established an evaluation index called PE index and constructed a control model on the web service client to manage the components of PE index in sports training. In order to implement this system, we adopted a three-layer architecture design, which is the network layer, information fusion layer, and data output layer. Below, we will provide a detailed introduction to the functions and roles of these three levels.

Firstly, the network layer is responsible for handling communication between the system and external devices. In our system, motion training data is obtained through the M flag in the RTP header and transmitted to the information fusion layer for processing. The network layer is also responsible for communicating with web service clients, receiving and sending relevant data. Next is the information fusion layer, which is the core part of the system. In this layer, we use XML to encode and decode motion training data, and determine the type of motion based on the M flag. By analyzing and calculating exercise data, we can obtain the evaluation results of physical fitness index [10]. At the same time, the information fusion layer is also responsible for transmitting these evaluation results to the data output layer for users to view and analyze. Finally, there is the data output layer, which presents the evaluation results of physical fitness index to users in a visual form. Through the web service client, users can easily view and monitor changes in their physical fitness index. In addition, we can also control and adjust the composition of the PE index according to user needs to achieve better training results.

Through a three-layer architecture design, we have achieved effective communication and data processing between the network layer, information fusion layer, and data output layer. This system provides a scientific, convenient, and visual monitoring method for sports training, which helps to improve training effectiveness and the physical fitness level of athletes. Figure 2.2 shows a three-layer structural system [11].

On the basis of the three layers structure of the system, the dynamic parameters of the physical fitness index are extracted. Based on the change of gas metabolism and critical threshold, this paper establishes the structural model for the monitoring of the physical fitness index. Based on the data collection of HRMS and the physical function analysis, the data is synthesized in the sensor. In the application layer, the exchange of data is carried out, and in the network layer, the information exchange, the data fusion, and the characteristic output of the physical fitness index are realized [12]. The system's functional module architecture is illustrated in Figure 2.3.

The physical fitness index is one of the important indicators for measuring a person's energy level. By extracting and monitoring the dynamic parameters of physical fitness index, it is possible to better understand and evaluate an individual's physical and health status. To achieve this goal, this article proposes a physical fitness index monitoring system based on a three-layer system structure [13].

Firstly, at the bottom of the system, we construct a structural model of the stereoenergy index by analyzing gas metabolism and changes in critical thresholds. Gas metabolism is an important indicator of energy metabolism during human movement. By monitoring and analyzing gas metabolism, the individual's physical

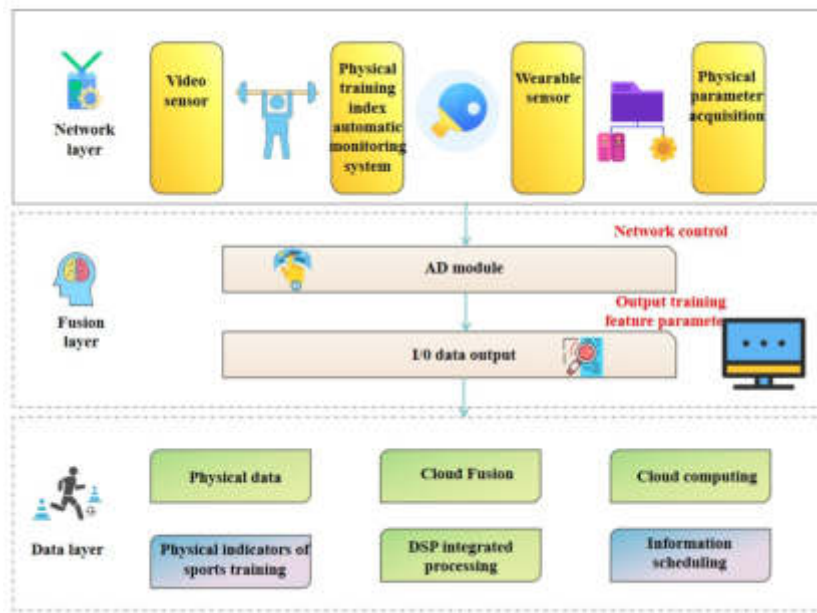


Fig. 2.2: The three-layer structural system of the system

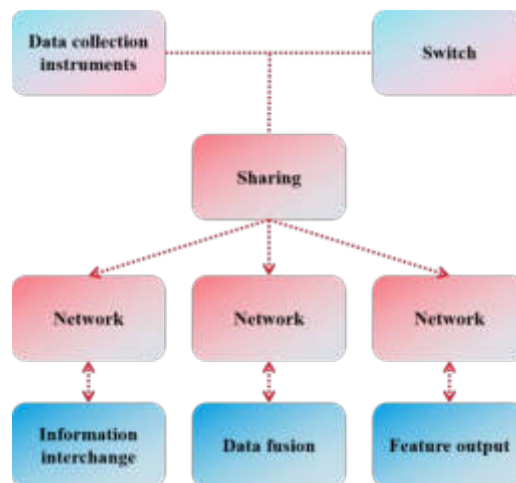


Fig. 2.3: System Function Module Structure

adaptability and ability level can be inferred. The critical threshold refers to the critical value of physical fitness index at a specific exercise intensity, exceeding which may lead to physical fatigue and overtraining. By comprehensively analyzing gas metabolism and critical thresholds, we can obtain an accurate physical fitness index model for monitoring an individual's physical condition [14].

Secondly, in order to collect and analyze data, we introduced HRMS (Exercise Physiology Monitoring System) as a data collection tool. HRMS can monitor individual physiological parameters such as heart rate, blood pressure, and body temperature in real-time, and transmit these data to the system's sensors. By analyzing and processing these data, we can obtain information on individual movement status and physical fitness. At the application layer of the system, we conducted data exchange and processing [15]. Through data

Table 3.1: Prior parameters of VO_{2max} monitoring distribution

Pilot projects	gender	VO_{2max}
laboratory experiment	male	0. 43
	female	0. 45
Outdoor experiment	male	0. 87
	female	0. 73

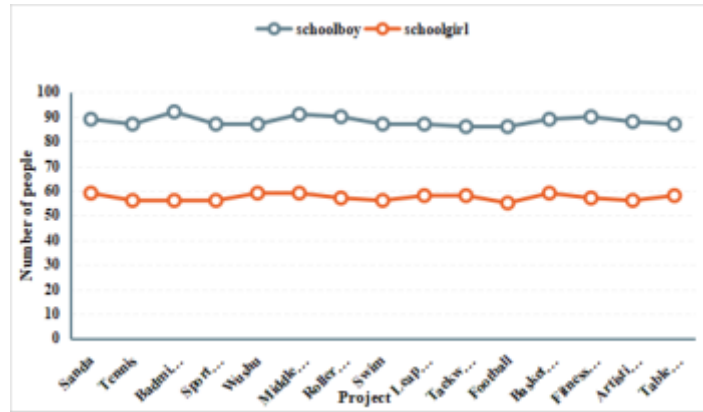


Fig. 3.1: Experimental Test Objects

transmission with HRMS, we can obtain real-time physiological parameter data of individuals and transmit it to the network layer for processing. At this level, we obtained the physical fitness index of individuals through data fusion and feature extraction. The physical fitness index is an evaluation index that takes into account individual physiological parameters, exercise ability, and physical condition, and can objectively reflect an individual's physical fitness level and health status.

Finally, at the network layer of the system, we achieved feature output for information exchange, data fusion, and physical fitness index. The network layer serves as a bridge for data transmission and processing, integrating and analyzing data from sensors, and outputting the final evaluation results. Through the operation of the network layer, we can achieve real-time monitoring and evaluation of individual physical fitness index [16].

3. System data analysis and experimental testing.

3.1. Experimental Method and Object. In the data analysis, the maximal oxygen uptake was measured by Cooper's method, and EQO_2 was used as the training target. In this paper, the distribution of physical fitness index was established, and the differences of the maximal oxygen uptake VO_{2max} , O_2 Pmax and $METS_{max}$ were analyzed. Physical fitness index monitoring was performed, and the original VO_{2MAX} monitoring profile of each group was given, as shown in Table 3.1.

Sports such as Sanda, tennis, badminton, sports dance, martial arts, and roller skating were selected as the test subjects, and the distribution of test participants is shown in Figure 3.1.

Based on the distribution of test subjects in Figure 3.1, this paper presents a method to monitor the performance of physical fitness index in the design of the automatic monitoring system. The participants were classified into two categories: technology, combat, long and short distance running, and ball. Figure 3.2 illustrates the distribution of monitoring information density for sensor sampling for various projects [17].

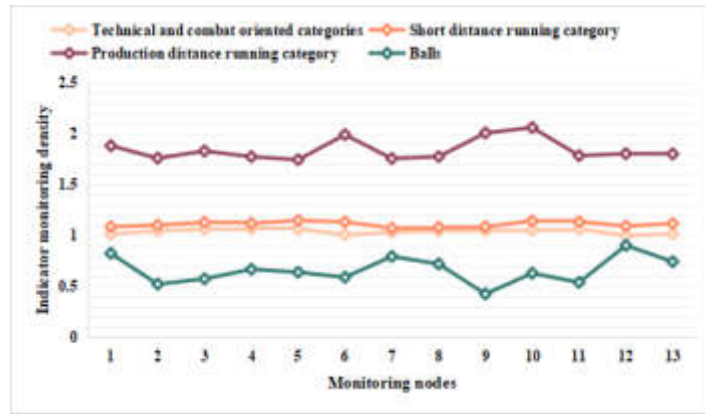


Fig. 3.2: Concentration distribution of monitoring information sampled by sensors of different projects

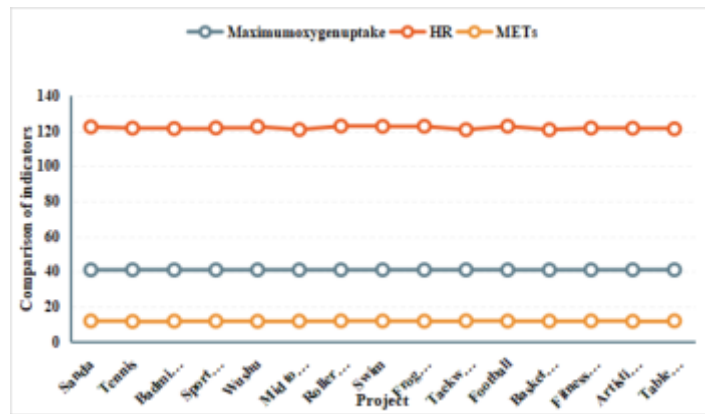


Fig. 3.3: Comparison of Various Indicators for Men’s Long Distance Running in Various Events

Table 3.2: Comparison of Various Indicators for Long Distance Running in Comprehensive Projects

	Technical and combat oriented categories	Short distance running category	Long distance running category	Balls
VO_{2max}	43.9 ± 2.54	32.4 ± 3	31.41 ± 2.23	30.3 ± 2.54
S	2032.46 ± 155.43	1852.11 ± 245.12	1546.46 ± 143.43	1653.32 ± 145.3
METs	932 ± 0.6	8.35 ± 1.6	8.12 ± 14.34	8.21 ± 0.46

According to the analysis chart, using this method for monitoring physical fitness indicators in sports training, the correlation between the feature distribution and $P < 0.05$ and $P < 0.01$. Meet the functional monitoring requirements for sports training indicators. Based on this, the comparison of various indicators for long-distance running in each project is shown in Figure 3.3. The comparison of various indicators for long-distance running in various projects is shown in Table 3.2 [18,19,20].

We test the accuracy of the training performance curve obtained as shown in Figure 3.4.

Through analyzing the above monitoring results, we can draw a conclusion that there is little difference between the monitoring of physical fitness index and the real one. The system can be used to automatically

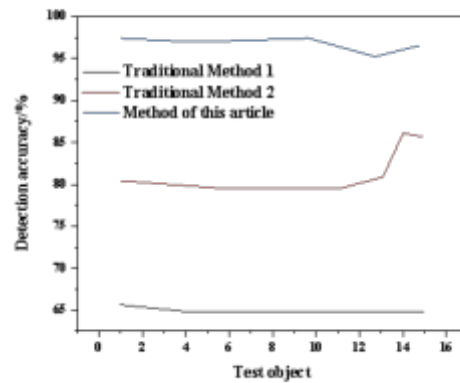


Fig. 3.4: Detection performance curve

monitor the physical performance index in physical training, and it is very adaptable to the environment and the individual.

4. Conclusion. An automatic monitoring system of physical fitness index in physical training is designed on the basis of multiple sensors. The whole structure and function modules of this system are expounded in detail. On this basis, a health index monitoring system based on wireless sensor network is proposed. Through the statistical analysis of large sample data, it is proved that the method has high credibility, can adapt to the personalized monitoring of large-scale sports activities and enhance the dynamic monitoring of BMI.

REFERENCES

- [1] Mei, Z. (2023). 3d image analysis of sports technical features and sports training methods based on artificial intelligence. *Journal of Testing and Evaluation: A Multidisciplinary Forum for Applied Sciences and Engineering*,54(7),69-72.
- [2] Wang, D., & Huang, G. (2022). Analysis of the influence of outward bound training based on data analysis in college physical training. *Computational intelligence and neuroscience*, 2022(33), 6488562.
- [3] Lee-Cultura, S., Sharma, K., & Giannakos, M. (2022). Children's play and problem-solving in motion-based learning technologies using a multi-modal mixed methods approach. *International Journal of Child-Computer Interaction*(Mar.), 31(74),23-25.
- [4] Zinnatullin, V., & Koledin, S. (2022). Visual development environment and visual programming as an effective tool for data collection and analysis. *2022 VIII International Conference on Information Technology and Nanotechnology (ITNT)*, 58(7),1-4.
- [5] Gandhi, V., Sardar, A., Wani, P., Borase, R., & Gawande, J. (2022). Iot based wireless data technology using lora and gsm. *ITM Web of Conferences*,96(4),102-105.
- [6] Zhang, Y., Liu, L., Wang, M., Wu, J., & Huang, H. (2022). An improved routing protocol for raw data collection in multihop wireless sensor networks. *Computer Communications*, 188(24), 66-80.
- [7] Hao, Y., Li, S., & Zhang, T. (2022). Multi-sensor optimal deployment based efficient and synchronous data acquisition in large three-dimensional physical similarity simulation. *Assembly Automation*,547(1), 42.
- [8] Cao, J., Li, H., & Zhang, X. (2022). Multitarget identification method for dual-plane detection based on data fusion and correlation analysis. *Microwave and Optical Technology Letters*,652(7),56-58.
- [9] Wang, D. (2022). Analysis and research on regeneration therapy of athlete tendon injury based on nanometre sensor technology. *International Journal of Nanotechnology*,63(74),54-57.
- [10] Su, Y. S., & Hu, Y. C. (2022). Applying cloud computing and internet of things technologies to develop a hydrological and subsidence monitoring platform. *Sensors and materials: An International Journal on Sensor Technology*,96(4 Pt.1), 34.
- [11] Chan, Y. K., Lai, J. C. M., Hsieh, M. Y., & Meen, T. H. (2023). Important elements of sensor technology and data management and related education. *Sensors and materials: An International Journal on Sensor Technology*,754(4 Pt.1), 35.
- [12] Zhang, R. (2022). College sports decision-making algorithm based on machine few-shot learning and health information mining technology. *Computational intelligence and neuroscience*,65(7),458-462.
- [13] Lai, S. C., Yang, M. L., Wang, R. J., Jhuang, J. Y., Ho, M. C., & Shiau, Y. C. (2022). Remote-control system for elevator with sensor technology. *Sensors and materials: An International Journal on Sensor Technology*,1784(5 Pt.1), 34.
- [14] Tahmid, K. T., Ahmed, K. R., Chowdhury, M. N., Mallik, K., Habiba, U., & Haque, H. M. Z. (2022). An integrated

- crowdsourcing application for embedded smartphone sensor data acquisition and mobility analysis. *Journal of Advances in Information Technology*, 96(5), 13.
- [15] Li, X., & Li, Y. (2022). Sports training strategies and interactive control methods based on neural network models. *Computational intelligence and neuroscience*, 2022(47), 7624578.
- [16] Chao, L. I., Tokgoz, K. K., Okumura, A., Bartels, J., Toda, K., & Matsushima, H., et al. (2022). A data augmentation method for cow behavior estimation systems using 3-axis acceleration data and neural network technology. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, E105.A(4), 655-663.
- [17] Rajagopal, V., Velusamy, B., & Rathinasamy, S. (2023). Double q-learning-based adaptive trajectory selection for energy-efficient data collection in wireless sensor networks. *International journal of communication systems*, 652(7), 526-528.
- [18] Kulkarni, A. R., Kumar, N., & Rao, K. R. (2023). Efficacy of bluetooth-based data collection for road traffic analysis and visualization using big data analytics. *Big Data Mining and Analytics*, 6(2), 139-153.
- [19] Navarro, M., Liang, Y., & Zhong, X. (2022). Energy-efficient and balanced routing in low-power wireless sensor networks for data collection. *Ad hoc networks*(Mar.), 127(74), 63-68.
- [20] Taami, T., Azizi, S., & Yarinezhad, R. (2023). Unequal sized cells based on cross shapes for data collection in green internet of things (iot) networks. *Wireless Networks*, 29(24), 2143 - 2160.

Edited by: Hailong Li

Special issue on: Deep Learning in Healthcare

Received: Jan 19, 2024

Accepted: Feb 22, 2024



INTRODUCTION TO THE SPECIAL ISSUE ON NEXT GENERATION PERVASIVE RECONFIGURABLE COMPUTING FOR HIGH PERFORMANCE REAL TIME APPLICATIONS

C. VENKATESAN*, YU-DONG ZHANG†, CHOW CHEE ONN‡ AND YONG SHI§

The evolution of scientific computing is reshaping the hardware and software requirements, emphasizing the need for high-performance platforms adaptable to real-time applications. Traditional methods with general-purpose processors often lack the agility needed for swift modifications and fast computations during real-time tasks. Reconfigurable computing offers a compelling solution by integrating hardware speed and software flexibility on a unified platform. This technique promises significant acceleration across diverse applications like image processing, encryption, decryption, runtime operations, and intensive computing tasks such as sequence searching and matching in smart environments.

For high-performance applications, software-programmed microprocessors offer versatility. Artificial Intelligence (AI) has proven more robust than traditional methods in noisy environments, relying on reconfigurable designs for efficient operation. Hybrid machine-learning techniques enhance system reliability without compromising performance. Reconfigurable computing systems have recently accelerated intensive algorithms compared to software-optimized versions, leveraging parallel topologies. Deep Neural Architectures with adaptable computation patterns excel in computer vision tasks. Artificial Neural Networks (ANNs) are crucial for pattern recognition, high-performance machine learning, data manipulation, security threats, data mining, signal processing, and other applications, driving ongoing research into reconfigurable hardware and software implementations for ANNs.

It is a privilege for us to introduce the Special Issue on “Next generation pervasive reconfigurable computing for high-performance real-time applications”. Among the numerous research papers we received (49 in total), we meticulously selected 22 papers for publication. The overarching objective of this special issue is to investigate the recent advancements and disseminate state-of-the-art research related to reconfigurable computing for high-performance real-time applications and the technologies that make this possible. This special issue represents a showcase of new dimensions of research, offering researchers and industry professionals an illuminating perspective on pervasive reconfigurable computing. We sincerely hope that the contributions in this special issue will not only inform but also inspire future research endeavours, leading to a deeper understanding of the multifaceted world of speed computation during real-time applications.

The paper titled “Vulnerability Detection in Computer Networks Using Virtual Reality Technology” by Songlin Liu, traditional network security challenges are tackled through virtual reality integration. Optimization calculations are employed to extract network security vulnerability attributes, with adjustments made using a web crawler and a detailed analysis of attack characteristics. This approach facilitates automated detection within a virtual reality framework. Empirical results show a significant reduction in detection delay to 75.33 milliseconds, compared to 290.11 milliseconds and 337.30 milliseconds in conventional methods, highlighting the efficiency of the proposed approach.

“Computer Malicious Code Signal Detection Based on Big Data Technology” by Xiaoteng Liu improves

*Department of Electronics and Communication Engineering HKBK College of Engineering, Bangalore, Karnataka, India. (venkatesanc.ec@hbk.edu.in)

†Chair in Knowledge Discovery and Machine Learning, Department of Informatics, University of Leicester, United Kingdom (yz461@leicester.ac.uk)

‡Department of Electrical Engineering, Faculty of Engineering, University of Malaya, Kuala Lumpur, Malaysia (cochow@um.edu.my)

§Department of Computer Science, Kennesaw State University, Marietta, United States of America. (yshi5@kennesaw.edu)

upon traditional methods by using big data for detecting malicious software behaviour. The approach tackles challenges in mobile malware detection through mean-variance feature selection and advanced techniques like PCA, KLT, and ICA for feature extraction. A decision tree-based multi-level classification model enhances accuracy and addresses data imbalance issues, leading to notable accuracy improvements of 3.36% to 6.41% across different Android detection methods, highlighting the effectiveness of the proposed malware detection technology.

"Network Security with VR-Based Antivirus Protection and Reduced Detection Delays" by Chunna Song et al., addresses delays in traditional systems by leveraging VR technology. It aims to decrease detection delays, enhance efficiency, and optimize security feature identification. The method includes web crawling for an injection list and virtual protection blocks to mitigate threats, achieving a detection delay of 75.33 milliseconds, surpassing traditional delays of 290.11 ms and 337.30 ms. Empirical evidence supports the efficacy of automatic detection in VR, promising improved network security responsiveness and effectiveness.

Xiaohong Li et al., in the paper titled "Computer Network Virus Defense with Data Mining-Based Active Protection" presents a novel approach to enhance computer network virus defence beyond traditional technologies. Utilizing Object-Oriented Analysis (OOA) mining, the method analyzes Win API call sequences in PE files to detect deformed and unknown viruses. Experimental results demonstrate the Data Mining-based Antivirus (DMAV) system's superiority with higher accuracy in deformed virus detection, effective defence against unknown viruses (92% recognition rate), improved efficiency, and reduced false alarms for non-virus files. The research introduces an OOA rule generator to optimize feature extraction, bolstering system intelligence and resilience in enhancing computer network security.

"Application of Nonlinear Big Data Analysis Techniques in Computer Software Reliability Prediction" by Li Gao and Hai Wang addresses challenges in using artificial neural networks for software reliability prediction, focusing on improving the PSO-SVM model. Comparative experiments with a Backpropagation (BP) model highlight the PSO-LSSVM model's rapid reduction in training error within 200 generations, compared to BP's 1,733 generations. The optimized PSO-LSSVM model demonstrates superior efficiency with small sample sizes, offering accelerated training and enhanced prediction accuracy for software reliability assessments.

"Enhancing Industrial Control Network Security through Vulnerability Detection and Attack Graph Analysis" by Yan Liao addresses communication attack defence gaps in industrial control networks. The study proposes using attack graphs to improve security and vulnerability assessments, providing detailed construction methodologies. Experimental evaluations using the "earthquake net" virus identify four main attack routes for the "Zhenwang" virus, each with specific loss values and attack success probabilities. This research emphasizes the need for systematic vulnerability analysis to enhance overall industrial control network security.

Chunmei Ji et al., in the paper titled "Improving Semantic Analysis in Visualization with Meta Network Representation and Parsing Algorithm" introduces the semantic Meta Network (MNet) for advanced semantic analysis in visualization. MNet employs a hierarchical framework to integrate semantic elements, relationships, and attributes, facilitating comprehensive understanding from phrases to complete texts. The study presents a construction algorithm for MNet and parsing methods tailored for natural language interface parsing, validated through empirical experiments. This research enhances semantic analysis capabilities, particularly in interpreting SCADA system instructions, contributing to improved natural language understanding and semantic analysis in visualization contexts.

"Hybrid optimization for high aspect ratio wings with convolutional neural networks and squirrel optimization algorithm" by Pengfei Li presents an efficient approach for optimizing lightweight high-aspect-ratio wings. The study combines a hybrid binary unified coding description, one-dimensional convolutional neural networks for aeroelastic modelling, and the squirrel optimization algorithm for computational efficiency. Experimental results demonstrate a 4.1% reduction in wing weight, showcasing the effectiveness of this hybrid method in optimizing complex wing structures.

