Scalable Computing: Practice and Experience

Scientific International Journal for Parallel and Distributed Computing

ISSN: 1895-1767

Universitatea de Vest
din Timişoara

Volume 25(3)

May 2024

SUBSCRIPTION INFORMATION: please visit `http://www.scpe.org`

Scalable Computing: Practice and Experience

Volume 25, Number 3, May 2024

TABLE OF CONTENTS

Papers in the special issue on Intelligent Cloud Technologies Enabled Solutions for Next Generation Smart Cities:

Hazardous Chemicals Logistics Internet of Vehicles Based on Encryption Algorithm **1287**
Yujian Tang, Ya Chen, Hoekyung Jung

Security Situation Awareness System Based on Artificial Intelligence **1301**
Hao Wu

Network-based Mechanical Vibration Fault Diagnosis System **1311**
Qiuchen Zhang, Xiaoxia Jin

Papers in the special issue on Next Generation Pervasive Reconfigurable Computing for High Performance Real Time Applications :

Minimizing Overhead through Blockchain for Establishing a Secure Smart City with IoT Model **1321**
Zhixiong Xiao

Target Image Processing based on Super-resolution Reconstruction and Machine Learning Algorithm **1332**
Chunmao Liu

Performance Evaluation of Micro Automatic Pressure Measurement Sensor for Enhanced Accuracy **1341**
Shuiquan Zhu

Papers in the special issue on Scalable Machine Learning for Health Care: Innovations and Applications :

Use of Topic Analysis for Enhancing Healthcare Technologies **1350**
Usha Patel, Preeti Kathiria, Chand Sahil Mansuri, Shriya Madhvani, Viranchi Parikh

Construction of an Intelligent Identification Model for Drugs in Near Infrared Spectroscopy and Research on Drog Classification based on Improved Deep Algorithms **1361**

Jiulin Xia

Performance Comparison of Apache Spark and Hadoop for Machine Learning based iterative GBTR on HIGGS and Covid-19 Datasets **1373**

Piyush Sewal, Hari Singh

Classification of Covid-19 using Differential Evolution Chaotic Whale Optimization based Convolutional Neural Network **1387**

D.P. Manoj Kumar, Sujata N Patil, Parameshachari Bidare Divakarachari, Przemysław Falkowski-Gilski, R. Suganthi

A Survey on AI-based Parkinson Disease Detection: Taxonomy, Case Study, and Research Challenges **1402**

Shivani Desai, Devam Patel, Kaju Patel, Alay Patel, Nilesh Kumar Jadav, Sudeep Tanwar, Hitesh Chhikaniwala

EHealth Innovation for Chronic Obstructive Pulmonary Disease: A Context-Aware Comprehensive Framework **1424**

Anam Iqbal, Shaima Qureshi, Mohammad Ahsan Chishti

Papers in the special issue on Machine Learning and Block-chain based Solution for Privacy and Access Control in IoT :

Research on Integrating Blockchain and Machine Learning LPP Algorithm in Online Education Platform Under COVID-19 Environment **1442**

Daojun Wang, Meishu Wang, Xinli Xing

Research on the Recommendation System of Music e-learning Resources with Blockchain based on Hybrid Deep Learning Model **1455**

Shasha Jin, Lei Zhang

Enhancing IoT Security in Russian Language Teaching: A Improved BPNN and Blockchain-Based Approach for Privacy and Access Control **1466**

Qi Jia

Blockchain Enhanced Student Physical Performance Analysis using Machine Learning-IoT and Apriori Algorithm in Physical Education Network Teaching **1478**

Jianing Li, Zheping Quan, Weijia Song

Papers in the special issue on Deep Learning-Based Advanced Research Trends in Scalable Computing :

Dynamic Scheduling of Multi-agent Electromechanical Production Lines based on Iterative Algorithms **1491**

Lulu Yuan

Application of Deep Learning Algorithm in Optimization of Engineering Intelligent Management Control **1499**

Shuai Zhou, Jing Guo

A Multi-level Power Grid Enhanced Identity Authentication Data Management Platform based on Filtering Algorithms **1508**