"Computer Software Maintenance and Optimization Based on Improved Genetic Algorithm" by Ming Lu aims to enhance software maintenance and network performance using an advanced genetic algorithm. The study refines network architecture through enhanced genetic operations and evaluates satisfaction and fitness index functions with controlled data iterations. Results show network reliability initially improves with iterations but stabilizes due to hardware limitations, highlighting a peak reliability of 0.894 achieved at 100

iterations. This research provides a foundational framework for optimizing computer network reliability with a balanced genetic algorithm approach.

Jing Wang et.al., in the paper titled "Research on Intelligent Transformation Platform of Scientific and Technological Achievements Based on Topic Model Algorithm and Its Application" enhances the conversion of scientific breakthroughs into practical applications using the LDA theme model. The study improves efficiency by extracting pivotal terms and thematic phrases, facilitating information management, retrieval, and recommendations for academic and business sectors. The platform evaluates transformation results and operational efficiency in advancing scientific and technological achievements.

"An Emotional Analysis of Korean Topics Based on Social Media Big Data Clustering" by Yanhong Jin introduces the Online Topic Emotion Recognition Model (OTSRM) to enhance emotional analysis accuracy in Korean social media. Utilizing the Online Latent Dirichlet Allocation (OLDA) model, OTSRM integrates emotion intensity and employs an innovative emotion iteration framework. It introduces an affective evolution channel and distribution matrices for characteristic and affective words, advancing understanding of emotional context. Validation experiments demonstrate OTSRM's effectiveness with emotion recognition accuracy rates of 85.56% and 81.03%, improving precise emotional dynamics assessment in Korean social media.

Xiangying Liu et. al., in the paper titled "Application of Artificial Intelligence Technology in Electromechanical Information Security Situation Awareness System" proposes leveraging AI and big data to enhance predictive capabilities in information security. Using LSTM-RNN and variant GRU models, the study achieves high accuracy with LSTM-RNN showing MAPE of 8.79%, RMSE of 0.1107, and RRMSE of 8.47% on test data. This research highlights AI's potential in developing robust information security situational awareness systems, comparing LSTM and GRU models for effectiveness and efficiency in predictive analysis.

Wenjuan Yang in the paper titled "Analysis and Application of Big Data Feature Extraction Based on Improved K-means Algorithm" addresses challenges in handling large volumes of data by proposing an enhanced K-means algorithm. Focused on mitigating errors, the study applies this algorithm to monitor power equipment, achieving an error rate below one per cent and consistently high accuracy exceeding 95%. The research underscores the algorithm's effectiveness in improving clustering accuracy and efficiency, emphasizing its practical application in big data analysis for energy systems.

"Intelligent Prediction of Network Security Situations Based on Deep Reinforcement Learning Algorithm" by Yan Lu et al. introduces a novel approach using deep learning to enhance network security assessment. The study develops a deep self-encoding model for effective network attack detection and integrates a unique oversampling weighting algorithm to improve pattern detection with limited training data. Experimental results demonstrate significant performance gains over traditional methods like DT, SVM, and LSTM, achieving higher F1 scores by approximately 2.77, 10.5, and 5.2, respectively. This research emphasizes improved accuracy and efficiency in predicting and measuring network security situations.

Min Yan and Hua Zhang in the paper titled "Interface Control and Status Monitoring of Electronic Information Equipment Based on Nonlinear Data Encryption" proposes an advanced system ensuring information fairness, objectivity, and security in traffic accident investigations. The study develops a robust security strategy with PC-based platforms for efficient data acquisition, secure processing, transmission, and storage using nonlinear data encryption methods. Performance tests with files ranging from 3MB to 10MB show a significant 25% average speed improvement over the original platform. This system enhances data integrity during transmission, aids in accurate equipment identification post-accident, and addresses critical security challenges in traffic accident investigations.

Yongqiang Shang in the paper titled "Detection and Prevention of Cyber Defense Attacks Using Machine Learning Algorithms" examines the rise in cyber threats fueled by enhanced computing power, big data utilization, and advanced machine learning techniques. The study explores both defensive and offensive uses of machine learning in cybersecurity, focusing on mitigating attacks targeting machine learning models. It underscores the pivotal role of artificial intelligence in enhancing digital security through tasks like malware analysis, network vulnerability assessment, and threat prediction amidst rapid global digitization.

Zhixiong Xiao in the paper titled "Minimizing Overhead Through Blockchain for Establishing a Secure Smart City with IoT Model" explores the integration of blockchain technology with IoT to enhance security while addressing resource constraints. Traditional blockchain deployments are impractical for IoT due to their

high storage and computational demands. This study proposes an optimized blockchain framework tailored for IoT devices, ensuring essential security measures like authenticity, credibility, confidentiality, availability, and non-repudiation. The approach aims to bolster security without imposing excessive resource burdens, making it suitable for smart city applications.

"Security and Privacy of 6G Wireless Communication Using Fog Computing and Multi-Access Edge Computing" by Ting Xu, Ning Wang, Qian Pang, and Xiqing Zhao addresses data confidentiality challenges in forthcoming 6G networks. The study explores blockchain's potential for enhancing security and incorporates machine learning (ML) to manage vast data volumes. It reviews strategies for protecting automotive communication systems and assesses confidentiality approaches within 6G architecture. The research emphasizes safeguarding private data in the Internet of Everything (IoE) era and proposes ML solutions for data processing challenges, underscoring blockchain's role in securing 6G communications.

"Sustainable Development in Medical Applications Using Neural Network Architecture" by Shuyi Jiang aims to enhance risk management in healthcare through machine learning. The study uses social media data analysis to identify and assess threats, aiding informed decision-making in healthcare management. It employs machine learning algorithms to visualize risk categories and simplifies data processing for efficiency. Empirical analysis of Consumer Value Stores (CVS) reveals operational, financial, and technological risks, categorized by severity. The study highlights the framework's effectiveness in threat identification and mitigation, offering insights for safer healthcare environments.

"Target Image Processing Based on Super-Resolution Reconstruction and Machine Learning Algorithm" by Chunmao Liu proposes a method to enhance the resolution of medical images using super-resolution reconstruction and machine learning. This approach integrates nonlocal autoregressive learning into the reconstruction model, exploiting inherent data similarities in medical images. Additionally, a clustering algorithm refines classification dictionaries to improve efficiency. Experimental results, conducted on CT/MR images, demonstrate significantly improved peak signal-to-noise ratios and structural similarity values compared to other methods, achieving a peak value of 31.49. This study validates the efficacy of combining super-resolution reconstruction and machine learning for enhancing medical image resolution.

"Group Intelligent City Mobile Communication Network's Control Strategy Based on Cellular Internet of Things" by Jiazheng Wei explores enhancing mobile communication networks in urban areas using an improved particle swarm optimization (PSO) algorithm. This study refines PSO to tackle specific network challenges and achieves robust optimization results through simulations. The Grad-PSO algorithm effectively improves network performance with optimized parameters, demonstrating its suitability for enhancing mobile communication efficiency in urban environments.

Shuiquan Zhu presented the paper titled "Performance evaluation of micro automatic pressure measurement sensor for enhanced accuracy" focuses on enhancing the accuracy of micro-automatic pressure measurement sensors through design improvements and rigorous performance testing. The sensor achieved high detection accuracy (0.0452%) and met reliability standards with features such as sensitivity (0.0582%), nonlinearity (0.0741%), hysteresis (0.0266%), and repeatability (0.0625%). It also demonstrated reduced response times compared to traditional sensors, though the study highlights the necessity for additional real-world testing to optimize its performance further.

In summary, this special issue of Scalable Computing: Practice and Experience explores emerging trends in scientific computing that influence hardware and software requirements. It underscores the need for high-performance platforms capable of real-time architectural updates. Traditional computing struggles with adaptability and speed in real-time applications, while reconfigurable computing integrates hardware speed and software flexibility in a unified platform. This approach accelerates applications like image processing, encryption, and machine learning, especially Artificial Neural Networks (ANNs). The issue highlights advancements in reconfigurable computing systems and their impact on enhancing performance in intensive computing domains such as signal processing and computer vision.



ML-CSFR: A UNIFIED CROP SELECTION AND FERTILIZER RECOMMENDATION FRAMEWORK BASED ON MACHINE LEARNING

AMIT BHOLA* AND PRABHAT KUMAR†

Abstract. Sustainable and substantial crop production is essential globally, especially considering the increasing population. To achieve this, selecting appropriate crops and applying necessary fertilizers are pivotal for ensuring satisfactory crop growth and productivity. Farmers have relied heavily on intuition when choosing which crops to cultivate and suitable fertilizers to use in a given season. However, this traditional approach often needs to consider the significant impact of current environmental and soil conditions on crop growth and yield. Overlooking these factors can have far-reaching consequences, impacting not just individual farmers and their households but also the entire agricultural sector. The integration of machine learning offers a promising avenue for addressing these challenges and providing practical solutions. The core contribution of this research lies in proposing a unified framework termed Machine Learning-enabled Crop Selection and Fertilizer Recommendation (ML-CSFR). This framework's primary objective is to predict appropriate crops accurately and subsequently suggest corresponding fertilizers based on specific agricultural conditions. The initial phase involves the identification of proper crops for individual farmlands, considering local input variables. This phase employs artificial neural networks (ANN) to filter crops effectively using the available choices. The next phase utilizes soil and environmental parameters to anticipate the optimal fertilizer for the selected crops. This phase leverages the XGBoost (XGB) model to predict the most suitable fertilizers accurately. This two-phase approach ensures a comprehensive and effective recommendation system for enhancing agricultural outcomes. Experimental results demonstrate the effectiveness of this framework, achieving an accuracy score of 99.10% using ANN and 97.66% for XGB. The framework's capability to deliver tailored recommendations for individual farms and its potential to integrate real-time sensor data positions it as an effective tool for improving agricultural decision-making.

Key words: Machine learning, Fertilizer recommendation, Digital agriculture, Crop selection, Neural network

1. Introduction. Agriculture stands as the foundation of our society, fulfilling the nutritional requirements of billions worldwide. With the relentless growth of the world's population, the assurance of a consistent and reliable food source has become paramount. The significant role of agriculture is highlighted by its substantial contribution to India's GDP, amounting to 18.3% during the fiscal year 2022-2023 [1]. Additionally, this sector serves as the source of livelihood for approximately 50% of the country's workforce [2]. Despite its importance, the agricultural industry has experienced a decline in its performance in recent times. According to the Food and Agriculture Organization of the United Nations (FAO), approximately one-third of all food produced for human consumption across the globe is lost or wasted each year [3]. Insufficient land holdings, soil depletion, improper crop choices, inadequate fertilization, climate variations, and plant disease influence these losses.

Emerging technologies like Machine Learning (ML) [4], [5], Deep Learning (DL) [6], [7], and the Internet of Things (IoT) [8] have proven highly advantageous for the agricultural sector. These advancements offer the potential to enhance productivity and simultaneously ensure ecological sustainability by fortifying conventional farming practices. However, many small and marginal farmers still employ traditional or customary agricultural methods. For instance, these farmers often rely on their rudimentary knowledge to select crops and suitable fertilizers, usually favoring traditional or popular crops within their locality. Consequently, this reliance on traditional methods can compromise crop yields and soil fertility [11] due to insufficient scientific insights [15]. An adverse consequence of such practices is increased soil acidity, inappropriate selection and crop-specific fertilizers, and inadequate soil nutrient management. Additionally, the quality and productivity of crops are influenced by environmental conditions, climate variations, soil attributes, and water levels, further underlining the complexity of agricultural outcomes.

*Computer Science and Engineering Department, National Institute of Technology Patna, Bihar, India. (amitb.phd19.cs@nitp.ac.in).

†Computer Science and Engineering Department, National Institute of Technology Patna, Bihar, India. (prabhat@nitp.ac.in).

Selecting suitable crops and appropriate fertilizers tailored to filtered crops is pivotal in augmenting agricultural output and enhancing quality. Driven by the challenges above, this study aims to identify and tackle the intricacies of crop production by proposing a machine learning enabled Crop Selection and Fertilizer Recommendation (ML-CSFR) framework. This framework is designed to navigate the complex decision-making process of crop selection and appropriate fertilizer recommendation. It achieves this by considering various influential factors, including temperature, humidity, pH, rainfall, and soil nutrient levels. Notably, soil nutrients like Nitrogen (N), Phosphorus (P), and Potassium (K) are paramount for promoting plant growth and enhancing yield [16]. Concurrently, pH governs chemical reactions within the soil by determining its acidic or alkaline nature. Moreover, the development of plants is significantly influenced by electrical conductivity (EC), which also indicates soil fertility, water quality, and salinity.

The government has taken initiatives to enhance agricultural productivity by issuing Soil Health Cards (SHC) to individual farmers after assessing their farm's soil composition. These cards provide details about the soil's macro and micro nutrient levels. However, the conventional farming practices followed by farmers often fail to capitalize on this valuable information, resulting in suboptimal agricultural productivity. The ML-CSFR framework is proposed as an accessible and cost-effective solution to address these challenges. This framework suggests crops and corresponding fertilizers based on localized parameters by leveraging machine learning techniques. The ML-CSFR framework encompasses two distinctive phases: crop filtration and fertilizer recommendation. In the initial phase, crops are filtered based on local and regional variables. This filtration process involves the selection of 'n' crops that align with the soil and weather conditions while excluding those less suitable. Subsequently, tailored fertilizer recommendations are provided for each filtered crop in the second phase, guided by the soil's specific parameters. This research offers several noteworthy contributions:

1. The study introduces a two-phase Machine Learning-based Crop Selection and Fertilizer Recommendation (ML-CSFR) framework designed to provide farmers with optimal crop selection and fertilizer recommendations to enhance their returns. The initial phase of this framework involves assisting in selecting appropriate crops by filtering them based on the specific soil nutrient levels and the prevailing regional weather conditions associated with each farmland. Subsequently, the second phase generates fertilizer predictions for each filtered crop, considering the localized soil parameters.
2. Extensive experiments are conducted to demonstrate the efficacy of the proposed framework.
3. The first phase is evaluated on six benchmark models, including Decision Tree (DT), Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbour (KNN), XGBoost (XGB), and Artificial Neural Networks (ANN), for crop filtration. The evaluation results highlight that the proposed Artificial Neural Network (ANN) model attains the highest accuracy scores of 99.07% for validation and 99.10% for testing, while 99.13% precision and 99.24% recall, surpassing the performance of all other studied models.
4. The second phase evaluates eight benchmark models for fertilizer prediction: Logistic Regression (LR), DT, Gaussian Naive Bayes (GaNb), LrSVM, XGB, and RF. The evaluation results emphasize that the proposed XGB model excels, achieving outstanding accuracy scores of 99.23% for training and 97.66% for testing.
5. This empirical analysis validates that the proposed crop filtration in the first phase aligns with improved crop intensity and productivity. Furthermore, the framework's fertilizer recommendations in the second phase undergo validation through credible sources referenced in citations, thereby justifying the credibility and reliability of the framework's output.

The subsequent sections of this paper are organized as follows: In Section 2, a comprehensive overview of related works is provided. Section 3 outlines the methodology for the proposed framework. Section 4 discusses the experimental setup, data analysis, and methods used. The results and discussion are presented in Section 5. Lastly, Section 6 concludes the findings and discusses potential avenues for future research.

2. Related works. Numerous researchers have made contributions to devising diverse solutions to enhance crop productivity. Machine learning [10] and Deep learning have made their way into agriculture to enable more innovative and accurate crop selection decisions and improve yield through precise fertilizer recommendations [12]. Automated systems equipped with IoT [27] are also facilitating farmers in monitoring processes and receiving alerts, including weather conditions and soil moisture updates [17].

Agriculture has seen notable advancements in its pursuit of real-time crop selection [36] and fertilizer prediction [13], [14]. Cheema et al. [18] introduced a diverse crop model leveraging multiple soil parameters for

identifying suitable crops. The proposed approach harnessed a Quantum Value-based Gravitational Search Algorithm (GSA) to discover optimal solutions. Soil attributes such as pH, salinity, texture, nitrogen, phosphorous, and potassium were explored and employed as inputs for crop selection. Bakhavatchalam et al. [19] presented a crop prediction system that relies on various attributes and utilizes a multilayer perceptron (MLP), JRip, and decision table classifier. A range of machine learning models was implemented on the WEKA platform, with the most effective MLP model achieving an accuracy of 98.22%. Jain et al. [20] devised a soil-based machine learning comparative analytical framework for predicting crop yield production. This framework employed soil characteristics and climatic factors to categorize crop yield predictions as high, low, or medium. Gupta et al. [21] introduced a crop recommendation system integrating MapReduce and K-means clustering. This model considered crop yields per acre for distinct regions based on different varieties cultivated in the target zone. Mariammal et al. [22] proposed a feature selection method called Modified Recursive Feature Elimination (MRFE) for crop prediction, aiming to select vital features from crop data. This method employed a ranking algorithm to identify significant attributes. The results highlighted that MRFE outperformed various wrapper-based feature selection techniques, achieving an accuracy of 95%.

Senapaty et al. [23] introduced an IoT-enabled soil nutrient classification and crop recommendation (IoTSNA-CR) model to assist farmers throughout the cultivation process. The model employs sensors to gather real-time data on soil moisture, temperature, water, and nutrient levels. This work is comparable to Bhola et al. [25], where the model helps in optimal crop selection and reduces fertilizer usage. The innovative hybrid algorithm, MSVM-DAG-FFO, which combines a multi-class SVM with directed acyclic graph optimization and fruit fly optimization, achieved a remarkable 97.3% accuracy, surpassing the SVM and Decision Tree performance. Swaminathan et al. [24] address the challenge of fertilizer recommendations based on soil nutrients through a nutrient-centered deep collaborative filtering approach. This research aims to create an advanced recommender system called the Nutrient-centered Deep Collaborative Filtering (NDCF) method. The proposed method achieved root mean square and mean absolute error of 0.8411 and 0.655 on the collected dataset, respectively. The study by Khan et al. [26] presents a real-time context-aware fertilizer recommendation system utilizing ML and IoT technology. The model suggests suitable fertilizers for specific soil and crop types by capturing real-time soil fertility context through IoT-assisted mapping. The proposed IoT-assisted fertility mapping aligns well with standard soil chemical analysis, demonstrating mean differences of 0.34, 0.36, and -0.13 for Nitrogen, Phosphorous, and Potassium, respectively. Machine learning models are employed for context-aware fertilizer recommendation, including Logistic Regression, Support Vector Machine, Gaussian Naïve Bayes, and K-Nearest Neighbor. The Gaussian Naïve Bayes model exhibited the highest accuracy, reaching 96% and 94% for training and testing datasets.

Further, integrating modern technologies like the IoT and Artificial Intelligence (AI) in agriculture is crucial for efficient and quality crop production. Swaminathan et al. [27] introduce a four-layer architectural model encompassing sensors, networks, services, and applications to establish an energy-efficient farming system. The application layer employs deep learning for a fertilizer recommendation system aligned with expert opinions. The entire system is accessible through a user-friendly mobile application for farmers. This work underscores the significance of IoT-driven agricultural sensors and AI applications to enhance crop yield and sustainability, addressing the increasing demands of a growing population. The proposed approach showcases its potential to improve crop yield through fertilizer recommendations based on chemical properties. It includes integrating more sensors for comprehensive farm management. Thorat et al. [34] focus on integrating AI and sensor technology in agriculture to enhance insecticide, fertilizer recommendation, and soil nutrient analysis. The proposed approach employs Transition Probability Function (TPF) and Convolutional Neural Network (CNN), achieving over 90% accuracy. Indeed, the suitability of a particular soil type for various crops is significant, but unfavorable crop selection in a given location can negatively impact crop yield. This study addresses the identified limitations by introducing suitable crop selection and crop-specific fertilizer prediction architecture.

3. Methodology. Farmers are grappling with reduced crop yields and profits due to a limited understanding of crop selection intricacies and the factors influencing crop growth. Given that crop selection and the corresponding fertilizer stand as pivotal factors for maximizing yield and profitability, the primary objective of this study is to devise a Crop selection and Fertilizer recommendation framework to enhance agricultural returns.

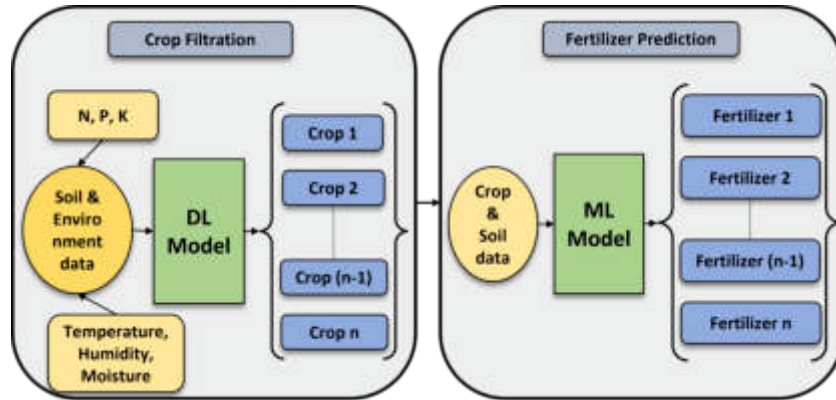


Fig. 3.1: Proposed framework

Consider a set of ‘ i ’ distinct crops denoted as $\{C_1, C_2, \dots, C_i\}$ and a collection of ‘ j ’ diverse farmlands denoted as $\{F_1, F_2, \dots, F_j\}$. It is assumed that each farmer possesses a Soil Health Card (SHC) for their farmland ‘ F_j ’, containing information about soil nutrient levels, alongside regular meteorological updates provided by government agencies. The objective is to determine appropriate crops and their corresponding fertilizers for each farmland based on inputs related to soil and weather conditions. The proposed framework, as depicted in Figure 3.1, employs a two-phase approach to suggest a variety of crops and associated fertilizers for each farm. In the initial phase, ‘ n ’ crops are filtered for each farmland ‘ F_j ’ from the pool of ‘ i ’ different crops. This phase involves assessing the compatibility of available crops with the local soil and weather conditions and filtering suitable crops. Further, the filtered crops are directed to the next stage, which identifies the suitable fertilizer for each crop for the farmer’s land. Implementing fertilizer tailored to the specific site offers benefits in terms of environmental sustainability, economic efficiency, and enhanced yield [35]. Each of these phases is further elaborated in the following subsections.

3.1. Crop filtration. Figure 3.2 illustrates the initial phase that focuses on filtering ‘ n ’ suitable crops. For every farmland ‘ F_j ’, consider $\{W_1(t), W_2(t), \dots, W_k(t)\}$ as weather conditions like temperature, rainfall, etc., at a time ‘ t ’, and $\{S_1(t), S_2(t), \dots, S_l(t)\}$ as soil attributes such as N, P, K, etc. Meteorological departments or local government agencies provide regular weather updates ‘ $W_k(t)$ ’ to farmers, aiding them in making informed decisions for their farming activities. Furthermore, the government provides a soil health card containing 12 vital soil macro and micro nutrients ‘ $S_l(t)$ ’, including pH, EC, Organic Carbon (OC), Nitrogen (N), Phosphorus (P), Potassium (K), Sulphur (S), Zinc (Zn), Boron (B), Iron (Fe), Manganese (Mn), and Copper (Cu). Since crop growth is intrinsically influenced by weather and soil conditions, these critical soil parameters are retrieved from the farmer’s soil health card and government weather updates to determine the most suitable crops. A proposed deep learning model computes probabilities $\{p_1, p_2, \dots, p_x\}$ based on these input parameters to rank crops. Subsequently, the top ‘ n ’ crops are selected and passed on to the next phase for further estimation.

Figure 3.4 portrays the architecture of the proposed artificial neural network model employed in the first phase. In this feed-forward backpropagation network, elements like weights, biases, the number of hidden layers, hidden neurons, learning rate, and training epochs play a pivotal role in determining prediction accuracy. These parameters are selected for precise predictions through a trial-and-error approach. A total of 7 inputs are fed into the input layer, and along with the bias, they are propagated to the hidden layer. The activation function ReLU is implemented for the hidden layers, while the output layers utilize the softmax activation function to predict probabilities. Furthermore, each hidden layer encompasses 512 neurons, the input layer consists of 7 neurons, and the output layer comprises 22 neurons, corresponding to each crop.

3.2. Fertilizer prediction. The second phase of the framework involves predicting fertilizer for each of the ‘ n ’ filtered crops obtained from the first phase. Figure 3.3 depicts the second phase of the ML-CSFR framework that predicts fertilizer using an ML model for each filtered crop individually on the available farmer’s

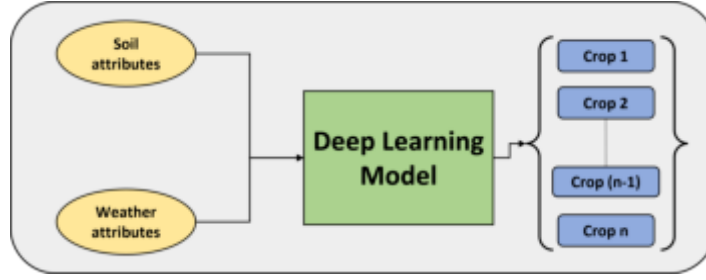


Fig. 3.2: Crop Filtration phase

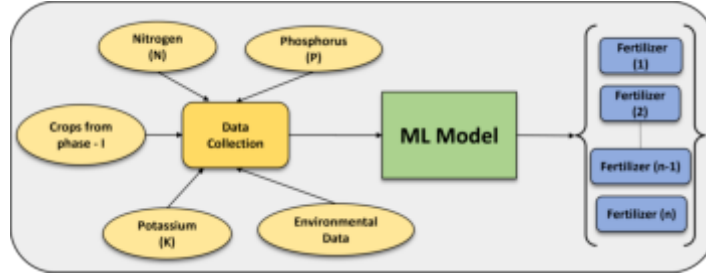


Fig. 3.3: Filter prediction phase

land. For every farmland ‘ F_j ’, filtered crops $\{C_1, C_2, \dots, C_i\}$, weather conditions $\{W_1(t), W_2(t), \dots, W_k(t)\}$, soil attributes $\{S_1(t), S_2(t), \dots, S_l(t)\}$, and $\{Fr_1, Fr_2, \dots, Fr_m\}$ be the available fertilizers.

The goal of the model is to recommend the appropriate fertilizer (y) for each filtered crop from the input set ‘ X ’, as described by Equation 3.1. The soil fertility level is determined by the quantities of nitrogen, phosphorus, and potassium in the soil. Equation 3.2 and Equation 3.3 represent the input and output feature set.

$$y = f(X) \tag{3.1}$$

$$X = \{W_k(t), S_l(t), C_m(t)\} \tag{3.2}$$

$$Y = F_n(t) \tag{3.3}$$

where, ‘ X ’ is the variables set representing weather, soil, and crop parameters; ‘ Y ’ represents the collection of commercially available fertilizers utilized as the model’s output for every combination of input variables. Each fertilizer available in the market possesses distinct compositions of essential nutrients like Nitrogen (N), Phosphorus (P), and Potassium (K), typically indicated in their names. The model inputs filtered crops, weather, and soil parameters and predicts suitable fertilizers for the land. Various models, including LR, DT, GaNB, LrSVM, XGB, and RF, are employed and compared to determine the most effective classification model for this phase.

3.3. Machine Learning models. Various machine learning models were employed and compared to determine the most effective classification model for fertilizer prediction. The models used include Logistic Regression, Decision Tree, Gaussian Naive Bayes, Support Vector Machine, Random Forest, Gradient Boosting, and XGBoost. Below is a detailed description of the setup for each model:

3.3.1. Logistic Regression. Logistic Regression is a linear model for binary classification. It models the probability that a given input point belongs to a particular class. The probability is modeled using a logistic

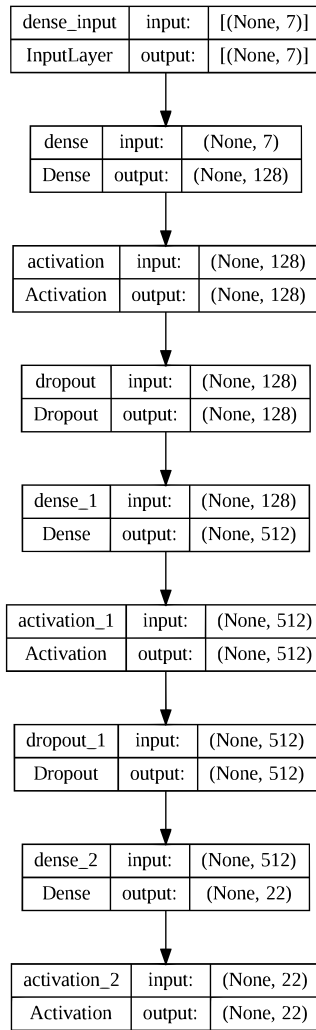


Fig. 3.4: Proposed ANN Architecture for Crop Filtration phase

function:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p)}} \tag{3.4}$$

where $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$ are the parameters to be estimated.

3.3.2. Decision Tree. Decision Trees are non-parametric models that split the data based on feature values. Each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. The Gini impurity or information gain is often used as a criterion to split the nodes:

$$Gini(D) = 1 - \sum_{i=1}^C P_i^2 \tag{3.5}$$

where P_i is the probability of an element being classified to a particular class.

3.3.3. Gaussian Naive Bayes. Gaussian Naive Bayes is based on Bayes' theorem and assumes independence between features. For continuous features, it assumes a Gaussian distribution:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}} \quad (3.6)$$

where μ_y and σ_y^2 are the mean and variance of the feature x_i for class y .

3.3.4. Support Vector Machine. SVMs find the hyperplane that best separates the classes by maximizing the margin between the nearest points of the classes (support vectors). For the linear case, the decision function is:

$$f(x) = \beta_0 + \sum_{i=1}^n \beta_i x_i \quad (3.7)$$

with the optimization objective:

$$\min_{\beta} \frac{1}{2} \|\beta\|^2 \quad \text{subject to} \quad y_i(\beta_0 + \sum_{j=1}^n \beta_j x_{ij}) \geq 1 \quad \forall i \quad (3.8)$$

3.3.5. Random Forest. Random Forest is an ensemble method that constructs multiple decision trees during training and outputs the mode of the classes (classification) of the individual trees. The model's prediction is the average prediction of the individual trees. The construction process involves:

$$\hat{f}(x) = \frac{1}{B} \sum_{b=1}^B f_b(x) \quad (3.9)$$

where B is the number of trees and $f_b(x)$ is the prediction of the b_{th} tree.

3.3.6. Gradient Boosting. Gradient Boosting builds trees sequentially, with each new tree aiming to correct the errors of the previous one. The key idea is to fit a new model to the residual errors made by the previous model:

$$F_m(x) = F_{m-1}(x) + \lambda h_m(x) \quad (3.10)$$

where $F_m(x)$ is the ensemble model at stage m , $h_m(x)$ is the new tree, and λ is the learning rate.

3.3.7. XGBoost. XGBoost is an advanced implementation of gradient boosting with additional regularization terms to control overfitting. The objective function is:

$$L(\phi) = \sum_{i=1}^n l(\hat{y}_i, y_i) + \sum_{k=1}^K \Omega(f_k) \quad (3.11)$$

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \|\omega\|^2 \quad (3.12)$$

where $\Omega(f_k)$ is a regularization term, T is the number of leaves, w is the leaf weight, and γ and λ are regularization parameters.

4. Experiment. This section provides a comprehensive evaluation of the proposed architecture through empirical analysis. It begins by outlining the experimental setup, detailing the dataset source, and discussing data analysis methods. Subsequently, the execution of the suggested model is discussed, followed by a comparative assessment of the attained outcomes compared to alternative machine learning models.

Table 4.1: Feature description (Crop filtration phase)

Features	Description	Unit
Nitrogen (N)	It is responsible for photosynthesis in the plant.	kg/ha
Phosphorus (P)	It is crucial to the crop's development.	kg/ha
Potassium (K)	It is required for the reproduction of crops.	kg/ha
pH level (pH)	It determines the availability of essential plant nutrients.	pH value
Temperature	Temperature is a key factor in plant growth and development.	degree Celsius
Humidity	Humidity is important for photosynthesis in plants.	%
Rainfall	The primary source of water for agricultural production.	mm

Table 4.2: Feature description (Fertilizer prediction phase)

Feature(s)	Description
Temperature	Degree Celsius
Humidity	%
Moisture	%
Soil type	5 types
Crop type	11 types
Nitrogen	Ratio
Potassium	Ratio
Phosphorous	Ratio
Fertilizer	7 types

4.1. Experimental setup. The experimental setup involves an Intel Core i5 processor with a quad-core x64-based architecture running at 3.6 GHz and 8 GB of RAM. The programming language used is Python, and the program was executed using the Google Colab notebook. Various standard software libraries such as Keras, Tensorflow, Matplotlib, and Numpy were utilized for implementation.

4.2. Dataset. Two distinct datasets are employed to assess the efficacy of the proposed ML-CSFR framework. The initial crop filtration phase dataset is sourced from [28]. This dataset categorizes lands and crops according to various attributes, encompassing soil characteristics such as nitrogen, phosphorus, potassium, and pH, and environmental factors impacting crop growth, such as humidity and rainfall. The collected dataset comprises 2200 land samples and encompasses 22 distinct crop types. Each crop category includes 100 individual land samples for analysis. The data has been divided into the ratio of 80:10:10 for training, validation, and testing sets. Table 4.1 describes the dataset features used in the first phase of the framework.

The second phase dataset [29] contains seven fertilizer varieties and eleven crops. The attributes of the collected dataset for the fertilizer dataset are shown in Table 4.2. The fertilizer used in executing the proposed approach is represented in Equation 4.1.

$$Y = \{Fr_1, Fr_2, \dots, Fr_m\} \quad (4.1)$$

The fertilizer names are used as ' Fr_1 ' for '20-20', ' Fr_2 ' for '14-35-14', ' Fr_3 ' for '26-28' fertilizer, ' Fr_4 ' for 'DAP', ' Fr_5 ' for '10-26-26', ' Fr_6 ' for 'Urea', and ' Fr_7 ' for '17-17-17'.

4.3. Dataset analysis. This section delves into the analysis of the soil and environmental data that influence both the crop filtration and fertilizer prediction processes. Initially, each dataset undergoes data cleaning procedures, as various attributes encompass distinct measurement scales. The Min-Max Scaler uses

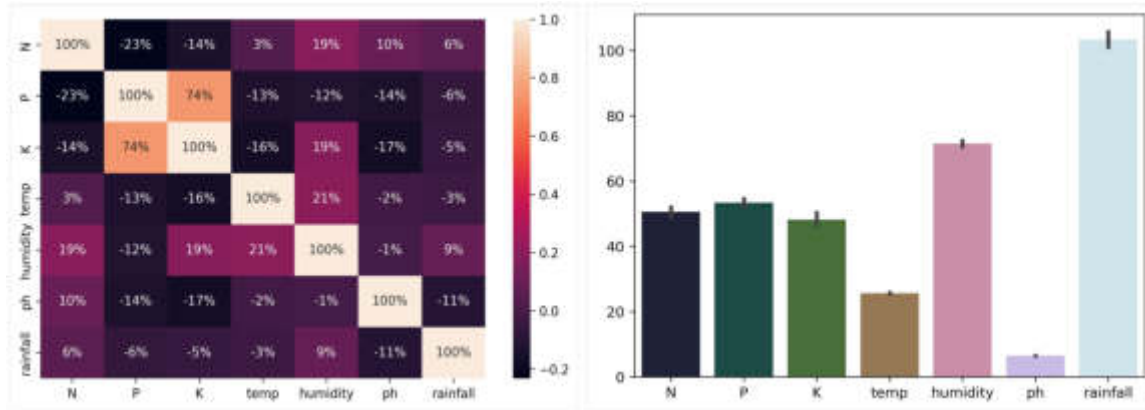


Fig. 4.1: Phase-1 dataset (a) Correlation matrix (b) Feature importance

the rescaled formula mentioned in Equation 4.2 for precise crop selection.

$$C = \frac{B - B_{min}}{B_{max} - B_{min}} \quad (4.2)$$

where C signifies the rescaled value, B is the feature value, B_{min} is the minimum value, and B_{max} is the maximum value. Further, the data has been apportioned into training and test sets, with an 80:10:10 split. The accuracy of these models is evaluated on both the training, validation, and test datasets.

The primary macronutrients, nitrogen, phosphorus, and potassium (N, P, and K), are pivotal in enhancing crop yield and quality. Figure 4.1(a) illustrates the correlation between the employed features, emphasizing the high correlation between potassium and phosphorous soil parameters and a moderate correlation between humidity and rainfall. On the other hand, Figure 4.1(b) provides insight into the crucial features within the crop filtration dataset, highlighting that among all the weather parameters, rainfall and humidity emerge as essential factors.

Figure 4.2 compares the nitrogen, phosphorus, and potassium values that different crops need. Among the crops, cotton, apple, and grapes exhibit the highest demand for macronutrients for optimal growth, whereas lentils, black gram, and oranges have the lowest requirements. The significance of various soil macronutrients such as N, P, and K holds a relatively uniform weightage across all crops. Notably, rainfall has the most significant importance among the considered parameters, while pH records are the least effective.

The dataset used for the study's second phase is sourced from [29]. The various commercial fertilizers used in the dataset and the distribution are classified in Figure 4.3. Many fertilizer products are labeled with the abbreviation NPK, followed by numerical values, such as NPK 10-26-26. The numbers following NPK indicate the percentage amounts of each nutrient in the fertilizer. For instance, an NPK value of 10-26-26 indicates that the fertilizer contains 10% nitrogen, 26% phosphorus, and 26% potassium [30]. The selected fertilizer and crop data type distribution in the dataset is visualized in Figure 4.4(a) and Figure 4.4(b), respectively.

Each type of soil has its unique attributes and composition. For instance, Urea is the most widely used solid nitrogen fertilizer and is usually applied as granules. Hence, its count is maximum in the dataset, as seen in Figure 4.4(a). Further, the crop has specific nutrient requirements for successful growth and enhanced production. Some crops necessitate lower nutrient levels, while others demand more. Consequently, in addition to soil type, crop type significantly influences fertilizer needs and recommendations.

The dataset distribution of the three macronutrients is categorized based on the provided classification. The application of fertilizer heavily relies on factors such as crop type, soil, and soil fertility regarding NPK nutrients. The soil nutrient level is interconnected with crops and fertilizers needed. Moreover, the relationships between crop type, soil type, and existing nutrient levels are even more intricate.

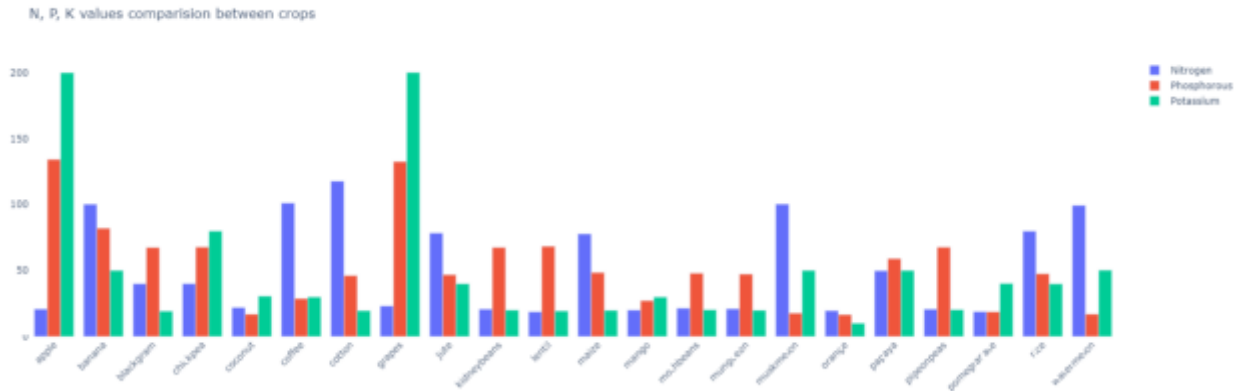


Fig. 4.2: N, P, K values required by different crops

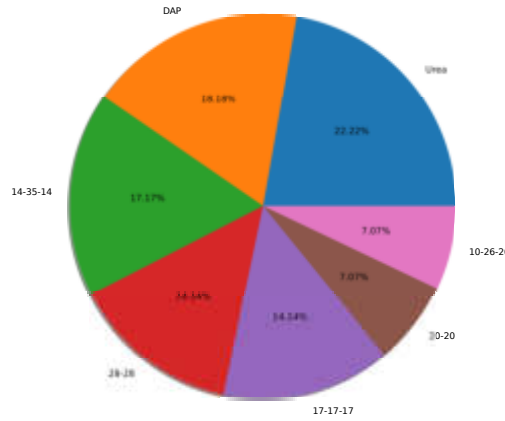


Fig. 4.3: Fertilizer class distribution

4.4. Algorithm for ML-CSFR framework. The primary goal of this experiment is to develop a recommendation model that will advise crop filtration and crop-specific fertilizer on various factors such as soil constituents, crop traits, and climate. Algorithm 1 presents the algorithm with the detailed steps involved in crop-based fertilizer prediction using ANN-XGB.

The algorithm is divided into two parts: (1) compute each crop’s rank and filter the top-n crops; (2) predict fertilizer for each crop corresponding to the farmer’s land. It requires soil health card details and environmental values concerning each land as input.

4.5. Evaluation Metrics. This study involves two separate datasets corresponding to the two phases of the framework. In the initial phase, the framework filters crops from multiple crop categories, while in the second phase, it predicts fertilizers from a range of different classes. As a result, a multi-class classification approach is utilized, and a confusion matrix is generated to calculate classification instances, including True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). In the context of multi-class classification, these instances can be interpreted as follows:

- True Positive (TP): The model correctly predicts the positive class.
- False Positive (FP): The model incorrectly predicts the positive class.
- True Negative (TN): The model correctly predicts the negative class.
- False Negative (FN): The model incorrectly predicts the negative class.

Algorithm 1 ML-CSFR: Crop Selection and Fertilizer Recommendation**Require:** Local input variables $X = \{x_1, x_2, \dots, x_n\}$, Soil parameters S , Weather parameters W **Ensure:** Top 3 recommended crops $C = \{c_1, c_2, c_3\}$, Recommended fertilizers F

```

1: Phase 1: Crop Selection using ANN
2: Initialize ANN parameters: number of layers  $L$ , neurons in each layer, activation functions
3: for epoch = 1 to  $N$  do
4:   for each batch in data do
5:     Forward Propagation:
6:     for layer  $l = 1$  to  $L$  do
7:        $z^{(l)} = W^{(l)} a^{(l-1)} + b^{(l)}$ 
8:        $a^{(l)} = \sigma(z^{(l)})$ 
9:     end for
10:    Output Layer:
11:     $\hat{y}_i = \frac{e^{a_i^{(L)}}}{\sum_{j=1}^m e^{a_j^{(L)}}}$ 
12:    Loss Function:
13:     $\mathcal{L} = -\sum_{i=1}^m y_i \log(\hat{y}_i)$ 
14:    Backward Propagation:
15:    Compute gradients:  $\frac{\partial \mathcal{L}}{\partial W^{(l)}}$  and  $\frac{\partial \mathcal{L}}{\partial b^{(l)}}$ 
16:    Update weights and biases:
17:     $W^{(l)} := W^{(l)} - \eta \frac{\partial \mathcal{L}}{\partial W^{(l)}}$ 
18:     $b^{(l)} := b^{(l)} - \eta \frac{\partial \mathcal{L}}{\partial b^{(l)}}$ 
19:    end for
20:  end for
21: Output top 3 recommended crops  $C = \{c_1, c_2, c_3\}$  based on highest probabilities  $\hat{y}_i$ 
22: Phase 2: Fertilizer Recommendation using XGBoost
23: Initialize XGBoost parameters: number of trees  $T$ , learning rate  $\eta$ , maximum depth  $d$ 
24: for  $t = 1$  to  $T$  do
25:   Compute predictions  $\hat{y}_t = \sum_{i=1}^t f_i(X)$ 
26:   Compute residuals  $r_t = y - \hat{y}_t$ 
27:   Fit regression tree to residuals:  $f_t = \arg \min_f \sum_{i=1}^n L(y_i, \hat{y}_{t-1} + f(x_i))$ 
28:   Update model:  $\hat{y}_t := \hat{y}_{t-1} + \eta f_t(X)$ 
29: end for
30: For each crop  $c \in C$ , predict the suitable fertilizer  $F$  using XGBoost
31: Output: Recommended crops  $C$  and fertilizers  $F$ 

```

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (4.3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4.4)$$

$$Recall = \frac{TP}{TP + FN} \quad (4.5)$$

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4.6)$$

Equation 4.3 represents accuracy, calculating the ratio of correctly predicted observations to total observations. Equation 4.4 defines precision, which measures the ratio of the true positive predictions to the total number of positive predictions. In contrast, Equation 4.5 represents recall, which determines the number of true positive predictions divided by the total number of relevant observations. The F_1 Score, as calculated in Equation 4.6, represents the harmonic mean of precision and recall.

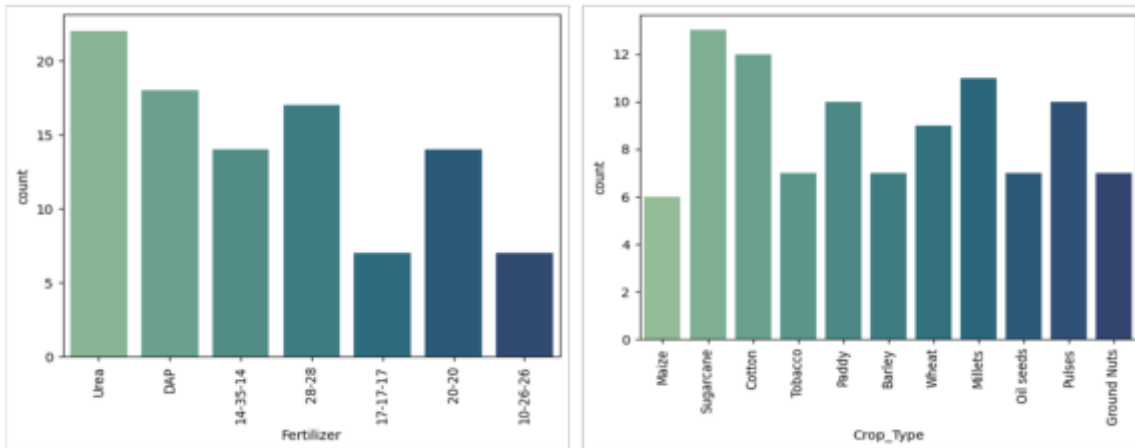


Fig. 4.4: Fertilizer data distribution (a) Fertilizer type (b) Crop type

Table 5.1: Comparative analysis (Crop filtration phase)

Accuracy / Models	DT	SVM	RF	KNN	XGBoost	ANN
Training Accuracy	0.985	0.985	0.992	0.988	0.992	0.992
Testing Accuracy	0.974	0.976	0.985	0.976	0.981	0.991
Validation Accuracy	0.972	0.975	0.983	0.975	0.980	0.990
Precision	0.972	0.975	0.983	0.975	0.981	0.991
Recall	0.974	0.976	0.985	0.976	0.982	0.992

Table 5.2: List of filtered crops for sample lands

Land	Top filtered crops
Land 1	Rice, Maize, Cotton
Land 2	Maize, Rice, Jute
Land 3	Maize, Chickpea, Lentil

5. Results. A series of comparative experiments were conducted to evaluate the performance of the proposed framework for crop filtration and crop-specific fertilizer recommendation. The results of these experiments are comprehensively analyzed and discussed in this section.

5.1. Result Analysis of Crop Filtration Phase. The crop filtration phase utilizes a classification model to filter crops based on their probabilities. The obtained results for the initial phase under different environmental conditions are summarized in Table 5.1.

The evaluation outcomes demonstrate that the proposed Artificial Neural Network (ANN) model achieves a validation accuracy of 99.07%, testing accuracy of 99.10%, precision of 99.13%, and recall of 99.24%. These results are the highest among all the studied models. Additionally, the Random Forest (RF) and XGBoost (XGB) models also performed well, with testing accuracies, precision, and recall all above 98%. Conversely, the Decision Tree (DT) recorded the lowest validation and testing accuracies of 97.20% and 98.50%, respectively. This lower accuracy could be attributed to variations in the dataset, as DT is very sensitive to small perturbations in the data.

Table 5.2 provides an overview of the crops the ANN model filtered for three distinct farmlands. While the present study filters only three crops, the number of crops can be customized according to the user's preferences.