Caiqin Nong, Hong Zou, Jiafa Zhang, Jiahao Liu, Pengfei Shao, Jiawei Jiang, Jiao Xie

Energy Optimization of the Multi-objective Control System for Pure Electric Vehicles based on Deep Learning **1517**

Bubo Zhu

Optimization of Unmanned Aerial Vehicle Flight Control Sensor Control System based on Deep Learning Model **1524**

Ji Liu

Big Data Analysis and Deep Learning Optimization in Artificial Intelligence Production of Information Enterprises **1533**

Na Gao, Qiuling Lu

An Intelligent Network Method for Analyzing Corporate Consumer Repurchase Behavior Using Deep Learning Neural Networks **1541**

Qiuping Lu

Application of Closed-loop Theory in Deep Learning Training Guided by High-strength Intelligent Machinery **1549**

Erfu Guo

Deep Machine Learning-Based Analysis for Intelligent Phonetic Language Recognition **1557**

Yumei Liu, Qiang Luo

Papers in the special issue on Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions:

Design and Implementation of Chinese Language Teaching System Based on Virtual Reality Technology **1564**

Tianya Yang, Jialing Wu

A Parameter Assessment of Teaching Quality Indicators Based on Data Class Mining Fuzzy K-Mean Type Clustering **1578**

Xinhua Huang, Yuzudi Tong

Prediction of English Teacher Career Development based on Data Mining and Time Series Model **1588**

Liping Fan

Research on the Construction of a Higher Education Knowledge Management Model Based on the Integration of Shadow Teams **1603**

Tian Xia

Intelligent Classroom Note-taking Application Software with Higher Performance **1616**

Li Zhuang

Mobile Smart App and its Application in Improving the Efficiency of English Homework Correction **1631**

Huiying Shao, Zan Liu

Papers in the special issue on Scalable Dew Computing for Future Generation IoT Systems :

A Lightweight Symmetric Cryptography based User Authentication Protocol for IoT based Applications **1647**

A. Mahesh Reddy, Durvasi Gudivada, M. Kameswara Rao

Scalable Video Fidelity Enhancement: Leveraging the SOTA AI Models **1658**

Ankit Das, Deven Prakash Paramaj, Shambhavi BR

Optimised ResNet50 for Multi-Class Classification of Brain Tumors **1667**

A Sravanthi Peddinti, Suman Maloji

Localization of Dielectric Anomalies with Multi-level Outlier Detection through Membership Function and Ensemble Classification Framework **1681**

Md. Najumunnisa, ASCS Sastry, B T P Madhav

SCHeMoS – Smart Cow Health Monitoring System: An IoT based Cow Hoof Detection and Healthcare Alert System by Using LSTM Network **1702**

Durairaj K, Dhilip Kumar V, Kanagachidambaresan GR

Performance Analysis for Optimized Light Weight CNN Model for Leukemia Detection and Classification using Microscopic Blood Smear images **1716**

Mahmoud Saed Alkhouli, Hiren Joshi

Papers in the special issue on Graph Powered Big Aerospace Data Processing :

Knowledge Graph Analysis for Chronic Diseases Nursing based on Visualization Technology and Literature Big Data **1728**

Siyu Duan, Yang Zhao

Remote Intelligent Medical Monitoring Data Transmission Network Optimization based on Deep Learning **1748**

Run Wang

Design of Smart Home System based on Wireless Sensor Network Link Status Awareness Algorithm **1757**

Rong Xu

Design of Smart Home System based on Wireless Sensor Network Link Status Awareness Algorithm **1765**

Zhiying Fan

Uncrewed Boat Path Planning Algorithm based on Evolutionary Potential Field Model in Dense Obstacle Environment **1773**

Wei Zheng, Xin Huang

Research on Inventory Control Method Based on Demand Response of Power Big Data **1781**

Huixuan Shi, Zhengping Gao, Li Fang, Jiqing Zhai, Hongzhi Sun

Application of Sports Video Image Analysis Based on Fuzzy Support Vector Machine **1790**

Licheng Gao, Yawen Zhao

A Multi-sentence Music Humming Retrieval Algorithm Based on Relative Features and Deep Learning **1799**

Yelin Zhang

Instinct, Suprresion and Catharsis: The Psychological Source and Guidance of Teenagers' Network Ideology in the Era of Big Data **1807**

Jiao Chen, Feng Du

Application of Big Data Analysis and Intelligent Algorithm in Power System Operation Optimization **1818**

Huichao Jin, Junyi Huo, Qingfen Wang, Dexiong Li

Text Classification and Cluster Analysis based on Deep Learning and Natural Language Processing **1826**

Hua Huang

Research on High-performance Computing Network Search System based on Computer Big Data **1833**

Xiaogang Chen, Dongmei Liu

Database Access Information Security Management Simulation under Big Data Platform **1841**

Zhaocui Li, Dan Wang

Database Access Information Security Management Simulation under Big Data Platform **1852**

Yi Liu, Yi Zhang

Optimization of Internal Control for Budget Operations in Public Institutions Based on Random Forest Algorithm **1863**

Yang Jin

Research and Design of an Automated Security Event Analysis and Handling Framework Based on Threat Intelligence **1872**

Linjiang Xie, Zhouyuan Liao, Hanruo Li

Application of Intelligent Algorithms and Big Data Analysis in Film and Television Creation **1882**

Weiwei Wu

Concept and Application of Industrial Heritage Planning Decision in Big Data Era **1894**

Xiaojuan Han, Xuange Zhu, Min Xie, Hui Chen

Electric Energy Metering Error Evaluation Method Based on Deep Learning **1904**

Tianfu Huang, Zhiwu Wu, Wen Zhan, Chunguang Wang, Tongyao Lin

Research on Data Mining and Reinforcement Learning in Recommendation Systems **1914**

Yueran Zhao, Huiyan Zhao

Context-Aware Music Recommendation Algorithm Combining Classification and Collaborative Filtering **1923**

Xiaoling Wu, Guodong Sun

High-Performance Computing Web Search System Based on Computer Big Data **1932**

Yingxi Kang, Beiping Tang, Xiaodong Hu

Research on Dynamic Optimization Algorithm of Warehousing Location Layout Based on Nonlinear Optimization **1940**

Guang Chen, Sheng Zhang, Jing Fang, Fan She

Papers in the special issue on Soft Computing and Artificial Intelligence for wire/wireless Human-Machine Interface:

BlockFog: A Blockchain-based Framework for Intrusion Defense in IOT Fog Computing **1950**

VG Prasuna, B. Ravindra Babu, Bhasha Pydala

Enhanced Feature-driven Multi-objective Learning for Optimal Cloud Resource Allocation **1963**

Uma Maheswara Rao I, Jkr Sastry

Optimal Feature Selection from High-dimensional Fusion of Blood Smear Images for Leukemia Diagnosis **1980**

G. Chinna Pullaiah, P.M. Ashok Kumar

Master Data Quality Management Framework: Content Validity **2001**

Azira Ibrahim, Ibrahim Mohamed, Nurhizam Safie Mohd Satar, Mohammad Kamrul Hasan

Papers in the special issue on Data-Driven Optimization Algorithms for Sustainable and Smart City :

Enterprise Audit Risk Assessment and Prevention based on AHP Analysis **2013**

Guoliang Sun, Bayi Guan

Economic Dispatch of Multi Regional Power Systems Based on CMOPSO Algorithm **2021**

Jinhua Guo

Product Design for the Elderly Based on Human-Computer Interaction in the Era of Big Data **2034**

Zhihong Liu

Detection Method of Tourist Flow in Scenic Spots based on Kalman Filter Prediction **2048**

Xiaoyan Xu, Li Zhang

ARIMA-BPNN Based Stock Price Prediction Model Based on Fusion News Sentiment Analysis **2062**

XiaoZhe Gong

Text Emotion Classification System Integrating Visual Communication and Deep Learning for Social Platform **2076**

Yan Liu

Design and Application of Automatic English Translation Grammar Error Detection System based on BERT Machine Vision **2088**