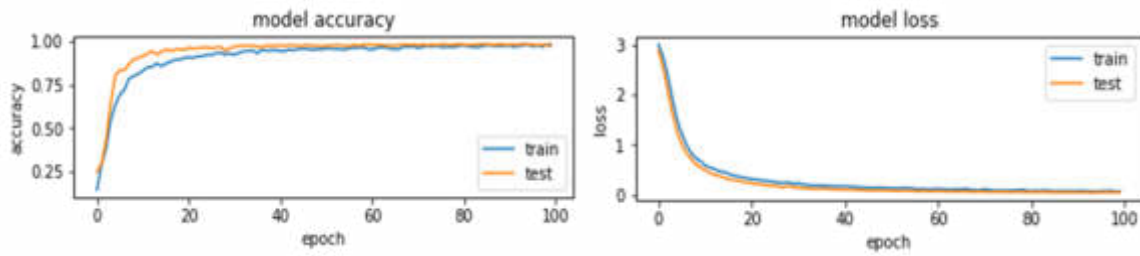


Fig. 5.1: ANN result (a) accuracy vs epoch (b) loss vs epoch

Table 5.3: Performance Comparison (Fertilizer Prediction Phase)

Model	Training Accuracy	Testing Accuracy	Precision	Recall	F1-Score	AUC
LR	0.92	0.70	0.71	0.70	0.70	0.75
DT	0.99	0.94	0.94	0.94	0.94	0.96
GaNB	0.96	0.50	0.52	0.50	0.49	0.55
LrSVM	0.94	0.85	0.86	0.85	0.85	0.88
XGB	0.99	0.97	0.97	0.97	0.97	0.98
RF	0.99	0.95	0.95	0.95	0.95	0.97

Figures 5.1(a) and (b) illustrate the graphs depicting accuracy versus epoch and loss versus epoch for the proposed Artificial Neural Network (ANN) model. Examining the loss curve reveals that the global optimal minimum is attained in the initial stages of iterations. The results justify the selection of ANN as the proposed model for crop filtration as its performance consistently improves and error reduces over time.

5.2. Result Analysis of Fertilizer Prediction Phase. In the fertilizer prediction phase, the performance of various machine learning models is evaluated, and the most effective model is selected for this phase. The most appropriate model is chosen by considering the performance metrics such as accuracy, precision, recall, and F1-score. The models, including LR, DT, GaNB, LrSVM, XGB, and RF, were applied to the preprocessed dataset for fertilizer prediction. The performance assessment of these models is detailed in Table 5.3.

The analysis revealed that tree-based models like RF, XGB, and DT exhibited more exceptional performance stability than other models. This stability arises from the fact that these models establish decision boundaries that are more consistent and accurate by aggregating the outcomes of multiple trees and utilizing majority voting to arrive at a precise prediction.

Notably, the XGB performs exceptionally well in recommending suitable fertilizers, achieving an accuracy of 99.23% and 97.66% on the training and test datasets, respectively. The classification report and confusion matrix for the best-performing model XGB are shown in Table 5.4 and Figure 5.2, respectively. The table indicates that all the fertilizers achieved 1.00 precision except 17-17-17, which achieved a precision of 0.67. Further, fertilizer 10-26-26 achieves a low recall of 0.67, indicating that the proposed model misclassifies 33 out of 100 classes. The low recall value can be attributed to the imbalance dataset, as depicted in Figure 4.3, where class 10-26-26 only accounts for 7.07% of the data.

Conversely, the GaNB model is the worst-performing, primarily due to its unsuitability for dealing with imbalanced class problems. When dealing with imbalanced datasets, where certain classes have much larger instances than others, GaNB can encounter difficulties delivering accurate outcomes. This problem may lead to biased predictions favoring the majority class and suboptimal performance for the minority class.

In summary, XGB performed better than other models in testing accuracy, precision, recall, F1-score, and AUC values. Table 5.4 present the classification report of the XGB model. The table depict a consistent performance in model's accuracy, with a reduced error over time, thereby justifying the selection of XGB as

Table 5.4: Classification report (XGB)

Classes	Code	Precision (%)	Recall (%)	F1-Score
10-26-26	0	1.00	0.67	0.80
14-35-14	1	1.00	1.00	1.00
17-17-17	2	0.67	1.00	0.80
20-20	3	1.00	1.00	1.00
28-28	4	1.00	1.00	1.00
DAP	5	1.00	1.00	1.00
Urea	6	1.00	1.00	1.00
Accuracy				0.97

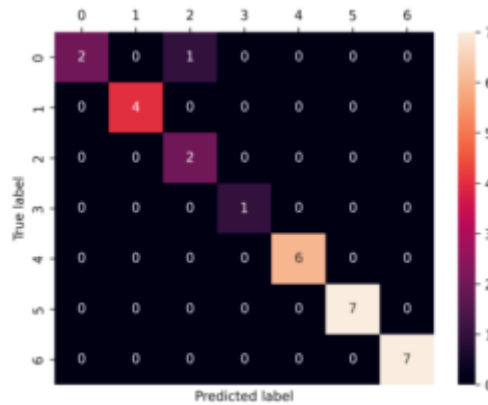


Fig. 5.2: Confusion matrix (XGB)

the proposed fertilizer recommendation model.

The fertilizer recommendations for each of the filtered crops for Land 1 are outlined in Table 5.5. The proposed model suggested DAP, 20-20, and 14-35-14 as the appropriate fertilizers for the predicted crops: Rice, Maize, and Cotton, respectively. These recommendations are substantiated by the references provided in the table. For example, the prediction of DAP or Urea fertilizer for the Rice crop aligns with established practices [31]. Moreover, the 28-28 fertilizer, a complex blend rich in Nitrogen and Phosphorus, is suitable for crops like Paddy, Cotton, Chillies, Sugarcane, and Vegetables, offering immediate and sustained greenness [32]. Similarly, the 14-35-14 fertilizer is optimal for Rice, Cotton, Groundnut, Chillies, Soya bean, and Potato crops, particularly those demanding high initial Phosphate [33]. Thus, the proposed framework is validated by its ability to suggest fitting crop-specific fertilizers, providing farmers with insights to enhance yields.

5.3. Discussion. The development of the proposed two-phase framework for crop filtration and fertilizer recommendation represents a significant advancement in precision agriculture. The experimental results underscore the framework's effectiveness, showcasing its ability to make accurate and reliable recommendations tailored to specific agricultural conditions.

In the crop filtration phase, the Artificial Neural Network (ANN) model demonstrated superior performance compared to other classification models, achieving a remarkable testing accuracy of 99.10%, precision of 99.13%, and recall of 99.24%. These metrics indicate the model's robustness in correctly identifying suitable crops under varying environmental conditions. The high precision and recall values further affirm the model's ability to make precise and consistent predictions, minimizing false positives and negatives.

The fertilizer prediction phase involved evaluating various machine learning models, where XGBoost (XGB) emerged as the best-performing model. It achieved training and testing accuracies of 99.23% and 97.66%, respec-

tively, along with a high F1-score and AUC, indicating its exceptional ability to distinguish between different classes of fertilizers. The classification report for XGB showed that most fertilizers achieved precision and recall of 1.00, except for fertilizer 17-17-17 and 10-26-26, which had a lower precision and recall. This discrepancy can be primarily attributed to class imbalance, as evidenced by the confusion matrix and distribution of the dataset.

Further, the framework iteratively follows a dynamic update mechanism, continuously updating its recommendations based on the latest soil and weather data. After each planting season, soil health is reassessed using an updated Soil Health Card or periodic soil testing. This updated soil data and ongoing weather information are fed back into the framework, which reruns the crop and fertilizer recommendation process. This dynamic feedback loop ensures that the recommendations remain relevant and accurate over time, adapting to changes in soil nutrient levels resulting from fertilizer applications and other environmental factors.

5.3.1. Error-Analysis. The error analysis from the first phase revealed that the Decision Tree (DT) model recorded the lowest validation and testing accuracies of 97.20% and 98.5%, respectively. This lower performance can be attributed to the model's sensitivity to small perturbations in the dataset, highlighting its limitations in handling imbalance data. Conversely, the Random Forest (RF) and XGBoost (XGB) models also performed well, with testing accuracies above 98%, but were slightly outperformed by the ANN model, suggesting that the latter's deeper architecture and non-linear learning capabilities offer significant advantages in this application.

Further, from the second phase error analysis is the model's occasional misclassification of minority class fertilizers, such as 10-26-26, which had a recall of 0.67. This misclassification suggests further data balancing techniques or augmenting the minority class samples to improve the model's performance on underrepresented classes.

5.3.2. Contributions. The significant contributions of this work include:

- *High Accuracy and Reliability:* The ANN model achieved exceptional accuracy in the crop filtration phase, and the XGB model demonstrated superior performance in the fertilizer prediction phase. These results highlight the models' effectiveness in handling complex agricultural data and making accurate recommendations.
- *Robustness to Environmental Variations:* The high precision and recall values across different models indicate the framework's robustness and adaptability to varying environmental conditions, ensuring reliable performance in diverse agricultural settings.
- *Scalability and Practicality:* The framework's architecture is designed to be scalable and capable of processing extensive agricultural data inputs in a cloud-based environment. This scalability ensures its applicability to small-scale farms and large agricultural enterprises.
- *Real-world Applicability:* The proposed framework's ability to provide farm-specific recommendations and its potential integration with real-time sensor data make it a practical tool for enhancing decision-making in agriculture. The accurate predictions of crop selection and fertilizer recommendations can lead to improved crop yields and optimized resource use.

Further, the proposed ML-CSFR framework is highly relevant to scalable computing due to its inherent ability to handle real-life problems. Using machine learning models such as ANN and XGBoost allows the framework to process extensive agricultural data inputs, including soil characteristics and weather conditions, in a scalable manner. This scalability ensures that the framework can be applied to diverse agricultural settings, from small-scale farms to large agricultural enterprises, that can be a versatile solution for improving crop selection and fertilizer recommendation processes. Additionally, the framework's architecture is designed to be deployable on cloud-based platforms, leveraging the power of distributed computing further to enhance its scalability and applicability in different agricultural scenarios.

Finally, the significant observation drawn from the results is the remarkable effectiveness of the proposed model in scenarios demanding farm-specific recommendations. These findings further confirm the practicality of the proposed framework, particularly in situations characterized by limited resources and constraints. Moreover, besides its simple architecture, the framework serves a dual role of crop selection and associated fertilizer recommendation, eliminating the requirement for separate applications for these tasks. This reduction in

Table 5.5: Recommended Fertilizer

Crop	Recommended Fertilizer	Ref.
Rice	DAP	[31]
Maize	28-28	[32]
Cotton	14-35-14	[33]

overhead enhances the model's efficiency and practicality. These findings underscore the adaptability of the proposed framework, positioning it as a valuable tool ready for implementation in real-world scenarios.

6. Conclusions and Future Works. The exponential growth of the world's population has boosted demand for both quantity and quality of food. Consequently, the agricultural sector is required to undergo a modern transformation to address this demand adequately. The integration of modern technologies with intelligent algorithms holds the potential to benefit the farming community significantly. This paper proposes ML-CSFR, a two-phase Crop Selection and Fertilizer Recommendation framework designed to provide better returns for the agricultural sector. The framework is a machine-learning tool that provides crop selection and fertilizer recommendations by leveraging local input variables. It harnesses various weather and soil data sources to deliver precise recommendations. The initial phase of crop filtration is executed using Artificial Neural Networks. The result justifies that ANN outperforms other ML models with an accuracy of 99.10%. In the second phase of fertilizer prediction, the results indicate that XGBoost is the most accurate machine learning model, achieving an accuracy of 99.23% for the training dataset and 97.66% for the test dataset. The proposed work contributes to increased accuracy in crop prediction, improved soil health with proper fertilization, and enhanced decision-making.

Additionally, potential areas for future exploration involve enlarging the datasets used and extending the application's utility to gain more comprehensive insights into crop and soil management. In terms of practical implications, the simple and lightweight design of the suggested framework offers potential for future integration with handheld devices. This integration could help farmers conveniently forecast crops and their corresponding fertilizers. Moreover, enhancing recommendations could involve integrating real-time sensor data collected from crop fields. This approach would enable precise determination of the exact quantities of macro and micro nutrients required, tailored to the specific demands of each crop.

REFERENCES

- [1] Agriculture & Farmers Welfare, M. Contribution of Agriculture Sector in GDP. (<https://www.pib.gov.in/PressReleasePage.aspx?PRID=1909213>), Accessed: 2023-08-03
- [2] Aayog, N. Workforce Changes and Employment. (https://www.niti.gov.in/sites/default/files/2023-02/Discussion_Paper_on_Workforce_05042022.pdf), Accessed: 2023-08-03
- [3] Food & United Nations, A. Food wastage: Key facts and figures. (<https://www.fao.org/news/story/en/item/196402/icode>), Accessed: 2023-08-03
- [4] Veeragandham, S. & Santhi, H. A review on the role of machine learning in agriculture. *Scalable Computing: Practice And Experience*. **21**, 583-589 (2020)
- [5] Bhola, A. & Kumar, P. Performance Evaluation of Different Machine Learning Models in Crop Selection. *Robotics, Control And Computer Vision*. pp. 207-217 (2023)
- [6] Verma, S., Kumar, P. & Singh, J. A Unified Lightweight CNN-based Model for Disease Detection and Identification in Corn, Rice, and Wheat. *IETE Journal Of Research*. pp. 1-12 (2023)
- [7] Bhola, A., Verma, S. & Kumar, P. A comparative analysis of deep learning models for cucumber disease classification using transfer learning. *Journal Of Current Science And Technology*. **13**, 23-35 (2023)
- [8] Sinha, A., Shrivastava, G. & Kumar, P. Architecting user-centric internet of things for smart agriculture. *Sustainable Computing: Informatics And Systems*. **23** pp. 88-102 (2019)
- [9] Swaminathan, B., Palani, S., Vairavasundaram, S., Kotecha, K., & Kumar, V. IoT-driven artificial intelligence technique for fertilizer recommendation model. *IEEE Consumer Electronics Magazine*, **12** pp. 109-117. (2022)
- [10] Janvier, NIYITEGEKA and Arcade, N & Eric, NGABOYERA & Jean, N. Machine Learning based Soil Fertility Prediction. *International Journal of Innovative Science, Engineering & Technology*, **8** pp. 141-146. (2021)

- [11] Sujatha, M and Jaidhar, CD Machine learning-based approaches to enhance the soil fertility—A review. *Expert Systems with Applications*. pp. 122557 (2023)
- [12] Bhattacharya, S., & Pandey, M. PCFRIMDS: Smart Next-Generation Approach for Precision Crop and Fertilizer Recommendations Using Integrated Multimodal Data Fusion for Sustainable Agriculture. *IEEE Transactions on Consumer Electronics*. (2024)
- [13] Wu, M., Xiong, J., Li, R., Dong, A., Lv, C., Sun, D., ... & Niu, W. Precision forecasting of fertilizer components' concentrations in mixed variable-rate fertigation through machine learning. *Agricultural Water Management*. **298** pp. 108859. (2024)
- [14] Reddy, G. V., Reddy, M. V. K., Spandana, K., Subbarayudu, Y., Albawi, A., Chandrashekar, R., ... & Praveen, P. Precision farming practices with data-driven analysis and machine learning-based crop and fertiliser recommendation system. *In E3S Web of Conferences*. **507**, pp. 01078. (2024)
- [15] Kang, M., Wang, X., Wang, H., Hua, J., Reffye, P. & Wang, F. The development of AgriVerse: Past, present, and future. *IEEE Transactions On Systems, Man, And Cybernetics: Systems*. (2023)
- [16] Modi, A., Sharma, P., Saraswat, D. & Mehta, R. Review of Crop Yield Estimation using Machine Learning and Deep Learning Techniques. *Scalable Computing: Practice And Experience*. **23**, 59-80 (2022)
- [17] Jani, K. & Chaubey, N. A novel model for optimization of resource utilization in smart agriculture system using IoT (SMAIoT). *IEEE Internet Of Things Journal*. **9**, 11275-11282 (2021)
- [18] Cheema, S., Singh, A. & Gritli, H. Optimal Crop Selection Using Gravitational Search Algorithm. *Mathematical Problems In Engineering*. **2021** (2021)
- [19] Bakthavatchalam, K., Karthik, B., Thiruvengadam, V., Muthal, S., Jose, D., Kotecha, K. & Varadarajan, V. IoT framework for measurement and precision agriculture: predicting the crop using machine learning algorithms. *Technologies*. **10**, 13 (2022)
- [20] Jain, K. & Choudhary, N. Comparative analysis of machine learning techniques for predicting production capability of crop yield. *International Journal Of System Assurance Engineering And Management*. **13**, 583-593 (2022)
- [21] Gupta, R., Sharma, A., Garg, O., Modi, K., Kasim, S., Baharum, Z., Mahdin, H. & Mostafa, S. WB-CPI: Weather based crop prediction in India using big data analytics. *IEEE Access*. **9** pp. 137869-137885 (2021)
- [22] Mariammal, G., Suruliandi, A., Raja, S. & Poongothai, E. Prediction of land suitability for crop cultivation based on soil and environmental characteristics using modified recursive feature elimination technique with various classifiers. *IEEE Transactions On Computational Social Systems*. **8**, 1132-1142 (2021)
- [23] Senapaty, M., Ray, A. & Padhy, N. IoT-Enabled Soil Nutrient Analysis and Crop Recommendation Model for Precision Agriculture. *Computers*. **12**, 61 (2023)
- [24] Swaminathan, B., Palani, S. & Vairavasundaram, S. Feature fusion based deep neural collaborative filtering model for fertilizer prediction. *Expert Systems With Applications*. **216** pp. 119441 (2023)
- [25] Bhola, A., & Kumar, P. Deep feature-support vector machine based hybrid model for multi-crop leaf disease identification in Corn, Rice, and Wheat. *Multimedia Tools and Applications*. pp. 1-21 (2024)
- [26] Khan, A., Faheem, M., Bashir, R., Wechtaison, C. & Abbas, M. Internet of things (IoT) assisted context aware fertilizer recommendation. *IEEE Access*. **10** pp. 129505-129519 (2022)
- [27] Swaminathan, B., Palani, S., Vairavasundaram, S., Kotecha, K. & Kumar, V. IoT-driven artificial intelligence technique for fertilizer recommendation model. *IEEE Consumer Electronics Magazine*. **12**, 109-117 (2022)
- [28] INGLE, A. Crop Recommendation Dataset. (<https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset>), Accessed: 2023-07-03
- [29] Abhishek, G. Fertilizer Prediction Dataset. (<https://www.kaggle.com/datasets/gdabhishek/fertilizer-prediction>), Accessed: 2023-07-03
- [30] University, T. Nutrient Management: Fertilizers. (https://agritech.tnau.ac.in/agriculture/agri_nutrientmgt_fertilizers.html), Accessed: 2023-08-03
- [31] Agriculture, I. Fertilizer Application for Rice. (https://agri.punjab.gov.in/sites/default/files/fertilizer_application_for_rice.pdf), Accessed: 2023-08-03
- [32] IndiaMart GROMOR 28-28-0. (https://www.indiamart.com/proddetail/gromor-28-28-0-fertilizer-10329114088.html?smp_widget=1), Accessed: 2023-08-03
- [33] IndiaMart GROMOR 14-35-14. (<https://www.indiamart.com/proddetail/gromor-14-35-14-10329096730.html>), Accessed: 2023-08-03
- [34] Thorat, T., Patle, B. & Kashyap, S. Intelligent insecticide and fertilizer recommendation system based on TPF-CNN for smart farming. *Smart Agricultural Technology*. **3** pp. 100114 (2023)
- [35] Sanches, G., Magalhães, P., Kolln, O., Otto, R., Rodrigues Jr, F., Cardoso, T., Chagas, M. & Franco, H. Agronomic, economic, and environmental assessment of site-specific fertilizer management of Brazilian sugarcane fields. *Geoderma Regional*. **24** pp. e00360 (2021)
- [36] Vandana, W. M., & Kavya, B. Soil Fertility Assessment and Crop Recommendation for Sustainable Farming using Machine Learning and Deep Learning. *In 2024 4th International Conference on Data Engineering and Communication Systems (ICDECS)* pp. 1-3. IEEE. (2024)

Edited by: Katarzyna Wasielewska-Michniewska

Regular paper

Received: Sep 7, 2023

Accepted: Jul 29, 2024



A NEW MULTI-ROBOTS SEARCH AND RESCUE STRATEGY BASED ON PENGUIN OPTIMIZATION ALGORITHM

OUARDA ZEDADRA ^{*}, AMINA ZEDADRA [†], ANTONIO GUERRIERI [‡], HAMID SERIDI [§] AND DOUAA GHELIS [¶]

Abstract. In response to the challenging conditions that arise after natural disasters, multi-robot systems are utilized as alternatives to humans for searching and rescuing victims. Exploring unknown environments is crucial in mobile robotics, serving as a foundational stage for applications such as search and rescue, cleaning tasks, and foraging. In our study, we introduced a novel search strategy for multi-robot search and rescue operations. This strategy draws inspiration from the hunting behavior of penguins and combines the Penguin Search Optimization Algorithm with the Random Walk Algorithm to regulate the global and local search behaviors of the robots. To assess the strategy's effectiveness, we implemented it in the ARGoS multi-robot simulator and conducted a series of experiments. The results clearly demonstrate the efficiency and effectiveness of our proposed search strategy.

Key words: Swarm Intelligence, Swarm Robotics, Search and Rescue Problem, Penguin Search Optimization Algorithm, Random Walk Algorithm.

1. Introduction. Exploring unknown environments is a fundamental concern in mobile robotics, crucial for applications like cleaning, search and rescue, and foraging. Search and rescue operations play a critical role in disaster management, aiming to ensure the safety of individuals and minimize rescue time. While these operations can be challenging for human rescuers, mobile robots are utilized to navigate and explore disaster-stricken areas, locate victims, and perform various tasks that enhance the effectiveness and efficiency of rescue operations. They can access hazardous or inaccessible areas, gather information, and provide assistance to humans.

In search and rescue operations, the effectiveness of multi-robot systems lies in their capacity to cover extensive areas, gather more data, and improve operational efficiency, particularly in tasks deemed perilous, time-consuming, or beyond human capabilities. Despite these advantages, coordinating and communicating among robots present significant challenges. Effective collaboration and information sharing are crucial for comprehensive search area coverage. Designing systems that facilitate seamless coordination and communication is essential to maximize the multi-robot team's effectiveness. Additionally, the ability of these systems to operate in environments with limited communication and navigation capabilities is crucial, considering that search and rescue scenarios often occur in areas where communication infrastructure may be damaged or nonexistent. Therefore, multi-robot systems need robust communication and adaptable search methods for such challenging environments.

Swarm intelligence has been a source of inspiration for various applications of mobile robotics. For example, some animals' exploration, return, and communication strategies have shown great effectiveness, especially in mobile robot exploration. In this work, we proposed a new search and rescue strategy based on swarm intelligence algorithms to achieve the following objectives:

1. Increase the explored areas.

^{*}LabSTIC Laboratory, Department of Computer Science, 8 May 1945 University, P.O. Box 401, Guelma, Algeria (zedadra.ouarda@univ-guelma.dz).

[†]LabSTIC Laboratory, Department of Computer Science, 8 May 1945 University, P.O. Box 401, Guelma, Algeria (zedadra.amina@univ-guelma.dz).

[‡]National Research Council of Italy, Institute for High Performance Computing and Networking (ICAR), Via P. Bucci 8/9C, 87036 Rende, Italy (antonio.guerrieri@icar.cnr.it).

[§]LabSTIC Laboratory, Department of Computer Science, 8 May 1945 University, P.O. Box 401, Guelma, Algeria (seridi.hamid@univ-guelma.dz).

[¶]Department of Computer Science, 8 May 1945 University, P.O. Box 401, Guelma, Algeria (douaaghelis41@gmail.com)

2. Disperse robots widely in their environment to increase the number of found victims.
3. Reduce search time.
4. Decrease the visits to already explored areas.