Yu Qing

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1287–1300, DOI 10.12694/scpe.v25i3.2064

HAZARDOUS CHEMICALS LOGISTICS INTERNET OF VEHICLES BASED ON ENCRYPTION ALGORITHM

YUJIAN TANG[*], YA CHEN[†], AND HOEKYUNG JUNG[‡]

Abstract. With the advancement of information and communication technology, vehicles role in people's lives is not just for transportation but also equipment for the carrier of mobile communication. The new direction is provided by the Internet for the automobiles development and upgrading. Internet of Vehicles (IoVs) includes smart vehicles, Autonomous Vehicles (AVs) as well as roadside units (RSUs) that communicate for providing the enhanced transportation services such as high traffic efficiency and reduced congestion and accidents. In the hazardous chemical logistics sector, the integration of Internet of Vehicles (IoVs) introduces significant security, privacy, and trust challenges. Vulnerabilities to cyberattacks, such as hacking and data manipulation, threaten the integrity of sensitive information regarding hazardous cargo, while concerns about location privacy and data minimization arise due to the tracking of vehicle movements. Ensuring authentication, authorization, and compliance with regulatory standards is essential for building trust within the IoV ecosystem, as any compromise in supply chain integrity could lead to safety hazards or legal liabilities. Robust encryption algorithms like AES and RSA play a crucial role in securing data transmission, but proper implementation and key management practices are necessary to prevent cryptographic weaknesses. Addressing these issues comprehensively is vital for safeguarding the transportation of hazardous chemicals and ensuring the safety of both the environment and the public. IoVs, suffer from security, privacy and trust issues. In order to solve the problem of hazardous chemical logistics of encryption algorithm, the author proposed a research on the information security of the Internet of Vehicles. Firstly, the information characteristics of dangerous chemicals vehicles are analyzed, then analyze the Data Encryption Standard (DES) algorithm. Advanced Encryption Standard (AES) algorithm is analyzed in the symmetric cryptosystem and then RSA algorithm is analyzed in the public key cryptosystem. Finally, the performance of the algorithm and the characteristics of vehicle information are comprehensively analyzed. The security of Internet of Vehicle's data transmission is efficiently improved by the proposed method.

Key words: Encryption algorithm; hazardous chemicals; Internet of Vehicles; vehicle logistics; security research; Autonomous Vehicles; Roadside units

1. Introduction. Internet of Vehicles at home and abroad With the continuous improvement of sensing technology, network communication technology, and data analysis and computing technology, in-vehicle system networks, cyber-physical systems and automotive Internet of Things are also gradually developing. But for the security risks brought by intelligent connected cars, it is still a major problem that scholars need to break through. The safety of vehicle network transportation of hazardous chemicals logistics vehicles is important and the current situation of hazardous chemicals transport vehicles in the Internet of Vehicles is analyzed, encrypt the information of hazardous chemicals transport vehicles, the concealment of vehicle information is guaranteed, and traffic accidents caused by the theft of hazardous chemicals can be prevented to a certain extent. The AES and RSA hybrid encryption algorithm is used to encrypt vehicle information, realize the concealment of vehicle information. On the basis of the original remote monitoring system for hazardous chemicals transport vehicles (Hazardous chemicals vehicle networking), according to the project's requirements for vehicle information encryption, using information encryption technology, the key information of the Internet of Vehicles is encrypted and transmitted, in order to achieve the purpose of information security of the Internet of Vehicles for hazardous chemicals [1].

To this end, the author aims at the network security problems existing in the existing vehicle networking of hazardous chemicals logistics vehicles, combined with the specific needs of the project, an encryption algorithm suitable for the information security of the Internet of Vehicles in the logistics of hazardous chemicals is designed,

[*] Department of information technology, GuangXi Police College, Nanning, GuangXi, 530028, China (yujiantang00234@yahoo.com).

[†] Department of information technology, GuangXi Police College, Nanning, GuangXi, 530028, China (YaChen39@163.com).

[‡] Department of Computer Engineering, PaiChai University, Daejeon, 35345 China (HoekyungJung@163.com).