The majority of research in the search and rescue field predominantly relies on old SI algorithms (mature algorithms such as Ant Colony Optimization, i.e., ACO, Particle Swarm Optimization, i.e., PSO, Artificial Bee Colony, i.e., ABC). The Penguin Optimization Algorithm (POA) is an optimization technique inspired by the cooperative hunting strategies of penguins, specifically their ability to work together to locate and capture fish in harsh and dynamic environments. POA simulates this behavior by modeling the search process as a collaborative effort among multiple agents, each representing a penguin, to explore the solution space and converge on optimal solutions. To the best of our knowledge, the Penguin Search Optimization Algorithm (PeSOA) was used to optimize mathematical benchmark functions and has not yet been used in multi-robot-related problems. That is why we adapted and used the PeSOA in search and rescue problems. The POA is particularly well-suited for the problem of multi-robot search and rescue for several reasons:

- *Cooperative Behavior*: Similar to how penguins coordinate their movements to find food efficiently, POA enables multiple robots to collaborate, share information, and optimize their search patterns in a coordinated manner, enhancing the overall efficiency of search and rescue operations.
- *Adaptability*: The dynamic nature of the POA allows it to adapt to changing environmental conditions and obstacles, which is critical in unpredictable search and rescue scenarios where obstacles and conditions can vary widely.
- *Robustness*: The algorithm's robustness in dealing with complex and dynamic environments ensures that the robots can effectively navigate and perform their tasks despite uncertainties.

The paper is organized as follows: in Section 2, we introduce the Penguin Optimization Algorithm and summarize some recent related works. In Section 3, we present the finite state machine and the pseudo-code of the proposed algorithm. In Section 4, we present and discuss the obtained results. Finally, in Section 5, we conclude the work and present some perspectives.

2. State of the art. In this Section, we first present a description of the Penguin Optimization Algorithm (PeSOA). Then, we summarize some relevant and recent works related to the search and rescue problem and present a qualitative comparison of the reviewed works based on some relevant characteristics.

2.1. Description of the Penguin Optimization Algorithm (PeSOA). The Penguin Optimization Algorithm is a meta-heuristic algorithm inspired by the collaborative hunting strategies of penguins. This bio-inspired algorithm mimics the way penguins hunt for fish in groups, optimizing their energy expenditure and maximizing their food intake.

Hunting Strategy of Penguins. Penguins hunt in groups and use a strategy that involves synchronized dives and communication to locate and capture fish. Each penguin searches for food individually and then communicates its findings to the group. This collaboration allows the group to identify the areas with the highest concentration of food and optimize their foraging efforts. The key behaviors modeled in PeSOA include [1]:

- **Group Hunting**: Penguins hunt in groups, coordinating their efforts to maximize efficiency.
- **Communication**: Penguins communicate their positions and the amount of food found to the group, enabling collective decision-making.
- **Synchronized Dives**: Penguins synchronize their dives to systematically search different depths and areas.

Algorithm Structure. PeSOA simulates these behaviors to solve optimization problems by following these steps [1]:

1. **Initialization**: Generate an initial population of solutions (penguins) grouped into several groups.
2. **Random Search**: Each penguin performs a random search within a defined range (hole and level) until its oxygen reserves are depleted.
3. **Communication**: After the search, penguins return to the surface and share their findings (location and amount of food) with their group.
4. **Update Positions**: Penguins update their positions and probabilities of finding food in different locations based on the shared information.

5. Iteration: The process repeats for a number of generations, with the group collectively moving towards the areas with the highest food concentration (optimal solutions).

Key Components. The algorithm is composed of the components below [2]:

- Population Groups: Penguins are divided into groups, and each group searches specific areas.
- Intra-Group Communication: Penguins within the same group share their findings to refine their search strategy.
- Inter-Group Communication: Groups with less success may follow the more successful groups to improve their chances of finding food.
- Probabilistic Search: The search process is guided by the probability of food presence, which is updated based on the findings.

Mathematical Formulation. The position of each penguin is updated using the following equation:

$$D_{new} = D_{last} + rand() \times |X_{localbest} - X_{last}| \quad (2.1)$$

where D_{new} is the new position, D_{last} is the last position, $rand()$ is a random number for distribution, $X_{localbest}$ is the best local solution, X_{last} is the last solution.

PeSOA has been validated against several benchmark functions, such as the Rastrigin and Schwefel functions, and compared with other optimization algorithms like Particle Swarm Optimization (PSO) and Genetic Algorithm (GA). The results showed that PeSOA is robust and efficient, particularly in avoiding local optima and finding global solutions.

2.2. Related works. In the literature, Firthous and Kumar [10] introduced a new search and rescue algorithm called *Artificial Bee Colony with Modified Evolutionary Programming (ABCMEP)*, combining ABC and Modified Evolutionary Programming (MEP). ABC is used for obstacle avoidance, while MEP optimizes the path for efficiency. Simulated in MATLAB, the algorithm was tested in various environments, comparing it with ABC and ABCEP (Artificial Bee Colony with Evolutionary Programming). The ABCMEP algorithm demonstrated superior performance, providing efficient target localization in unknown areas by avoiding obstacles. The study presents the best cost values for each algorithm in different environments, highlighting the effectiveness of ABCMEP.

Garg et al. [7] introduce the Velocity-inspired Robotic Fruit Fly Algorithm (VRFA) as a search strategy for efficient victim search and real-time assessment in search and rescue operations. VRFA combines the Fruit Fly Optimization algorithm (FOA) for target search and tracking and the PSO algorithm for updating positions and velocities. The independent robot in the system performs activities such as local data processing, data sharing, movement plan formulation, and timeline generation. The study compares VRFA with other techniques using centralized and decentralized cooperation strategies for both static and dynamic targets. Results indicate that VRFA outperforms other algorithms, particularly in decentralized cooperation, demonstrating reduced search and rescue time and superior performance in dynamic scenarios.

Huang et al. [9] propose a novel algorithm for task allocation in a multi-robot cooperative rescue system, integrating the Ant Colony Contract Net Protocol (ACCNP). The algorithm uses the ACO algorithm to determine the initial allocation results and the CNP algorithm for dynamic adjustments in changing environments. Notable features include parallel computation, time efficiency, and adaptability to environmental changes. Simulated experiments in a fire rescue scenario reveal the CNP algorithm's superiority over ACO in terms of computation time, average task completion cost, and average task completion time. The results emphasize the effectiveness of CNP in solving multi-robot task assignment challenges within dynamic environments.

In their paper [12], the authors introduce a novel strategy called Distributed Particle Swarm Optimization (DPSO) for multi-robot exploration in search and rescue operations. The focus is on guiding a robot swarm to navigate the search space, avoid obstacles, and locate victims. Key concepts include robots avoiding static and dynamic obstacles, victims scattered randomly, robots sensing local environments, and sharing positions. DPSO addresses the drawbacks of classical PSO by incorporating an artificial potential function to attract forces toward unknown areas and victims, along with a repulsion forces function for collision avoidance, divided into intra and inter-repulsion forces. Experiments were conducted using Python and VRep for multi-agent model implementation and environmental simulation. Four experiments demonstrated that the DPSO algorithm

enables robots to escape local minima, find alternative paths, and navigate through disaster scenarios effectively, leading to optimal solutions.

Dah-Achinanon et al. [5] present a search and rescue algorithm utilizing ad-hoc networks with sporadic connectivity. The algorithm aims to bridge the gap between theoretical concepts and practical applications, emphasizing autonomy, decentralization, and effective research strategy. It facilitates communication among swarm members to share information about target findings and locations, enabling a base station to coordinate rescue efforts. The search method employs belief space exploration, incorporating key information from authorities, such as the last known locations of targets. The algorithm's principles include individual drones conducting searches based on available belief information, updating and distributing belief maps to prevent redundant searches, and using virtual stigmergy for belief map distribution. Validation of the algorithm involves tests in the ARGOS simulator platform and real-world experiments, assessing the time required for target discovery and relay chain establishment. Results from simulations and real-world tests confirm the feasibility and effectiveness of the proposed approach, demonstrating superiority over a random walk strategy.

In [13], authors describe an adverse search and rescue Multi-Agent Reinforcement Learning (MARL) approach to learning efficient coordination and collaboration strategies to achieve search and rescue mission objectives in the presence of adverse search and rescue communications. A Centralized Training approach with Decentralized Execution is used in this training approach. The aim is to coordinate the search to avoid visiting the same location multiple times and reduce the overall target exploration time by cooperative agents in the presence of adverse search and rescue. To investigate how adverse search and rescue interference can derail the performance of cooperative agents in a search and rescue mission and to what extent the proposed training algorithm can mitigate adverse search and rescue actions, the authors conducted four case studies to evaluate the proposed model. According to each case's results, the agents could locate the targets more successfully in the case with the modified reward structure.

The difficulties of search and rescue operations when the space is difficult to cover by human rescuers and ground robots have led to the emergence of a new technology for use, namely Unmanned Aerial Vehicles (UAVs). Authors in [14] proposed a Self-Organizing Mesh Network that is optimized for area coverage by maximizing the search area and maintaining wireless communication. This new system lays the groundwork for the behavior of more than two unmanned aerial vehicles. Whenever a UAV is connected to more than one other UAV, it will attempt to remain within the range of all of them. Additionally, this system prevents UAVs belonging to the same swarm from colliding with one another. Rather than having a leading coordinator UAV, all of the UAVs adjust based on each other. The authors proposed a mesh reliance architecture, which works by assigning the swarm a connectivity matrix, with each UAV connection being a rep and a swarm controller based on a Genetic Algorithm (GA) and Neural Network (NN). Based on the simulation results, the proposed methods are valid for organizing swarms. However, they are not as effective for swarm travel, and adding each drone did increase computation time. GA runs very slowly, making it a poor choice in emergencies.

In order to solve the problems of full-coverage path planning in search and rescue operations, which strive to completely cover the area of interest in a limited time and avoid obstacles autonomously in unknown areas, the authors in [16] proposed a new Neural Network Algorithm through the ABC algorithm. The Neural Network takes as input information about obstacles and coverage in the five directions and gives as output the speed of the left and right wheels. In the initial situation, the parameters are randomly generated, giving the path the neural network planned a lower score. According to the advantages of the ABC algorithm, such as the increased probability of finding the optimal solution and the robustness of the system, the authors used this method to optimize the parameters of the Neural Network and improve the training efficiency and effects of the entire system. To analyze the algorithm's performance, the authors define an evaluation function, which is divided into three parts: the coverage rate, the path repetition rate, and the failure rate. Based on experimental results, the ABC method, combined with a Neural Network path planning algorithm, can effectively control rescue robots to plan complete coverage paths and can be migrated to various environments with high robustness.

Naval operations such as marine search and rescue operations require a higher performance and efficient control to rescue survivors. The authors [15] presented a new method for visual navigation and Unmanned Surface Vehicle (USV) control named Conventional Convolutional Neural Network Spatial Softmax (CNN-SS). The authors divided their work into two main parts, summarized as follows:

- Deep learning-based visual navigation architecture for USV and floating object positioning in USV-UAV operations. This approach improves visual positioning accuracy and computational efficiency by introducing a soft-max spatial layer and a two-stage structure.
- Reinforcement Learning (RL)based USV control strategy approaches and encircles a floating object. The trained control policy can improve the control system's performance under wave disturbances by adding observable wave features to the state space.

The proposed visual navigation architecture consists of two stages: the first consists of estimating the position of the USV and the floating object, and the second stage consists of estimating the heading angle of the USV. This two-stage architecture is built based on CNN and the network structure. A USV controller based on RL has been developed to avoid collisions with floating objects. The basic principle of RL is to maximize the accumulated reward at every step by iteratively optimizing a policy. The authors conducted several experiments to evaluate the performance of the visual navigation model and the USV controller model. They evaluated four network models (CNN-SS, Conventional Convolutional Neural Network (CNN), Fully Convolutional Network (FCN), YOLOv5) for the first proposed algorithm and designed three tasks to evaluate the second algorithm. Therefore, the proposed visual navigation architecture can provide high-accuracy position, velocity, and heading angle estimations under most weather conditions. However, this latter still has some limitations in heavy foggy weather. Finally, after the simulation, the authors demonstrated the effectiveness and superiority of the proposed algorithms.

Authors in [11] developed a new algorithm called the Ant Search Path with Visibility algorithm (ASPV) for the NP-hard optimal search path problem with visibility. This algorithm draws inspiration from ACO principles, which define 96 variants based on four key components of traditional ACO: pheromone initialization, pheromone update, restart, and boosting. The authors conducted several experiments to identify the best variant of the proposed ACO algorithm variants. In these experiments, three phases were conducted:

- Configuration phase, which determines the optimal parameter pairs for each variant.
- multi-run evaluation phase allows for determining how the performance of an ACO varies between runs on a given instance.
- The across-instance evaluation phase provides a better understanding of an ACO's performance in a variety of instances.

The authors compared the performance of the best ASPV algorithm variants with that obtained through a Mixed-Integer Linear Program (MILP) and with a simple greedy heuristic. Results indicated that the proposed algorithm produced search paths with higher success probabilities in a shorter period.

In the study by AI et al. [3], a novel autonomous coverage planning model for maritime search and rescue operations is introduced, leveraging reinforcement learning. The key contributions include (1) complex environment modeling, and (2) introducing a reinforcement learning algorithm with a multi-objective reward function. This function considers avoiding obstacles, preventing repeated paths, and favoring high-probability areas. The algorithm effectively addresses issues like sparse rewards, sub-goal conflicts, and reduces overlapping paths during learning, (3) developing an action selection policy that balances exploitation and exploration to determine the global-optimal solution. Experiments in a simulation area demonstrate the model's validity. Comparisons with various trajectory planning algorithms, including Boustrophedon Motion (BM), Boustrophedon Cellular Decomposition (BCD), and Boustrophedon Motions with the A* search (BA*), show that the proposed algorithm generates search paths covering the entire search and rescue area safely, with low repetition rates and prioritization of high-probability areas for searching.

Authors in [6] proposed a new extension of the A* algorithm for the 3D search and rescue environment. This algorithm is based on the A* and Task Allocation algorithms to obtain a faster and more efficient path-planning method. The objectives achieved by the author are summarized below:

- General 2D and 3D trajectory planning of UAVs.
- Granting UAVs the ability to avoid obstacles when performing tasks.
- Identifying the shortest path for all UAVs in a minimum of time.

In this paper, the task allocation algorithm provides advantages to the system, such as a parallel search of the environment, allowing the drones to complete the entire task by distributing the assignments to each drone, helping to reduce the total route planning time and memory space of each drone. Moreover, the A* algorithm

is generally used for the path planning optimization problem. To evaluate the performance of the proposed algorithm, the author created 2D and 3D simulation scenarios to test the functionality of the task allocation algorithm and performed experiments in a 3D environment to verify and evaluate the performance of the improved A* algorithm compared to the classical A* algorithm. According to the results obtained from the simulation, it can be observed that the improved algorithm has better running time, speed, and efficiency than the classical algorithm and has higher efficiency in the coverage environment, and reduces the amount of storage required.

Because of the increase in the number of casualties in the maritime sector, maritime authorities and operation centers are trying to develop a quick search for survivors at sea using different technologies such as USV, Unmanned Underwater Vehicle (UUV), and UAV. [4] introduced a new contribution by using a UAV swarm. This latter was divided into a two-phase method to solve the (CPP) problem. The first phase presents a methodology for transforming the search area into a graph consisting of vertices and edges using grid-based area decomposition. Hence, the proposed method takes into account the angle at which the area is decomposed to minimize the size of the coverage area. The second phase focuses on determining the optimal path for multiple UAVs based on the results of the first phase to reduce the completion time. The authors developed a Mixed-Integer Linear Programming (MILP) model that minimizes the time required to cover the search area. As a constraint, the region (position of nodes, the angle between nodes, etc.) and the dynamics of the robot covering the nodes are considered. The researchers in this paper observed that the proposed MILP model was time-consuming for large-scale real-world problems. To solve this problem, a new algorithm called Randomness Search Heuristic (RSH) was developed. This approach consists of three phases in each iteration:

1. Random construction: constructs the coverage path of UAVs by sequentially selecting the adjacent nodes.
2. Repair: repairs the generated roads if all nodes are uncovered to recover the feasibility.
3. Local search: improves the constructed path for each UAV to obtain high-quality solutions that are close to the optimal road.

This paper considered numerical experiments to validate the efficiency and effectiveness of the proposed approach and real-world experiments with two UAVs to verify the validity of the proposed algorithm in the real world. According to the results of the numerical experiments, the RSH algorithm can produce near-optimal solutions in a much shorter computational time than a commercial solver. In contrast to the real-world experiments, the mission execution time was longer due to the influence of wind or network communication uncertainty.

After summarizing the different related works, we conclude some important points:

1. Based on Table 2.1, we can group the research according to the scenarios used in the different search and rescue operations (based-ground search and rescue, maritime search and rescue, and aerial search and rescue).
2. Some works used a single sink with a fixed position, the others did not specify the number of sinks and their positions.
3. Most of the works have used static targets with a random position.
4. The majority of the works used multiple robots with a random position.
5. All research done in the based-ground search and rescue focuses on homogeneous robots and is based on decentralized control.
6. Most research focuses on cooperation between robots using direct communication.
7. The majority of studies take into account the lack of redundancy in space exploration, and all of these studies use different search methods.
8. In the search and rescue based-ground, all research experiments were conducted by simulation using different simulators.
9. As for all the works, the use of swarm intelligence-based algorithms is very weak and needs to be investigated since these algorithms provide scalable, flexible, and robust solutions to systems like search and rescue ones.

3. Proposed Algorithm: Modified Penguin Search Optimization Algorithm (MPeSOA). In this Section, we present the proposed search and rescue algorithm named: The Modified Penguin Search Optimization Algorithm (MPeSOA). During its life cycle, the robot can be in one of the four behaviors: **Global**

Table 2.1: Qualitative comparison of the summarized related works.

			[10]	[7]	[8]	[9]	[16]	[12]	[13]	[14]	[5]	[15]	[11]	[3]	[6]	[4]
Environment	Complexity	with obstacles	X	X			X	X	X		X			X		
		Free of obstacles			X	X					X	X	X	X	X	X
Object	Distribution	Random		X				X	X		X	X	X			X
		fixed	X			X										
	clustered				X											
	Nature	Static	X	X	X	X		X	X		X	X				X
	Dynamic		X								X	X				
Position	fixed	X			X	X									X	X
	random			X	X		X	X	X	X	X	X	X	X	X	X
Homogeneity	Yes	X	X	X	X	X	X	X	X	X				X	X	X
	No											X				
Robot	Number	single	X				X								X	
		multiple		X	X	X		X	X	X	X	X	X	X	X	X
control	centralized			X	X											
	decentralized	X	X		X		X	X	X	X	X	X	X	X	X	X
cooperation	Yes		X	X	X		X	X	X	X	X	X	X	X	X	X
	No	X				X									X	X
Strategy	communication	direct		X	X	X		X		X	X	X	X	X	X	X
		indirect							X							
exploration	Yes			X		X								X		
	No	X		X	X		X	X	X	X			X	X	X	X
exploration type	ABCEMP	X														
	VRFA		X													
	CNP				X											
	MARL							X								
	ABC					X										
	ANN					X										
	belief based search										X					
	GA									X						
	leader follower			X												
	D-PSO						X									
	DL											X				
	RL											X		X		
	ASPV											X				
	enhanced A*														X	
RSH															X	
Real world	Real experiments										X					X
	Simulator	ArGOS									X					
Simulations	Netlogo			X												
	Python				X	X		X			X		X	X	X	X
	Pybullet				X			X								
	VRep						X									
	MATLAB		X													X
	C++													X		

Search, Local Search, Migration, and Obstacle Avoidance. The robots start all from the central depot and change their behaviors according to input data gathered by their sensors. The state machine representing the behavior of the robots is given by Figure 3.1.

We also present in this section a pseudo code of the **proposed MPeSOA algorithm** (Algorithm 1), and the pseudo-codes of the four key states: **Global search** (Algorithm 2), **Local search** (Algorithm 3), **Obstacle avoidance** (Algorithm 5), and **Migration** (Algorithm 4).

Below, we explain the different behaviors of the state machine. Throughout the subsequent rules, the oxygen reserve $O(t)$ will be treated as a constant value set to 1.

Global search. During this phase, each group of robots undergoes a relocation process to a new position, guided by the LocalBest solution obtained from the previous dive. This relocation is executed by employing the penguin search position equation as depicted in Equation 3.1. Individual robots may encounter obstacles or victims throughout the search process, prompting distinct decision-making. For instance, when faced with

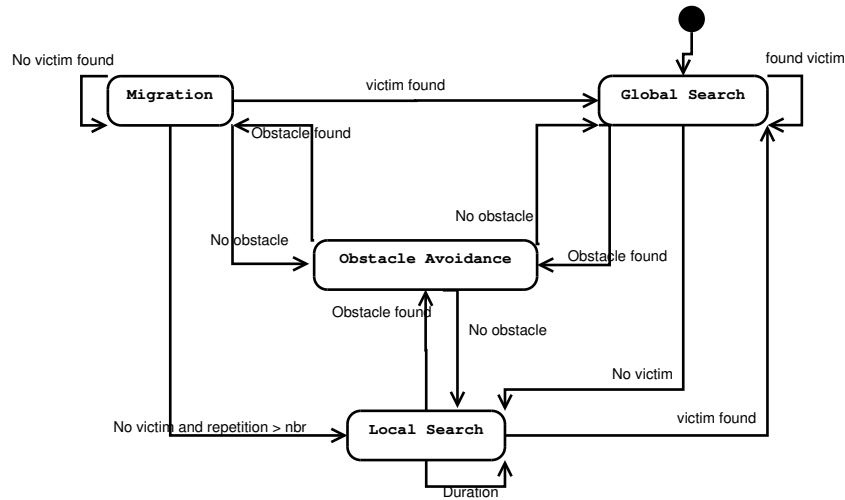


Fig. 3.1: MPeSOA's State Machine.

```

1 Initialize the number of groups;
2 Initialize the total number of robots;
3 Generate a random local best for each group;
4 Distribute randomly the robots in groups;
5 while (stopping criteria is not reached) do
6   for each group i do
7     Goto Global Search (Algorithm 2);
8     Update the food abundance degree for this group using Equation 3.4;
9   Update the total eating food;
10  Update the global best solution;
11  Update membership function values for each group by Equation 3.3;

```

Algorithm 1: Pseudo code of the MPeSOA Algorithm.

obstacles, a robot must avoid them using the *obstacle avoidance state*. Conversely, the robot must prioritize its rescue and subsequent safety upon discovering victims. Once the group reaches a new position, all group members must shift their state to the Local Search state. Algorithm 2 presents the pseudo-code of the Global search state.

$$x_j^i(t+1) = x_j^i(t) + rand() \times |x_{Localbest}^i - x_j^i(t)| \quad (3.1)$$

where: $x_j^i(t)$ is the position of penguin j allocated to the i^{th} group at t^{th} instant. $x_{localbest}^i$ is the best solution found by i^{th} group. $rand(a, b)$ is a random number drawing from $(0, 1)$

Local Search. Before starting the local search, the robots must await the arrival of all group members, ensuring the collective initiation of the search procedure employing the Random Walk Algorithm (RWA) (Equation 3.2). As in the Global search, the robots have to get around obstacles and effectively rescue victims encountered along the way. However, the key distinction lies in the search duration. In this case, the robots apply the RWA for a predefined period and subsequently repeat this process in the absence of victim discoveries, thereby providing additional opportunities to locate victims. In instances where no victims are detected, the robots must undergo migration by reverting to the Migration state. Conversely, upon successful victim detection, the group updates their local best and transitions to the Global search state. The pseudo-code of this state is

```

1 calculate new position using Equation 3.1;
2 while the group does not reach the new position do
3   if  $\exists$  obstacle then
4     Goto Obstacle Avoidance (Algorithm 5);
5   else
6     if  $\exists$  victim then
7       get the victim;
8       update the number of victims rescued;
9       wait a few time before updating localbest for this group;
10      if the time is expired then
11        update localbest for this group;
12        update the quantity of eating fish for this group;
13        Goto Global Search (Algorithm 2);
14 if the group reached the new position then
15   Goto local Search (Algorithm 3);

```

Algorithm 2: Pseudo code of the global search Algorithm.

represented by Algorithm 3.

$$sl_i = rand(a, b) \quad (3.2)$$

where: sl_i is the step length for the i^{th} , and $rand(a, b)$ is a random number drawing from (a, b)

```

1 while repetition process > 0 do
2   while the duration has not expired do
3     while  $\nexists$  obstacle or victim or pheromone do
4       Random walk using Equation 3.2;
5     if ( $\exists$  obstacle) then
6       Goto Obstacle Avoidance (Algorithm 5);
7     else
8       if ( $\exists$  victim) then
9         Get the victim;
10        Update the number of victims rescued;
11        Update the quantity of eaten fish for this group;
12        Update duration for this group;
13    Update LocalBest for this group;
14    if new localbest=the last localbest then
15      decrease repetition process;
16      if repetition process =0 then
17        Goto Migration (Algorithm 4);
18    else
19      Goto Global Search (Algorithm 2);

```

Algorithm 3: Pseudo code of the Local Search Algorithm.

Migration State. If victims remain undetected, the group will initiate migration and merge with another group. Before the migration process, it is necessary to search and rescue to compute the probabilities of the existence of fish for all groups using Equation 3.3. However, if all computed probabilities are found to be less than 0.5, the current group will refrain from migrating and take a random localbest. Conversely, if at least one probability exceeds or equals 0.5, the group will initiate migration based on the aforementioned approach

outlined in the Migration State. The equation governing the update of the Quantity of Eating Fish is equation 3.4:

$$P_j(t+1) = \frac{QEF^i(t)}{\sum_{i=1}^k QEF^j(t)} \quad (3.3)$$

$$QEF^i(t) = \sum_{j=1}^{di} E_j^i(t) \quad (3.4)$$

where: $E_j^i(t)$ represents the quantity of eaten fish of the j^{th} robot of i^{th} group at t^{th} instance.

```

1 Return to the best position;
2 calculate Pi using Equation 3.3;
3 if all probabilities < 0.5 then
4   set a random localbest;
5   Goto Global Search (Algorithm 2);
6 else
7   divide the group into two subgroups;
8   give the subgroups opposite directions;
9   while  $\nexists$  obstacle or victim or pheromone do
10    use the same random step for all members within the same group until they reach the hunting group;
11   if  $\exists$  obstacle then
12    Goto Obstacle Avoidance (Algorithm 5);
13   else
14     if  $\exists$  victim then
15       get the victim;
16       update the number of victims rescued;
17       update localbest for this group;
18       update the quantity of eating fish for this group;
19       Goto Global Search (Algorithm 2);

```

Algorithm 4: Pseudo code of the Migration Algorithm.

Obstacle avoidance. Our explorer robot is equipped with 24 IR proximity sensors positioned strategically around its structure. These sensors possess detect obstacles within a proximity range of 30 cm. In accordance with the readings obtained from these sensors, if the robot identifies the presence of an obstacle, it retrieves the angle information and adjusts its movement accordingly. When encountering a negative angle, the robot executes a right turn, whereas a positive angle prompts a left turn.

```

1 Get readings from all proximity sensors;
2 Accumulate all readings and get the accumulated value and the angle;
3 if accumulated value > 0 then
4   if accumulated angle > 0 then
5     Turn left;
6   else
7     Turn right;

```

Algorithm 5: Pseudo code of the obstacle avoidance Algorithm.

Table 4.1: The proposed simulation scenarios

Scenario 1: the influence of robot's number	Scenario 2: the influence of group's number
number of robots = 30, 40, 50, 60 victim distribution: clusters number of groups: 6 number of clusters: 12 number of victims: 910 obstacle density: 10% environment size: 120m X 120m	number of robots = 50 victim distribution: clusters number of groups: 3, 5, 6, 8 number of clusters: 12 number of victims: 910 obstacle density: 10% environment size: 120m X 120m
Scenario 3: the influence of cluster's number	Scenario 4: the influence of environment size
number of robots = 50 victim distribution: clusters number of groups: 6 number of clusters: 2, 4, 8, 16 number of victims: 1000 obstacle density: 10% environment size: 120m X 120m	number of robots = 50 victim distribution: clusters number of groups: 6 number of clusters: 12 number of victims: 910 obstacle density: 10% environment size: 80m X 80m, 100m X 100m, 150m X 150m, 190m X 190m

4. Results and Discussions. In order to validate the proposed algorithm (PeSOA), we implemented it by using the ARGoS mobile robotics simulator. Finding the right parameters for the algorithm will certainly increase its efficiency. We conducted various simulations to test the influence of the different parameters on the performance of the proposed algorithm. The considered parameters comprehend (i) the number of robots, (ii) the number of groups, (iii) the number of clusters, and (iv) the environment size. Moreover, we used as a performance criterion the number of collected victims in a fixed time (20 minutes).

4.1. Simulation scenarios. We provide in this section an overview of four simulation scenarios. We test the influence of robot numbers, group numbers, cluster numbers, and environment sizes on the performance of the proposed strategy by using these scenarios. Table 4.1 shows the different proposed scenarios:

4.2. Results and discussions. We introduce and analyze in this section the results obtained by applying this algorithm using scenarios presented in Table 4.1. It is important to note that the reported results represent the average outcome from five simulations.

Results of scenario 1. As the number of robots increases, the total number of found victims also increases. Increasing the number of robots positively impacts search and rescue operations. However, the rate of increase in the number of found victims is non-linear with the increase in the number of robots. The rate of increase diminishes as the number of robots increases. When the number of robots increases from 30 to 40 robots, the number of found victims is increased by 27, while going from 50 to 60 robots, it is increased by 156. After 60 robots, the performances did not significantly improve. Table 4.2 and Figure 4.1.a show the results obtained in this scenario.

Table 4.2: Results of scenario 1

Number of robots	30	40	50	60
MPeSOA	40%	43%	48%	65%

Results of scenario 2. With a smaller number of groups, there might be fewer opportunities for efficient coverage of the search area. Increasing the number of groups leads to a notable improvement in the number of found victims. This indicates that the algorithm benefits from finer granularity in dividing the search area into smaller sections, enabling more thorough exploration and detection of victims. However, the rate of improvement starts to diminish compared to the previous increase from 3 to 5 groups. The results indicate that

there's an optimal number of groups for maximizing the effectiveness of the search and rescue operation. Also, increasing the number of groups allows for better coverage of the search area but also introduces challenges in coordination among the groups. Table 4.3 and Figure 4.1.b show the results obtained in this scenario.

Table 4.3: Results of scenario 2

Number of groups	3	5	6	8
MPeSOA	38%	48%	64%	72%

Results of scenario 3. The algorithm achieves the maximum possible number of found victims when all clusters are thoroughly searched. This suggests that having a smaller number of clusters simplifies the search process and maximizes coverage within each cluster. Further increasing the number of clusters leads to a significant decrease in the total number of found victims. With eight clusters, the search area becomes more fragmented, making it challenging for the robots to cover each cluster thoroughly. Additionally, coordinating movements and managing resources across more clusters might introduce inefficiencies or delays in the search and rescue process. Table 4.4 and Figure 4.1.c show the results obtained in this scenario.

Table 4.4: Results of scenario 3

Number of clusters	2	4	8	16
MPeSOA	100%	75%	60%	57%

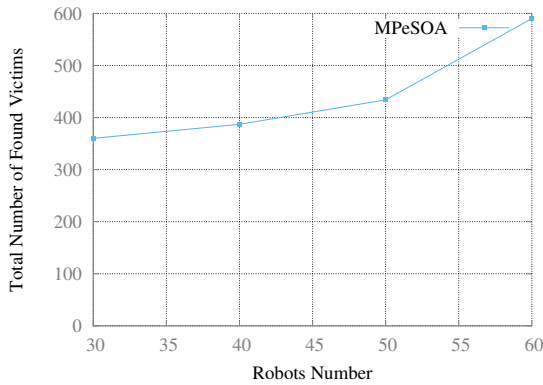
Results of scenario 4. The results indicate that the environment size has a significant impact on the algorithm's performance in terms of finding victims. As the environment size increases, the total number of found victims decreases consistently. When the environment size increases from 80x80 to 100x100, there's a noticeable decrease in the number of found victims. This reduction suggests that expanding the search space slightly beyond a certain threshold can lead to decreased search efficiency. The trend continues as the environment size increases to 150x150 and 190x190, with a corresponding decrease in the number of found victims. The rate of decrease appears to accelerate as the environment size grows larger. This indicates diminishing returns in search efficiency with further expansion of the search area. Table 4.4 and Figure 4.1.d show the results obtained in this scenario.

Table 4.5: Results of scenario 4

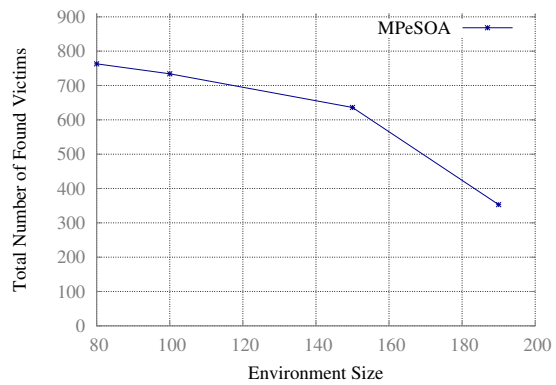
Environment size	80X80	100X100	150X150	190X190
MPeSOA	84%	81%	70%	39%

4.3. Recommendations for parameters settings. Based on the analysis of the results from the experiments detailed in the document, here are some extended recommendations on how to set the studied parameters to enhance the effectiveness of the search and rescue strategy using the Penguin Optimization Algorithm (PeSOA):

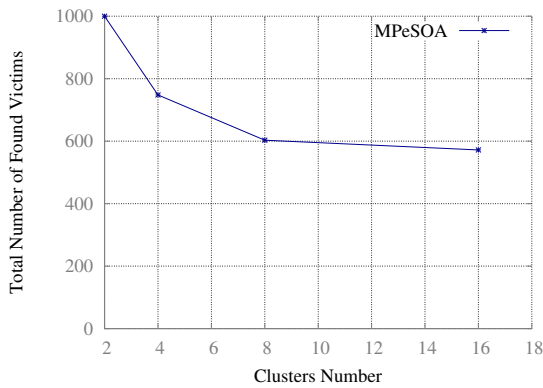
1. *Number of Robots:* Deploy a fleet of around 50-60 robots for optimal performance without unnecessary redundancy. This range maximizes the number of found victims while maintaining efficiency.
2. *Number of Groups:* Use 6-8 groups to balance the coverage area and coordination complexity. This setting ensures thorough exploration and effective victim detection while managing inter-group communication efficiently.
3. *Number of Clusters:* Keep the number of clusters low (2-4 clusters). This approach simplifies the search process, allows more thorough coverage of each cluster, and enhances the efficiency of the overall search and rescue operation.



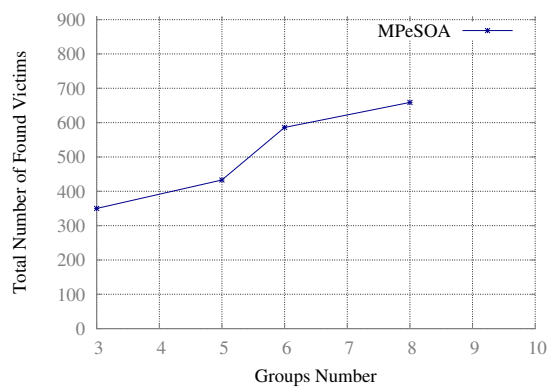
(a) Results of scenario 1



(b) Results of scenario 2



(c) Results of scenario 3



(d) Results of scenario 4

Fig. 4.1: Simulation results

4. *Environment Size*: For environments larger than 100x100 meters, consider deploying additional robots or segmenting the environment into smaller manageable areas. For optimal performance, environments around 100x100 meters provide a good balance between search efficiency and area coverage.
5. *Enhanced Local Search*: Develop a more efficient local search strategy beyond the Random Walk Algorithm (RWA) to improve detection rates within local clusters. This can involve using heuristic or machine learning-based approaches to optimize the search paths dynamically.
6. *Redundancy Reduction*: Implement mechanisms to prevent redundant exploration of already searched areas. Techniques such as virtual stigmergy or communication protocols that share explored regions among robots can significantly reduce overall search time and increase efficiency.
7. *Obstacle Avoidance*: Ensure robust obstacle avoidance mechanisms are in place. The use of multiple proximity sensors and adaptive movement strategies based on sensor input can enhance the robots' ability to navigate complex environments effectively.

5. Conclusion and future work. Our focus in this study lies in the dispersion of robot groups achieved through an implicit division of the environment based on the positions of each group. To address this, we proposed a multi-robot search and rescue algorithm, PeSOA (Penguin-inspired Search Optimization Algorithm). The algorithm has been successfully implemented within the ARGoS simulation platform, and the obtained results are promising.

Future perspectives may include: (1) Further analysis is needed to evaluate the robustness of the proposed algorithm under different scenarios, such as varying cluster densities, uneven distribution of victims, or obstacles in the search area, (2) developing a more efficient local search strategy that surpasses the randomness of the current approach, (3) preventing redundant exploration of previously explored areas and consequently reducing the overall search time.

REFERENCES

- [1] GHERAIBIA, Y., MOUSSAOUI, A. *Penguins search optimization algorithm (PeSOA)*. In Recent Trends in Applied Artificial Intelligence: 26th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2013, Amsterdam, The Netherlands, June 17-21, 2013. pp. 222-231. Springer Berlin Heidelberg.
- [2] MANSOURI, A., AMINNEJAD, B., AHMADI, H.. *Introducing modified version of penguins search optimization algorithm (PeSOA) and its application in optimal operation of reservoir systems*. Water Science and Technology: Water Supply, 18(4), pp. 1484-1496.
- [3] A. BO, J. MAOXIN, X. HANWEN, X. JIANGLING, W. ZHEN, L. BENSHUAI AND Z. DAN, *Coverage path planning for maritime search and rescue using reinforcement learning*, Ocean Engineering, 241 (2021), pp. 117–134.
- [4] C. S. WON, P. H. JI, L. HANSEOB, S. D. HYUNCHUL AND K. S. YOUNG, *Coverage path planning for multiple unmanned aerial vehicles in maritime search and rescue operations*, Computers and Industrial Engineering, 161 (2021), pp. 107612.
- [5] D. ULRICH, M. BAJESTANI, S. EHSAN, L. P. YVES AND B. GIOVANNI, *Search and rescue with sparsely connected swarms*, Autonomous Robots, 47 (2023), pp. 849–863.
- [6] D. YUWEN, *Multi UAV Search and Rescue with Enhanced A* Algorithm Path Planning in 3D Environment*, International Journal of Aerospace Engineering, 2023 (2023).
- [7] G. VIKRAM, T. RITU AND S. ANUPAM, *Comparative Analysis of Fruit Fly-Inspired Multi-Robot Cooperative Algorithm for Target Search and Rescue*, in IEEE World Conference on Applied Intelligence and Computing, AIC 2022, pp. 444–450.
- [8] N. GOMEZ, N. PENA, S. RINCON, S. AMAYA, AND J. CALDERON, *Leader-follower Behavior in Multi-agent Systems for Search and Rescue Based on PSO Approach*, in IEEE SOUTHEASTCON Conference, 2022, pp. 413–420.
- [9] H. JIE, S. QUANJUN AND X. ZHANNAN, *Multi robot cooperative rescue based on two-stage task allocation algorithm*, *Journal of Physics: Conference Series*, 2310 (2022).
- [10] R. KUMAR, M. FIRTHOUS AND R. KUMAR, *Multiple oriented robots for search and rescue operations*, in IOP Conference Series: Materials Science and Engineering, 912(2022).
- [11] M. MICHAEL, A. Z. IRÈNE AND Q. C GUY, *Ant colony optimization for path planning in search and rescue operations*, European Journal of Operational Research, 305 (2023), pp. 53–63.
- [12] PAEZ, D., ROMERO, J. P., NORIEGA, B., CARDONA, G. A., AND CALDERON, J. M., *Distributed particle swarm optimization for multi-robot system in search and rescue operations*, IFAC-PapersOnLine, 54 (2021), pp. 1–6.
- [13] RAHMAN, A., BHATTACHARYA, A., RAMACHANDRAN, T., MUKHERJEE, S., SHARMA, H., FUJIMOTO, T., AND CHATTERJEE, *Adversersearch and rescue: Adversersearch and rescueial Search and Rescue via Multi-Agent Reinforcement Learning*, IEEE International Symposium on Technologies for Homeland Security, HST 2022.
- [14] RUETTEN, L., REGIS, P. A., FEIL-SEIFER, D., AND SENGUPTA, S., *Area-Optimized UAV Swarm Network for Search and Rescue Operations*, 10th Annual Computing and Communication Workshop and Conference, CCWC 2020, pp. 613–618.
- [15] WANG, Y., LIU, W., LIU, J., AND SUN, C., *Cooperative UAV marine search and rescue with visual navigation and reinforcement learning-based control*, ISA transactions, 137 (2023), pp. 222–235.
- [16] YANG, L., XING, B., LI, C., AND WANG, W. , *Research on Artificial Bee Colony Method Based Complete Coverage Path Planning Algorithm for Search and Rescue Robot*, In 2022 5th International Symposium on Autonomous Systems (ISAS), pp. 1–6.

Edited by: Katarzyna Wasielewska-Michniewska

Regular paper

Received: Mar 18, 2024

Accepted: Jul 29, 2024



FEDERATED LEARNING FOR INTERNET OF MEDICAL HEALTHCARE: ISSUES AND CHALLENGES

NIKITA CHELANI*, SHIVAM TRIPATHY†, MALARAM KUMHAR‡, JITENDRA BHATIA§, VARUN SAXENA¶, SUDEEP TANWAR|| AND ANAND NAYYAR**

Abstract. Federated Learning is a decentralized machine learning method that allows collaborative model training across several devices or institutions while maintaining the privacy and localization of data. Since the raw data is used locally, this collaborative method enables the development of a strong and precise global model without jeopardizing the privacy and security of sensitive data. The healthcare sector is an important one that focuses on preserving and enhancing people's health through medical services, diagnoses, treatments, and preventative measures. Efficient evaluation of Federated Learning in the Internet of Medical Things (IoMT) enables breakthroughs in medical image analysis, electronic health record analysis, personalized treatment planning, and drug development by enabling institutions to train models locally on sensitive patient information without sharing raw data. This paper presents the role of Federated Learning in healthcare and current trends in Federated Learning-based healthcare. A case study is presented on deep Federated Learning for privacy-preserving in healthcare. Finally, challenges and future research directions are discussed in the paper.

Key words: Federated Learning, Healthcare, Data Privacy, Machine Learning, Medical Image Analysis, Electronic Health Records, Data Security.

1. Introduction. A powerful machine learning-based health protection system utilizes the doctor's clinical judgment and the computer's massive processing power. Machine learning is crucial in healthcare, especially in areas such as computer-aided diagnosis, image annotation, image-guided medical help, image database retrieval, multimodal image fusion, and medical image segmentation, where inadequacies may be fatal. There are more options to create a system for patient recovery thanks to the improvement of health-related information. In the healthcare sector, machine learning has had limited social impact. Machine learning is the key to decreasing healthcare costs and encouraging improved patient-clinician communication. Multiple health-related uses of ML solutions include assisting physicians in locating multiple patient-specific drugs and therapies and assisting patients in deciding when and if to arrange follow-up visits [2]. Various ML algorithms have been used in healthcare environment. It vary in their implementation in terms of their methodology, the nature of input and output data [37]. The authors in [38] reviewed IoT frameworks and ML algorithms in healthcare sector, specifically on voice pathology. It also covers the impact of ML in various healthcare applications such as blood pressure and oxygen saturation monitoring using smartphones. ML technologies continue to go deeper into the healthcare environment, which also places higher demands on intelligent healthcare [36]. Figure 1.1 shows the various applications of ML in healthcare.

Federated Learning is an innovative approach for training a global machine learning model using data scattered across many data groups, removing the necessity for raw data exchange. Once a global model is shared

*Department of Computer Science and Engineering, Institute of Technology, Nirma University Ahmedabad, Gujarat, India. (22mced03@nirmauni.ac.in).

†Department of Information Technology, L. J. Institute of Engineering and Technology, Ahmedabad, India. (spt3009@gmail.com).

‡Department of Computer Science and Engineering, Institute of Technology, Nirma University Ahmedabad, Gujarat, India. (Corresponding author, malaram.kumhar@nirmauni.ac.in)

§Department of Computer Science and Engineering, Institute of Technology, Nirma University Ahmedabad, Gujarat, India. (Corresponding author, jitendra.bhatia@nirmauni.ac.in)

¶Department of Computer Engineering, Govt. Mahila Engineering College, Ajmer, India. (vps@gweca.ac.in).

||Department of Computer Science and Engineering, Institute of Technology, Nirma University Ahmedabad, Gujarat, India. (sudeep.tanwar@nirmauni.ac.in).

**School of Computer Science, Duy Tan University, Da Nang, Vietnam. (anandnayyar@duytan.edu.vn).

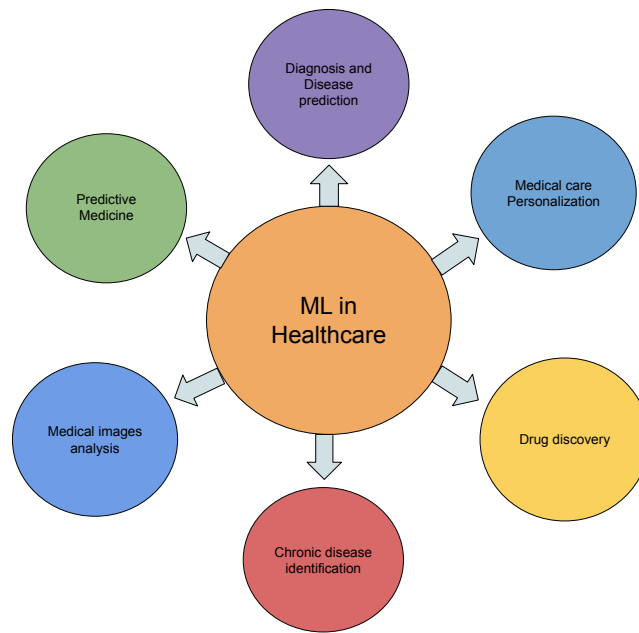


Fig. 1.1: Machine Learning applications in Healthcare

between different locations, the model is trained at each site using local data. The local models' parameter changes are then transmitted to an aggregation server and integrated into the final model. This is repeated until the global model's convergence requirement is met. Figure 1.2 presents the basic federated learning architecture. The benefits of Federated Learning have lately been proven in many practical applications, such as language modeling and picture classification. It is particularly pertinent in healthcare applications where data is replete with sensitive, personally identifiable information, and data analysis techniques must adhere to legal and regulatory standards [1].

1.1. Motivation. The potential for the Federated Learning technique to change the area of healthcare is what led to choosing it for research. The study aims to explore the intersection between cutting-edge machine learning methods and the complexity of healthcare data administration. The primary problem of protecting patient privacy and data security in healthcare research is addressed by Federated Learning's capability to support collaborative model training across diverse data sources. This compatibility between decentralized data analysis and the need for data secrecy provides a strong justification for this research work. The anticipated results, like improved diagnosis accuracy, customized treatment plans, and expedited medical research, highlight the profound change that Federated Learning might bring to the healthcare industry. Through the examination of Federated Learning applications, the research is conducted in hopes of assisting in the development of a healthcare environment that is more secure, effective, and patient-centered. Figure 1.3 shows various key technologies used for implementing Federated Learning in healthcare. Today's biggest difficulty for AI researchers and practitioners is how to legally address the issue of data fragmentation and isolation [3]. By developing a global model without distributing raw data among sites, Federated Learning offers a first degree of privacy protection. Federated Learning, however, may occasionally be exposed to inference attacks.