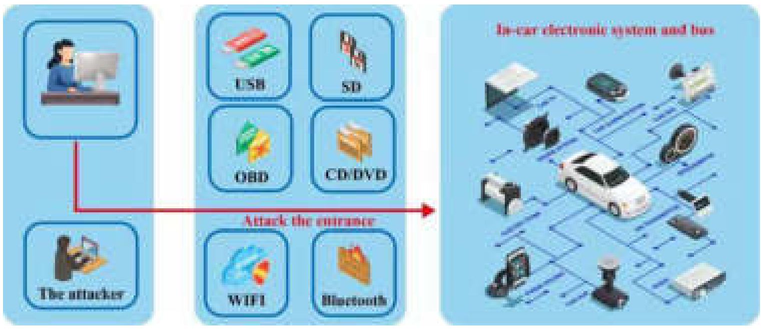


Fig. 1.1: The Internet of Vehicles for Logistics Vehicles

and the encryption algorithm is applied to the Internet of Vehicles in the logistics of hazardous chemicals to realize the confidential communication of the Internet of Vehicles(See Figure 1.1).

The National Highway Traffic Safety Administration (NHTSA) published a proposition for all new manufactured vehicles. Aside from vehicles, it is also essential for 5G communications. The vehicles achievement is the functionality aim by sharing vehicle driving-related information. The security is the design and privacy foundation and it is a priority. Unfortunately, relevant research is scarce concerning in the vehicular communication field. The quantum computers emergence will disrupt traditional cryptographic communications, which will increasingly require physical layer security needs in communications. The possibility of V2V communications suggested by the study and utilizing a secrecy capacity defined as the confidential data. They investigated secrecy capacity limited to vehicle communication and confirm that security parameters that can be controlled. Existing studies have attempted for secrecy capacity calculation by the system model but these efforts have failed for providing meaningful information. The vehicle secrecy capacity is defined with only Signal-to-Noise Ratio values provided in existing wireless communications to perform vehicle communication using the vehicle's defined secrecy capacity. The secure vehicle communication defined by the vehicle secrecy capacity within a security cluster.

In dangerous chemical logistics inside Internet of Vehicles (IoVs), encryption algorithms play a basic part in tending to security concerns. These concerns stem from the sensitivity of information related to the transportation and dealing with of dangerous materials. Encryption algorithms guarantee that delicate data, such as vehicle routes and communication between vehicles and central frameworks, is ensured from unauthorized access. Encryption algorithms defend information judgment by encoding it in a way that can as it were be unscrambled by authorized parties with the comparing keys. This anticipates unauthorized access or altering of basic data, decreasing the chance of robbery, attack, or unauthorized get to to dangerous materials. By scrambling information transmitted between vehicles and central frameworks, encryption algorithms ensure the security of people included in unsafe chemical logistics. This guarantees that individual data, such as driver personalities or cargo substance, remains secret and blocked off to unauthorized parties. Encryption algorithms improve the reliability of communication inside IoVs frameworks by avoiding listening in or information control. This guarantees that communications between vehicles, control centers, and other partners stay secure and solid, encouraging secure and productive transportation of perilous chemicals. By and large, encryption algorithms serve as a essential component of security measures in IoVs for perilous chemical logistics, defending delicate information, ensuring security, and improving the reliability of communication channels.

Contribution. The research aims to upgrade the data security of the Web of Vehicles (IoVs), especially concerning dangerous chemical logistics.

1. Firstly, it analyzes the special data characteristics of vehicles transporting unsafe chemicals, taken after by an examination of the Information Encryption Standard (DES) and Advanced Encryption Standard

(AES) algorithms inside the symmetric cryptosystem.

2. Furthermore, it digs into the RSA calculation inside the open key cryptosystem.
3. At last, the consider comprehensively assesses the execution of these calculations and the particular qualities of vehicle data.
4. Moreover, a crossover encryption calculation plot is proposed, wherein AES scrambles vehicle data and RSA is utilized to scramble AES keys.