1.2. Scope of the paper. The importance of data privacy and security has emerged as a major global problem as huge organizations compromise user privacy and data security. Public media and governments are very concerned about reports of data leaks. Federated Learning aims to enable ML from non-located data, hence addressing privacy and data governance issues. Each data controller in an Federated Learning context establishes its governance procedures and privacy guidelines, managing data access and having the authority to withdraw it. It covers both the validation phase and the training phase. Federated Learning might provide new

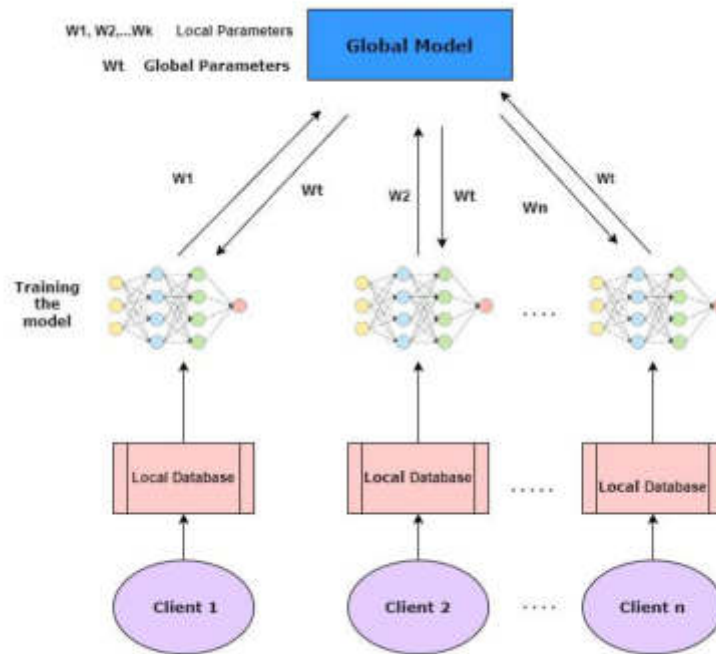


Fig. 1.2: Basic Federated Learning Architecture

opportunities in this way [4]. In this paper, the authors explore the role of Federated Learning in healthcare and review the current trends.

1.3. Organization of the paper. This section discusses the organization of the survey. The rest of the paper is organized as follows: Section 3 examines earlier research on collaborative machine learning that protects privacy with an emphasis on medical applications. Section 2 gives a brief introduction about the topic and its various use cases in medical field. Section 4 discusses the RPDFL method for digital healthcare applications. It deals with problems with conventional Federated Learning, such as gradient leaking and data silos in healthcare companies. Section 5 outlines machine learning’s potential to benefit healthcare and its transformational role in clinical decision-making, healthcare research, and tailored therapy while maintaining data security and privacy.

2. Background. Technological breakthroughs have revolutionised the healthcare business over the last several decades, ushering in a new era of customised, efficient, and data-driven medical services. At the centre of this transformation are several interconnected technologies, including IoT, SDN, Fog Computing, IoMT, and Machine Learning. Each of these technologies is critical to the transformation of healthcare, the improvement of patient outcomes, and the simplification of administrative operations.

2.1. Healthcare. Healthcare is one of the industries that is connected to everyone’s life. It is a necessary part of people’s life, encompassing physical, emotional, and mental aspects. It entails detecting, treating, and dispensing vaccinations. The healthcare business has incorporated technological advancements to make it more efficient and effective. There are various stages in the evolution of healthcare that might be labelled as “Healthcare X.0.” These stages represent significant developments and advancements in healthcare practices, techniques, and delivery systems. Figure 2.1 depicts the evolution of healthcare from 1.0 to 5.0, and the

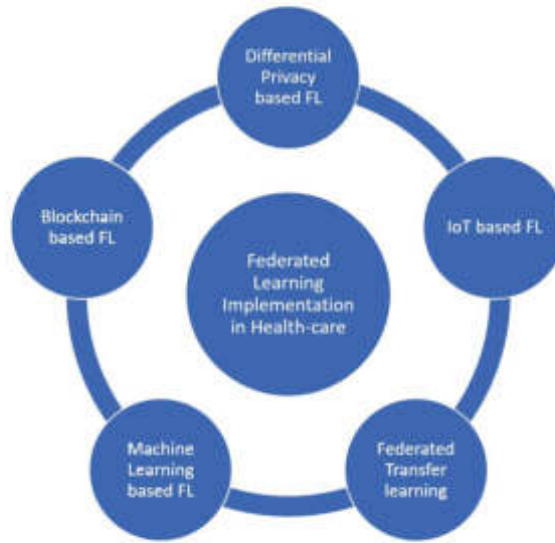


Fig. 1.3: Federated Learning Implementation in Healthcare

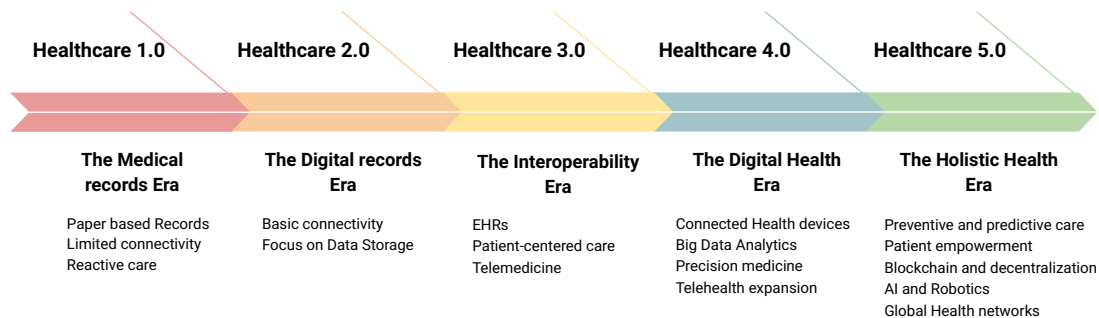


Fig. 2.1: Healthcare Domain Transformation [19]

significant advancements in healthcare are highlighted in this section.

- Healthcare 1.0: Healthcare was first focused on home medicines, traditional healing practices, and a lack of medical expertise. Medical treatments were not standardized, and healthcare was frequently given by local healers and herbal experts.
- Healthcare 2.0: This period saw improvements in medical knowledge, technology, and hospital development. During this time, the standardization of medical education and practices began. The doctor-patient connection was formalized.
- Healthcare 3.0: The digitization of healthcare records and the implementation of electronic health records (EHRs) occurred during this period. The emphasis has shifted to data-driven decision making and evidence-based medicine. Technologies for telemedicine and remote monitoring began to emerge.
- Healthcare 4.0: During this stage, big data, artificial intelligence, and machine learning are being integrated into medical practises. These technologies provide predictive and preventative treatment,

allowing for more personalised and accurate therapy. Because of the growth of wearable technology and sensors, patients may now measure their wellbeing and physical activity.

- **Healthcare 5.0:** Healthcare 5.0 is at the forefront of healthcare change. It focuses on personalized medicine, which involves adapting therapies to individuals' genetic, genomic, and lifestyle characteristics [21]. Advances in genomics, artificial intelligence, and big data analytics are critical in more precisely identifying and treating illnesses. Patient empowerment and active involvement in healthcare decision-making remain critical.

2.2. Federated Learning. Federated Learning, a revolutionary machine learning paradigm, has emerged as a possible answer to these problems in the healthcare industry. In contrast to traditional data aggregation methods, Federated Learning allows for the cooperative training of machine learning models across several institutions or devices while preserving the localization of raw data [24]. Data privacy is met by this dispersed strategy, which also protects patient data and complies with strict legal requirements like HIPAA. Federated Learning offers several benefits, including privacy preservation, reduced communication costs, and improved scalability [25].

- **Privacy preservation:** Since the data remains on the device, it is unnecessary to send it to a central server for processing, ensuring the data's privacy.
- **Reduced communication costs:** Since only model updates are sent to the central server, the communication costs are significantly lower than the traditional machine learning approach.
- **Improved scalability:** Since the computation is distributed among multiple devices, Federated Learning can handle multiple devices without needing extra resources.

The way people control their diabetes has been transformed by Continuous Glucose Monitoring (CGM) technology. For the treatment of diabetes, CGM systems continually assess glucose levels in real time. However, the volume and complexity of CGM data produced by these devices provide major hurdles, particularly in making accurate forecasts and enhancing patient outcomes. An original solution to these problems is the decentralized fusion of CGM data via Federated Learning, which combines the power of CGM data from many sources while maintaining individual privacy and security.

With the help of Federated Learning, CGM data owners may build a global glucose prediction model without disclosing their raw data, protecting their privacy [26]. Local updates from participating CGM devices are combined to train a global model. Diverse data sources are advantageous for this model aggregation. The global model can offer particular people specific insights and suggestions for managing diabetes. The decentralized method improves data security by lowering the possibility of data breaches or illegal access. The use of decentralized CGM data fusion and Federated Learning shows great potential for overcoming the difficulties in adequately maintaining and exploiting CGM data. It makes it possible to create precise, privacy-preserving models that provide people with diabetes the ability to take care of their health while advancing study and treatment in diabetes management. This technology has the potential to fundamentally alter how we perceive and manage diabetes as it develops. Figure 2.2 shows the process of implementing Federated Learning in healthcare.

2.3. Internet of Medical Things (IoMT). The Internet of Medical Things (IoMT) integrates medical devices with the Internet of Things (IoT). IoMT represents the future of modern healthcare systems, where all medical devices will be interconnected and monitored online by healthcare professionals [22]. This development promises to enhance the speed and reduce the costs of healthcare services as it progresses. IoMT enhances the volume of health data accessible to caregivers, diversifies its sources, and accelerates the processes of collection, transmission, and analysis. The increased flow of data aids in improving decision-making for both patients and healthcare providers. This network of interconnected technology allows for the constant collection, analysis, and transmission of health data, enabling real-time monitoring and management of patients' health. IoMT devices encompass wearable fitness trackers, smart implants, remote patient monitoring systems, and connected diagnostic instruments [20]. IoMT streamlines data flow and automates routine tasks, alleviating administrative burdens and enabling medical staff to concentrate more on patient care, thus boosting operational efficiency. It helps lower healthcare costs by decreasing hospital readmissions, reducing the need for in-person visits through remote monitoring, and optimizing resource allocation. IoMT also enhances personalized healthcare through comprehensive data, increases patient engagement with intuitive interfaces, and supports telemedicine, thereby

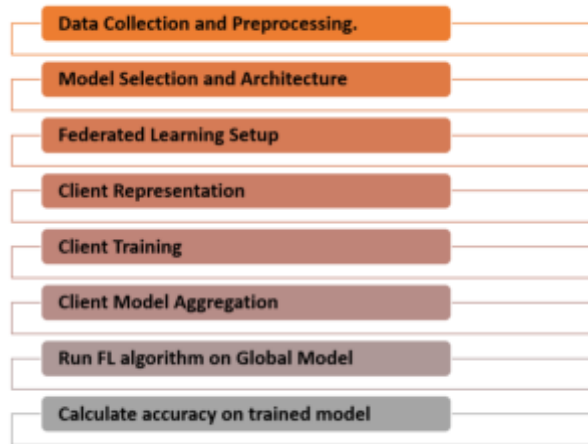


Fig. 2.2: General Implementation Process

improving healthcare accessibility, particularly in underserved regions [23]. Furthermore, IoMT ensures patient safety and adherence to prescribed treatments through reminders and alerts and enhances data analytics, propelling medical research and innovation.

3. Related Work. In the realm of Federated Learning, numerous studies have delved into the healthcare domain. This section reviews prior research on Federated Learning in healthcare, including new frameworks and architectures in this area and existing surveys.

Choudhury et al. [16] applied three classification algorithms—the perceptron, support vector machine (SVM), and logistic regression—that are amenable to distributed solutions using gradient descent for ADR and mortality prediction tasks. The models’ usefulness is assessed using the F1 score before and after applying a privacy-preserving technique. It is demonstrated that FL, even without differential privacy, may deliver model performance near the theoretical scenario of total data centralization. The performance of the developed global FL model for a specific level of privacy is next examined with differential privacy. Contrasting the outcomes with baseline techniques demonstrates how the adaptive contribution weighting strategy enhances learning accuracy and convergence. The tests also validate the privacy-preserving techniques and show the framework’s potential for usage in real-world smart healthcare applications.

Gupta et al. [8] examine using digital twins and hierarchical Federated Learning for anomaly detection in smart healthcare. It emphasizes the value of anomaly detection in healthcare and the difficulties with centralized, conventional methods. To enable data sharing and collaborative model training while preserving privacy, the authors suggest a hierarchical Federated Learning architecture that uses digital twins and virtual representations of actual items. The review gives a general description of the framework, goes through its benefits, and shows experimental findings to show how well it works at spotting abnormalities in healthcare data. The authors also examine possible uses for digital twins and hierarchical Federated Learning and potential future studies in smart healthcare.

Xu et al. [9] thoroughly investigate Federated Learning’s use in healthcare informatics. It discusses Federated Learning’s ideas, architectures, algorithms, and privacy-preserving techniques, among other topics. To illustrate the efficacy and promise of Federated Learning in various fields, the authors look at various healthcare informatics applications, including population health analysis, illness detection, therapy suggestion, and predictive modeling. The paper also emphasizes the significance of model aggregation approaches, data heterogeneity, and data quality in Federated Learning for healthcare. Overall, the literature analysis highlights research gaps, offers insightful information about the current status of Federated Learning in healthcare informatics, and makes recommendations for future developments.

Stripelis et al. [10] emphasize using Federated Learning to scale neuroscience research. Due to privacy issues and restricted access to massive datasets, they discuss the difficulties traditional centralized systems in neuroscience research confront. The paper demonstrates how Federated Learning might facilitate group analysis of dispersed neuroimaging data while protecting individual privacy. The authors also discuss the advantages of Federated Learning, including better data sharing, bigger sample sizes, and greater study reproducibility. Additionally, they give examples of Federated Learning applications in neuroscience, such as cognitive modeling, brain imaging processing, and research into neurodegenerative diseases.

Liu et al. [12] used a traditional method to assess the contributions of members inside a coalition accurately Shapley Value (SV). Furthermore, the coalition's overall utility is to be calculated because FL may use it as a fair contribution evaluation principle. The average of all marginal contributions a participant makes over all coalitional permutations is that individual's SV. To investigate CAREFL's suitability for further intelligent healthcare application situations. Adding contribution-based data pricing mechanisms to the CAREFL framework would help the development of a market for FL-based healthcare data sharing. The ultimate objective is to include these capabilities in the free and open-source FATE framework so that they may be used by more programmers, scholars, and practitioners.

Le Sun and Jin Wu [13] proposed a model that comprises a parameter protection method to prevent catastrophic forgetting, a lightweight 1-D CNN-based feature extractor, and a PCC mechanism to facilitate class scaling. To increase the effectiveness of training models for new tasks, SCALT also implements two significant model transfer methods. The trained SCALT models' mean parameter values from previous tasks are utilized as model initialization for a new task. To eliminate fuzzy boundaries between various classes, a KL-divergence-based sharing knowledge removal technique is used in the transferring process.

Patel et al. [14] suggested FL in healthcare for improved privacy and security of user data. A centralized intelligence system must cope with many constraints, including resource constraints, data update delays, a lack of high precision and accuracy, and privacy and security issues. The suggested FL-EHR case study incorporates differential privacy, which guarantees learning data privacy, which is advantageous in real-world settings. It also eliminates data bias throughout the training phase, increasing the model's accuracy. As a result, the suggested study enhances model validation, the foundation for efficient analytics setups in diverse medical ecosystems compared to present healthcare systems. The FL training scheme calculates the local slopes. Then, using techniques like FedAvg, Fed-Prox, and FedSGD, all local gradients are combined at the aggregation level. Here, a 6G network is being used for communication. Effective optical wireless communication channels in 6G networks can be used to establish direct, close connectivity with healthcare infrastructure. The prediction model is then sent to numerous clients via the aggregation layer using a HE to FL training strategy.

Wu et al. [15] proposed FedHome, a cloud-edge FL system. FedHome employs a novel generative convolutional autoencoder architecture. To compensate for the loss in prediction performance caused by an unbalanced and non-IID distribution of user data, GCAE synthesizes minority class samples and retrains the user's local model using the resultant class-balanced dataset. The Federated Learning framework safeguards data privacy by keeping user data local and learning a common global model in the cloud from several network edge homes. To deal with the unbalanced and non-IID distribution inherent in user monitoring data, a generative convolutional autoencoder (GCAE) is created with the goal of achieving accurate and personalised health monitoring by refining the model with a generated class-balanced dataset from user-specific data.

Lu et al. [16] use a FedAP as a weighted, tailored, batch-normalized federated transfer learning technique for the healthcare industry. FedAP combines data from several companies while maintaining privacy and security, and by taking into account similarities and maintaining local batch normalization, it provides reasonably customized model learning. The statistical mean of the relevant layer's inputs and the running mean of the BN layer are positively correlated. As a result, the researchers may run many rounds of FedBN while maintaining local batch normalization and use the parameters of BN layers to substitute the statistics when a pre-trained model is unavailable. This variety is known as f-FedAP. The authors suggest FedAP, a personalized Federated Learning method for healthcare that uses adaptive batch normalization to collect data from several clients without sacrificing security and privacy and develop customized models for each client.

Islam et al. [17] gave a system for universal data sharing for Federated Learning that uses differentiated privacy to forecast specific heart failure conditions. The system minimizes the total data dimension through

feature selection, lowering the noise addition for differential privacy while keeping effective scores. The authors propose a feature selection strategy based on the correlation value of the data to reduce dimension and boost the utility of the ML models. The authors employ Federated Learning infrastructure to enable private data sharing among collaborating parties in a differentially private manner. The basic structure of this Federated Learning model is comprised of two components. A model manager with whom all data owners communicate and the owner(s) of the raw data used to train the model. Furthermore, the model is protected by a second privacy layer that employs differential privacy, and the authors assume that the aggregator server is truthful yet interested in recovering the raw data of a data owner from it. The same training data was utilized in both Naive Bayes and Random Forest in a two-party configuration where the data is divided into two groups. The data's correlation value serves as the foundation for feature selection. The authors then added noise using Laplace transformation to the statistics they had generated using Differential Privacy (DP) techniques. The generated noisy data are finally sent to the aggregator server in place of the raw data. The central aggregator server feeds the machine learning framework to create a global model that is used to forecast the likelihood of cardiac failure.

Li et al. [5] focus on developing a smart healthcare system that preserves privacy using Federated Learning techniques. The proposed approach aims to solve concerns about the privacy and security of sensitive patient data while utilizing the benefits of machine learning in healthcare. Patient data must frequently be centralized in traditional healthcare systems, raising privacy concerns and the possibility of data breaches. However, Federated Learning allows training models collaboratively without transferring raw data. Multiple edge devices, including wearables and smartphones, are included in the system to gather patient data.

Yang et al. [3] introduced the concept of federated machine learning (FML), and its different applications are explored. A distributed approach to machine learning called FML enables numerous parties to jointly train a single model without disclosing their personal information. The authors outline the FML systems' structure, consisting of a dispersed client base and a central coordination server. The study discusses methods and strategies to handle technological difficulties such as communication effectiveness, privacy protection, and system robustness. In addition, practical uses in healthcare, finance, and smart cities are investigated, demonstrating how FML promotes collaborative learning while protecting data privacy. Overall, the paper illustrates the importance of FML in dealing with privacy issues and thoroughly discusses its idea, design, difficulties, and practical applications.

Abdulrahman et al. [6] provides an in-depth overview of Federated Learning, tracing its development from centralized to distributed on-site learning and exploring its advancements. It covers various aspects of Federated Learning, including architectures, communication protocols, optimization algorithms, privacy-preserving mechanisms, and model aggregation methods. The survey discusses applications in healthcare, IoT, edge computing, and autonomous vehicles, highlighting challenges and open research areas. Additionally, it explores recent advancements such as federated transfer learning and reinforcement learning and emerging trends like edge intelligence and blockchain-based Federated Learning. The paper serves as a comprehensive resource for researchers and practitioners interested in understanding the current state and future directions of Federated Learning.

Prayitno and Shyu [7] thoroughly assess the literature on Federated Learning in the healthcare industry, emphasizing data characteristics and applications. To illustrate the usefulness and promise of Federated Learning in tackling healthcare issues, the authors look at various healthcare applications, including illness prediction, medical picture analysis, electronic health record analysis, and wearable device data analysis. The study intends to present insightful analyses of current research, identify knowledge gaps, and suggest future research avenues in Federated Learning in healthcare.

Table 3.1 summarizes the existing research work on Federated Learning in healthcare. This literature review on Federated Learning in healthcare finds an emerging area with a strong emphasis on protecting patient privacy and exploiting decentralized data for model training. To handle the variety of healthcare data, researchers are building specialized Federated Learning strategies focusing on multiple data types and matching machine learning models. Key use cases like illness prediction, medical picture analysis, and personalized therapy recommendations highlight the potential to improve patient outcomes and healthcare efficiency. However, issues such as communication overhead, model convergence, data quality, scalability, and regulation compliance

Table 3.1: Survey of Existing Research Works

Author	Year	Key Contributions	Evaluation Parameters	Limitations and Future Scope
Dai <i>et al.</i> [31]	2020	Presented a blockchain-enabled framework to address the security and privacy issues with IoMT systems	Security and Privacy	Improving the performance by implementing more scalable consensus algorithms is not easy. Deep learning can be used to enhance the performance
Choudhury <i>et al.</i> [11]	2020	Federated Learning framework with two levels of privacy protection that can learn a global model from dispersed health data.	The global Federated Learning model's performance is then evaluated with respect to differential privacy at a specific privacy level.	Differential privacy hampers Federated Learning's prediction capacity due to excessive noise.
Gupta <i>et al.</i> [8]	2021	A threat model for centralized anomaly detection with particular threat scenarios and suggested a privacy-preserving model.	Accuracy, sensitivity, and efficiency of the anomaly detection system.	Scalability, Data Quality and Algorithmic Challenges.
Xu <i>et al.</i> [9]	2021	Privacy-Preserving Techniques in Healthcare Informatics and Model Aggregation Approaches	Accuracy, privacy preservation, model performance, scalability, and Computational efficiency.	Real-world Deployments and Interoperability
Stripelis <i>et al.</i> [10]	2021	A brain age prediction model based on structural MRI scans from several locations with varying quantities of data and subject (age) distributions.	Elapsed time and a Mean Absolute Error (MAE)	The proposed technique demonstrates distinct relative performance of the multiple training approaches and, in the future, federated transfer learning in neuroimaging might be investigated.
Ali <i>et al.</i> [34]	2022	Presented the role of FL in IoMT for detecting privacy threats	Privacy and Accuracy	Need robust and universal FL architecture from privacy perspective.
Arya <i>et al.</i> [35]	2022	Proposed an ensemble FL approach for training a DL model to classify decentralized data streams in IoMT	Accuracy, Precision, Recall and F1-score	Since ensemble learning is carried out on a server, so there can be a threat for data privacy.
Liu <i>et al.</i> [12]	2022	Adaptive contribution weighting mechanism, privacy-preserving methods, contribution-aware Federated Learning framework.	Accuracy and balance between performance and privacy.	CAREFL's potential for other healthcare applications
Le Sun and Jin Wu [13]	2022	A scalable and transferable classification framework, SCALT to protect privacy of patients data	Measured accuracy of ECG, EEG, and PPG data	SCALT's space complexity can be reduced by minimizing the feature extractor's volume. Sophisticated model transfer methods will enhance new model training.
Patel <i>et al.</i> [14]	2022	FL in HI for enhanced privacy and security and a centralized intelligence systems to overcome resource constraints, data delays, and privacy issues.	The FL-EHR case study incorporates differential privacy for data privacy and reduces bias during training, enhancing model accuracy and validation.	The framework examines FL performance on healthcare datasets with different sensitivity levels. It introduces a DP model for FL-HI, ensuring high secrecy and accurate diagnosis.
Wu <i>et al.</i> [15]	2022	Developed a Generative Convolutional Autoencoder (GCAE) to provide accurate and personalized health monitoring by improving the model using a newly constructed class-balanced dataset.	Local minibatch size, training passes on each client's data; and number of participating clients per communication round.	GCAE's few parameters reduce communication overhead during model transfer.
Lu <i>et al.</i> [16]	2022	FedAP is a weighted, batch-normalized federated transfer learning for healthcare.	BN layer parameters to replace statistics, enabling multiple rounds of FedBN with local batch normalization when a pre-trained model is not available.	Consider implementing more accurate methods for calculating and updating client similarity.
Islam <i>et al.</i> [17]	2022	Feature selection based on correlation values minimizes data dimension, preserving model effectiveness.	Model manager and data owner(s), with no direct message exchange between providers. Naive Bayes and Random Forest were trained.	Data anonymization and differential privacy should be included, and the score should be further enhanced using a better feature selection technique.
Li <i>et al.</i> [5]	2022	A convenient and privacy-preserving system named ADDETECTOR to detect Alzheimer's disease (AD)	Accuracy, Time overhead, F-score	Find additional useful features to reflect the characteristics of Alzheimer's disease and test the viability of ADDETECTOR on a bigger dataset.
Tian <i>et al.</i> [18]	2023	Ring Structure for Federated Learning, Threshold Secret Sharing Mechanism and edge-dropout handling	Accuracy and convergence pace	To boost parallelism to lower the computational cost and communication overhead of the RPDFL training scheme.
Rani <i>et al.</i> [32]	2023	An exhaustive survey on the security of FL-based IoMT applications	Data Security and Privacy	Real world health data analytics with integrity of the data.
Alamleh <i>et al.</i> [33]	2023	Developed an multicriteria decision-making (MCDM) framework for standardising and benchmarking the ML-based IDSs for FL architecture of IoMT applications	Systematic ranking and Computational cost	More investigation on MCDM methods that consider the experts' importance is needed.

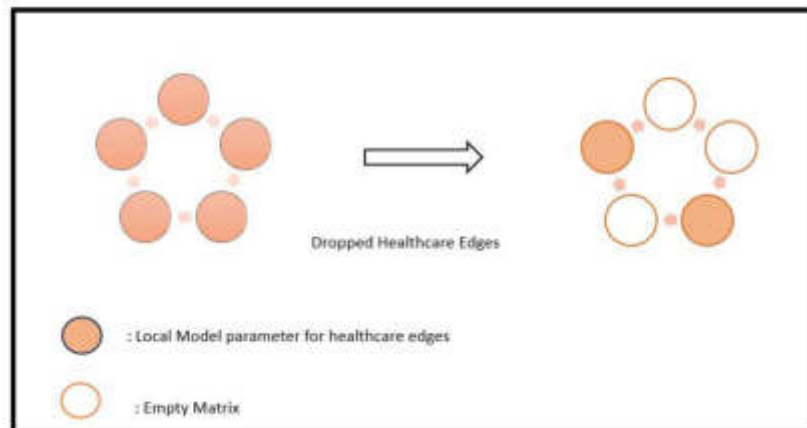


Fig. 4.1: Data sharing procedure of Ring AllReduce

prompt continuous research. Interdisciplinary research, the creation of Federated Learning frameworks, and an increasing interest in scalable and efficient methodologies are all exciting futures for Federated Learning's revolutionary role in healthcare.

4. Case Study. The potential to improve patient care and medical research has been demonstrated by using deep neural networks in digital healthcare applications. Deep learning in healthcare, however, has a set of issues, the main among them is patient data protection. Because healthcare companies are spread, and patient information is sensitive, traditional machine learning algorithms sometimes ask for the centralization of data, which poses challenges. Federated Learning has become recognized as a viable response to these issues. Federated Learning allows training machine learning models without exchanging and centralizing raw data.

In this case study, we have examined a unique strategy called Robust Privacy-Preserving Decentralized Federated Learning (RPDFL) proposed by Tian et al. [18] in the context of digital healthcare. "Data silos," where healthcare institutions keep patient data in separate databases, frequently plague digital healthcare systems. Given that the data is scattered around several institutions, developing comprehensive and efficient machine-learning models is difficult due to these data silos. Data interoperability and cooperation are hampered by this fragmentation, making it challenging to fully utilize deep learning, which benefits from extensive and varied training datasets. Traditional centralized Federated Learning approaches also have their problems. These include communication bottlenecks between clients, the possibility of malicious servers attempting to deduce gradients, and single points of failure in the central server.

By grouping Federated Learning clients into a logical ring structure, RPDFL presents a novel strategy motivated by the widely utilized distributed computing Ring-AllReduce technique. This structure gives Each client a left and a right neighbor node. The fundamental idea is that each client should only communicate information with the proper neighbors to ensure synchronized global model updates and reduce vulnerability to malicious servers.

The strong privacy protection system of RPDFL is one of its key characteristics. It uses a threshold secret sharing mechanism, a cryptographic method to protect gradient privacy. This method maintains data security even if healthcare edge nodes gracefully leave during Federated Learning training. Clients exchange gradients from their local models. This sharing mechanism must preserve the privacy of each client's gradients. The remaining clients can complete the program uninterrupted if a client withdraws or disconnects, providing continual instruction. This privacy protection measure is essential to guarantee that sensitive patient data is kept private throughout the Federated Learning procedure.

Using their local datasets, RPDFL edge nodes in the healthcare industry produce local models across a number of training rounds. Stochastic Gradient Descent (SGD) is used to update these local models under the direction of the acquired global model. The Ring-Allreduce technique is used to broadcast gradients, which

stand in for the updates needed to the model parameters. Because of the threshold secret sharing technique, privacy is maintained during the whole procedure. Gradient averaging is used to produce new global models once gradients have been compiled and global model modifications have been made. For the upcoming training cycle, these upgraded global models will take the place of the local models. Up till the target level of model performance is attained, this iterative procedure is continued. Gradient averaging is used to generate new global models once gradients have been accumulated and global models have been modified. These improved global models will replace the local models in the future training cycle. This iterative method is repeated until the desired level of model performance is achieved.

For the performance evaluation of RPDFL on healthcare data, the UCI-HAR dataset from UCI was used, which captures daily human activities via smartphones and was evaluated using a fully connected network architecture. This capability enables RPDFL to effectively overcome information barriers between healthcare organizations, facilitating highly efficient decision-making in complex scenarios.

Thus, a new data-sharing scheme was implemented by updating the execution processes of the CRT in the threshold secret sharing scheme, allowing healthcare edges to drop out during training without causing data leakage and ensuring the robustness of the RPDFL training. The RPDFL scheme also supports edge dropout during the training process while preserving local gradient privacy. Security analysis demonstrates that RPDFL is highly secure under the HbC security model.

RPDFL exhibits promising performance in terms of privacy protection and accuracy. RPDFL obtains an accuracy level of around 98% by the fifth training round, which is an impressive accomplishment given its emphasis on privacy protection. Despite slower convergence after this first phase, RPDFL's accuracy is nevertheless on par with the conventional federated averaging (FedAvg) method. It demonstrates how RPDFL can maintain good performance while guarding against gradient privacy leaks. RPDFL has had a thorough security evaluation and has been shown to be extremely secure. According to security studies, RPDFL is resilient against harmful attacks and data breaches, protecting patient information's privacy and confidentiality.

RPDFL is a revolutionary technique for addressing the critical concerns of privacy and scalability in digital healthcare systems. RPDFL integrates a novel ring Federated Learning structure, a Ring-AllReduce-based data sharing mechanism, and a robust threshold secret sharing mechanism to provide a secure and effective solution for Federated Learning in the healthcare business. Figure 4.1 shows the data sharing process of Ring-AllReduce. Because of its extensive security features and promising performance, it is an excellent choice for healthcare organizations wishing to utilize the potential of deep learning while preserving patient privacy. As digital healthcare evolves, RPDFL is a valuable tool for harnessing the potential of machine learning in improving patient care and promoting medical research.

Then the performance of RPDFL in comparison to FedAvg and Gossip Learning is evaluated. Based on available information, Gossip Learning is the first implementation of decentralized learning. However, it is important to note that neither FedAvg nor Gossip Learning consider the privacy of gradients.

The training loss and testing accuracy for RPDFL, FedAvg, and Gossip Learning were recorded. Since FedAvg and Gossip Learning do not support edge dropout during the training process, RPDFL was also evaluated without edge dropout for comparison. It was observed that the accuracy of the RPDFL training model significantly surpasses that of Gossip Learning and is comparable to that of FedAvg. Additionally, RPDFL utilizes an enhanced CRT-based threshold secret sharing protocol to ensure gradient privacy. This protocol involves a truncation process, which results in some accuracy loss and a slower convergence speed. For instance, RPDFL converges slightly slower than FedAvg, but achieves an accuracy of 98% by the fifth round and maintains similar accuracy levels thereafter. Consequently, RPDFL offers excellent efficiency while safeguarding gradient privacy. This ensures that RPDFL can prevent malicious users from inferring the privacy of others in practical applications.

5. Open Research Challenges and Future Research Directions. In the rapidly evolving world of healthcare technology, data holds the key to ground-breaking advancements. This paper explores the dynamic world of Federated Learning in the healthcare domain, revealing the unique challenges posed by the distributed nature of medical data.

5.1. Non-IID Data Management. Data's inherent non-IID (Non-Independent and Identically Distributed) nature across various healthcare organizations presents substantial hurdles for Federated Learning in

the healthcare industry. Healthcare data is frequently gathered from several sources with variances in patient demographics, illness incidence, and treatment procedures, unlike typical centralized machine learning, where data is homogeneous and evenly disseminated. Using common Federated Learning techniques is challenging due to this heterogeneity directly. Innovative approaches like federated transfer learning and domain adaptation are being explored to tackle this issue [27]. These methods seek to account for the differences in data distributions while adapting and transferring information gained from one institution's data to another. Federated transfer learning makes it possible to share knowledge effectively by utilizing pre-trained models and fine-tuning them using data relevant to the institution.

5.2. Privacy-Preservation. The delicate nature of patient data needs sophisticated privacy-preserving strategies in Federated Learning since privacy is of the utmost significance in the healthcare industry. Traditional Federated Learning guarantees that data stays local to the institutions and is decentralized, but more developments are needed to improve privacy [28]. Differential privacy, a potential method, adds controlled noise to the model updates to secure patient information while allowing for useful learning. Homomorphic encryption is one of the most sophisticated encryption methods that can contribute to the protection of data throughout transmission and aggregation procedures. By implementing these privacy-enhancing strategies, Federated Learning may preserve patient confidentiality and adhere to stringent data protection laws.

5.3. Cross-Modal Data Fusion. The integration of multi-modal data develops as a crucial endeavor in the scene of healthcare improvement. It requires integrating many data types, from electronic health records to complex medical images, while respecting the fundamental principles of privacy protection. Modern approaches like federated transfer learning and meta-learning are simultaneously integrated, taking center stage [29]. By utilizing these methods to their full potential, the healthcare industry benefits from quick and seamless knowledge transfer, effective information translation and application across various contexts and activities, and strategic sharing of insights and expertise across institutions.

5.4. Interoperability and Data Harmonization. Integrating interoperability and data harmonization techniques is a crucial requirement in the healthcare field. These methods are essential for reducing inequities from disparate data formats and semantic details used by various healthcare facilities. Addressing the requirements for explainability and interpretability in the context of Federated Learning simultaneously becomes crucial. It requires developing models that produce reliable outcomes and foster mutual understanding and trust between patients and medical professionals. Model interpretability is a goal that has increased importance since it facilitates a broader understanding and acceptance of the results of Federated Learning projects.

5.5. Resource Management and Communication Efficiency. The successful deployment of Federated Learning in the healthcare sector depends on the effective allocation and use of resources. Notably, the need for effective communication is emphasized. It calls for developing novel strategies like decentralized aggregation and compressed model updates, which are key to reducing bandwidth requirements and boosting communication effectiveness. Healthcare systems can balance resource conservation and the seamless flow of vital information by carefully improving these strategies, eventually paving the way for a strong and effective Federated Learning framework.

The advancement of Federated Learning in healthcare, overcoming difficulties and supporting its effective application for improved patient care and medical insights [30]. By focusing on these future research routes, Federated Learning in healthcare may evolve further and contribute to better patient outcomes, customized treatment, and collaborative medical research as long as privacy, security, and ethical considerations are addressed.

6. Conclusion. Federated Learning in the healthcare industry has enormous potential to improve clinical decision-making, medical research, and customized treatment. Federated Learning makes it possible to use machine learning on decentralized healthcare data while protecting patient privacy by utilizing the strength of distributed computing and safe data sharing. This article has offered a comprehensive review of the current trends and issues facing the field of Federated Learning in healthcare. We explored how Federated Learning has emerged as a potential method for leveraging decentralized healthcare data while maintaining privacy and security. A case study on privacy preservation is presented in the context of digital healthcare. We have

also covered the key challenges that must be overcome, such as non-IID data management, interoperability, communication efficiency, and multi-model data integration. Federated Learning has a promising future in the medical field. Medical professionals and researchers should keep working collectively to create novel approaches, improve current techniques, and solve the legal and ethical challenges involving the sharing of medical data. To guarantee that Federated Learning helps the healthcare business and the persons it serves, it is critical to prioritize its appropriate and ethical deployment.

REFERENCES

- [1] Choudhury, O., Gkoulalas-Divanis, A., Saloniadis, T., Sylla, I., Park, Y., Hsu, G. & Das, A. Differential Privacy-enabled Federated Learning for Sensitive Health Data. (2020)
- [2] Shailaja, K., Seetharamulu, B. & Jabbar, M. Machine Learning in Healthcare: A Review. *2018 Second International Conference On Electronics, Communication And Aerospace Technology (ICECA)*. pp. 910-914 (2018)
- [3] Yang, Q., Liu, Y., Chen, T. & Tong, Y. Federated Machine Learning: Concept and Applications. (Association for Computing Machinery, 2019), <https://doi.org/10.1145/3298981>
- [4] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H., Albarqouni, S., Bakas, S., Galtier, M., Landman, B., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R., Trask, A., Xu, D., Baust, M. & Cardoso, M. The future of digital health with federated learning. *Npj Digital Medicine*. **3** (2020,12)
- [5] Li, J., Meng, Y., Ma, L., Du, S., Zhu, H., Pei, Q. & Shen, X. A Federated Learning Based Privacy-Preserving Smart Healthcare System. *IEEE Transactions On Industrial Informatics*. **18**, 2021-2031 (2022)
- [6] Abdulrahman, S., Tout, H., Ould-Slimane, H., Mourad, A., Talhi, C. & Guizani, M. A Survey on Federated Learning: The Journey From Centralized to Distributed On-Site Learning and Beyond. *IEEE Internet Of Things Journal*. **PP** (2020,10)
- [7] Prayitno, Shyu, C., Putra, K., Chen, H., Tsai, Y., Hossain, K., Jiang, W. & Shae, Z. A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications. *Applied Sciences*. **11** (2021), <https://www.mdpi.com/2076-3417/11/23/11191>
- [8] Gupta, D., Kayode, O., Bhatt, S., Gupta, M. & Tosun, A. Hierarchical Federated Learning based Anomaly Detection using Digital Twins for Smart Healthcare. (2021,11)
- [9] Xu, J., Glicksberg, B., Su, C., Walker, P., Bian, J. & Wang, F. Federated Learning for Healthcare Informatics. *Journal Of Healthcare Informatics Research*. **5** pp. 1-19 (2021,3)
- [10] Stripelis, D., Ambite, J., Lam, P. & Thompson, P. Scaling Neuroscience Research using Federated Learning. (2021,2)
- [11] Choudhury, O., Gkoulalas-Divanis, A., Saloniadis, T., Sylla, I., Park, Y., Hsu, G. & Das, A. Differential Privacy-enabled Federated Learning for Sensitive Health Data. *ArXiv*. **abs/1910.02578** (2019)
- [12] Liu, Z., Chen, Y., Zhao, Y., Yu, H., Liu, Y., Bao, R., Jiang, J., Nie, Z., Xu, Q. & Yang, Q. Contribution-aware federated learning for smart healthcare. *Proceedings Of The AAAI Conference On Artificial Intelligence*. **36**, 12396-12404 (2022)
- [13] Sun, L. & Wu, J. A Scalable and Transferable Federated Learning System for Classifying Healthcare Sensor Data. *IEEE Journal Of Biomedical And Health Informatics*. **27**, 866-877 (2023)
- [14] Patel, V., Bhattacharya, P., Tanwar, S., Gupta, R., Sharma, G., Bokoro, P. & Sharma, R. Adoption of federated learning for healthcare informatics: Emerging applications and future direction. *IEEE Access*. (2022)
- [15] Wu, Q., Chen, X., Zhou, Z. & Zhang, J. FedHome: Cloud-Edge Based Personalized Federated Learning for In-Home Health Monitoring. *IEEE Transactions On Mobile Computing*. **21**, 2818-2832 (2022,8)
- [16] Lu, W., Wang, J., Chen, Y., Qin, X., Xu, R., Dimitriadis, D. & Qin, T. Personalized federated learning with adaptive batchnorm for healthcare. *IEEE Transactions On Big Data*. (2022)
- [17] Islam, T., Ghasemi, R. & Mohammed, N. Privacy-preserving federated learning model for healthcare data. *2022 IEEE 12th Annual Computing And Communication Workshop And Conference (CCWC)*. pp. 0281-0287 (2022)
- [18] Tian, Y., Wang, S., Xiong, J., Bi, R., Zhou, Z. & Bhuiyan, M. Robust and Privacy-Preserving Decentralized Deep Federated Learning Training: Focusing on Digital Healthcare Applications. *IEEE/ACM Transactions On Computational Biology And Bioinformatics*. pp. 1-12 (2023)
- [19] Kaur, J., Verma, R., Alharbe, N., Agrawal, A. & Khan, P. Importance of Fog Computing in Healthcare 4.0. (2020,8)
- [20] Razdan, S., & Sharma, S. (2022). Internet of Medical Things (IoMT): Overview, Emerging Technologies, and Case Studies. *IETE Technical Review*, 39(4), 775–788. Taylor & Francis. doi:10.1080/02564602.2021.1927863.
- [21] Mohanta, B., Das, P. & Patnaik, S. Healthcare 5.0: A Paradigm Shift in Digital Healthcare System Using Artificial Intelligence, IOT and 5G Communication. *2019 International Conference On Applied Machine Learning (ICAML)*. pp. 191-196 (2019)
- [22] Vishnu, S., Ramson, S. & Jegan, R. Internet of Medical Things (IoMT) - An overview. *2020 5th International Conference On Devices, Circuits And Systems (ICDCS)*. pp. 101-104 (2020)
- [23] Pournik, O., Ghalichi, L., Gallos, P. & Arvanitis, T. The Internet of Medical Things: Opportunities, Benefits, Challenges and Concerns. *Studies In Health Technology And Informatics*. **309** (2023,10)
- [24] Dhade, P. & Shirke, P. Federated Learning for Healthcare: A Comprehensive Review. *Engineering Proceedings*. **59** (2023), <https://www.mdpi.com/2673-4591/59/1/230>
- [25] Prayitno, Shyu, C., Putra, K., Chen, H., Tsai, Y., Hossain, K., Jiang, W. & Shae, Z. A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications. *Applied Sciences*. **11** (2021), <https://www.mdpi.com/2076-3417/11/23/11191>
- [26] De Falco, I., Della Cioppa, A., Koutny, T., Ubl, M., Krcma, M., Scafuri, U. & Tarantino, E. A Federated Learning-

- Inspired Evolutionary Algorithm: Application to Glucose Prediction. *Sensors*. **23** (2023), <https://www.mdpi.com/1424-8220/23/6/2957>
- [27] Lu, Z., Pan, H., Dai, Y., Si, X. & Zhang, Y. Federated Learning With Non-IID Data: A Survey. *IEEE Internet Of Things Journal*. **11**, 19188-19209 (2024)
- [28] Maurya, J. & Prakash, S. Privacy Preservation in Federated Learning: its Attacks and Defenses. *2023 3rd International Conference On Pervasive Computing And Social Networking (ICPCSN)*. pp. 1042-1047 (2023)
- [29] Ahmed, S., Alam, M., Afrin, S., Raza, S., Raza, N. & Gandomi, A. Insights into Internet of Medical Things (IoMT): Data fusion, security issues and potential solutions. *Information Fusion*. **102** pp. 102060 (2024), <https://www.sciencedirect.com/science/article/pii/S1566253523003767>
- [30] Muazu, T., Yingchi, M., Muhammad, A., Ibrahim, M., Samuel, O. & Tiwari, P. IoMT: A Medical Resource Management System Using Edge Empowered Blockchain Federated Learning. *IEEE Transactions On Network And Service Management*. **21**, 517-534 (2024)
- [31] Dai, H., Imran, M. & Haider, N. Blockchain-Enabled Internet of Medical Things to Combat COVID-19. *IEEE Internet Of Things Magazine*. **3**, 52-57 (2020)
- [32] Rani, S., Kataria, A., Kumar, S. & Tiwari, P. Federated learning for secure IoMT-applications in smart healthcare systems: A comprehensive review. *Knowledge-Based Systems*. **274** pp. 110658 (2023), <https://www.sciencedirect.com/science/article/pii/S0950705123004082>
- [33] Alamleh, A., Albahri, O., Zaidan, A., Albahri, A., Alamoody, A., Zaidan, B., Qahtan, S., Alsatat, H., Al-Samarraay, M. & Jasim, A. Federated Learning for IoMT Applications: A Standardization and Benchmarking Framework of Intrusion Detection Systems. *IEEE Journal Of Biomedical And Health Informatics*. **27**, 878-887 (2023)
- [34] Ali, M., Tariq, M., Naem, F. & Kaddoum, G. Federated Learning for Privacy Preservation in Smart Healthcare Systems: A Comprehensive Survey. *IEEE Journal Of Biomedical And Health Informatics*. **PP** (2022,6)
- [35] Arya, M. & G, H. Ensemble Federated Learning for Classifying IoMT Data Streams. *2022 IEEE 7th International Conference For Convergence In Technology (I2CT)*. pp. 1-5 (2022)
- [36] Lin, H., Han, J., Wu, P., Zhu, L., Wang, J. & Tu, J. Machine Learning and Human-Machine Trust in Healthcare: A Systematic Survey. *SSRN Electronic Journal*. (2023), <https://api.semanticscholar.org/CorpusID:257704958>
- [37] Durga, S., Nag, R. & Daniel, E. Survey on Machine Learning and Deep Learning Algorithms used in Internet of Things (IoT) Healthcare. *2019 3rd International Conference On Computing Methodologies And Communication (ICCMC)*. pp. 1018-1022 (2019)
- [38] Al-Dhief, F., Latiff, N., Malik, N., Salim, N., Baki, M., Albadr, M. & Mohammed, M. A Survey of Voice Pathology Surveillance Systems Based on Internet of Things and Machine Learning Algorithms. *IEEE Access*. **8** pp. 64514-64533 (2020)

Edited by: Katarzyna Wasielewska-Michniewska

Review paper

Received: Nov 23, 2023

Accepted: Jul 29, 2024

AIMS AND SCOPE

The area of scalable computing has matured and reached a point where new issues and trends require a professional forum. SCPE will provide this avenue by publishing original refereed papers that address the present as well as the future of parallel and distributed computing. The journal will focus on algorithm development, implementation and execution on real-world parallel architectures, and application of parallel and distributed computing to the solution of real-life problems. Of particular interest are:

Expressiveness:

- high level languages,
- object oriented techniques,
- compiler technology for parallel computing,
- implementation techniques and their efficiency.

System engineering:

- programming environments,
- debugging tools,
- software libraries.

Performance:

- performance measurement: metrics, evaluation, visualization,
- performance improvement: resource allocation and scheduling, I/O, network throughput.

Applications:

- database,
- control systems,
- embedded systems,
- fault tolerance,
- industrial and business,
- real-time,
- scientific computing,
- visualization.

Future:

- limitations of current approaches,
- engineering trends and their consequences,
- novel parallel architectures.

Taking into account the extremely rapid pace of changes in the field SCPE is committed to fast turnaround of papers and a short publication time of accepted papers.

INSTRUCTIONS FOR CONTRIBUTORS

Proposals of Special Issues should be submitted to the editor-in-chief.

The language of the journal is English. SCPE publishes three categories of papers: overview papers, research papers and short communications. Electronic submissions are preferred. Overview papers and short communications should be submitted to the editor-in-chief. Research papers should be submitted to the editor whose research interests match the subject of the paper most closely. The list of editors' research interests can be found at the journal WWW site (<http://www.scpe.org>). Each paper appropriate to the journal will be refereed by a minimum of two referees.

There is no a priori limit on the length of overview papers. Research papers should be limited to approximately 20 pages, while short communications should not exceed 5 pages. A 50–100 word abstract should be included.

Upon acceptance the authors will be asked to transfer copyright of the article to the publisher. The authors will be required to prepare the text in $\text{\LaTeX} 2_{\epsilon}$ using the journal document class file (based on the SIAM's `siamltex.clo` document class, available at the journal WWW site). Figures must be prepared in encapsulated PostScript and appropriately incorporated into the text. The bibliography should be formatted using the SIAM convention. Detailed instructions for the Authors are available on the SCPE WWW site at <http://www.scpe.org>.

Contributions are accepted for review on the understanding that the same work has not been published and that it is not being considered for publication elsewhere. Technical reports can be submitted. Substantially revised versions of papers published in not easily accessible conference proceedings can also be submitted. The editor-in-chief should be notified at the time of submission and the author is responsible for obtaining the necessary copyright releases for all copyrighted material.