Experimental outcomes illustrate the adequacy of this strategy in essentially moving forward the security of information transmission inside the Web of Vehicles.

The rest of the paper is organized as follows. Section II provides an overview of the exiting techniques. The research methodology is discussed in section III. Results and discussions are provided section IV and Finally, concluding remarks are provided in Section V.

2. Literature Review. The Internet of Vehicles technology is under the background of the increasingly perfect transportation infrastructure and the increasing difficulty of vehicle management, the concept established on the basis of the Internet of Things is the integration of traditional automobile industry and mobile Internet technology, it can be regarded as a specially optimized mobile ad hoc network. The Internet of Vehicles utilizes the integration of technologies such as GPS, RFID, wireless communication, and sensor networks, through information processing, wireless communication and information sharing between vehicles and X (people, roads, environment, etc.) In this way, it can realize the regulation of vehicles, improve traffic conditions, and at the same time provide users with comprehensive services such as safety and entertainment. The car networking system takes the car as the central node, and connects the car to the network through modern information means, realize the interconnected perception of vehicles and vehicles, vehicles and roads, vehicles and the environment. In today's Internet of Vehicles, the openness of wireless channels and the high-speed mobility of vehicles make vehicle information more vulnerable to attacks, for example, injecting erroneous beacon information into the communication link, tampering with and retransmitting previously broadcast beacon information, etc., can cause harm to the vehicle or driver. For example, Sun, Y, through security and privacy protection protocols and technologies, solve the problems of information security and privacy protection encountered by the Internet of Vehicles in the process of intelligent transportation [2]. Long, N. T. et al. proposed a security protection system for the Internet of Vehicles based on the case of security incidents of the Internet of Vehicles [3]. After analyzing the characteristics of the transmission channel of the Internet of Vehicles, Yu-Mei analyzed vehicle privacy and location privacy, and proposed schemes based on roadside nodes and mutual cooperation between vehicles [4]. Xin-Gang et al. designed a set of intelligent management security positioning system that can perceive in real time, realize precise positioning and security monitoring of data, and ensure the safe transmission of information [5]. Xu, B et al. explored the application of encryption technology in e-commerce in 2009 [6]. Liu, J. et al. studied encryption in aviation in 2009 [7]. In 2002, Wen-Yan conducted energy attack research on the AES algorithm, and the attack complexity ranged from 267 to 2131, which further reduced the scale of the attack [8]. In the same year, Wu, X improved the AES algorithm to make it have Square attack capability and good performance [9]. Based on the current research, the author proposes a research on the information security of the Internet of Vehicles. With the rapid development of intelligent transportation, the Internet of Vehicles as the core has huge development prospects. As a national strategic emerging industry, the safety of the Internet of Vehicles has become the focus, which is the premise and foundation of the development of the Internet of Vehicles. Aiming at the security threat of Internet of Vehicles information transmission, the author proposes a data transmission protection method based on encryption algorithm. Authors discussed the Vehicle-to-Vehicle communications that is suggested to secure in a secure cluster that refers to a vehicles group having a certain level of secrecy capacity. There are problems in secrecy capacity, but vehicular secrecy capacity is defined for the vehicle by SNR. The vehicular secrecy capacity is effective in achieving physical layer and may be changed by antenna related parameters. The vehicle-related parameters are addressed such as vehicle speed, safety distance etc. The vehicle-related secrecy parameters and secrecy capacity through modeling relationship is confirmed in traffic situations. The vehicular secrecy capacity is utilized to achieve secure vehicle communications that enables economic, and effective physical layer security [10]. Author in this paper discussed the Block-chain technology that has been emerged as a decentralized approach. The benefits of trustworthiness and mitigates the problem of single point of failure are offered by Blockchain. Author gives

Blockchain-enabled IoVs (BIoV) on their applications such as crowdsourcing-based applications, and accident avoidance and infotainment and content cashing. In-depth applications federated learning (FL) applications for BIoVs are also presented by the author [11]. Author in this paper presented a three-layer framework through which automotive security can be understood better. The vehicle dynamics and environmental sensors made up the sensing layer for jamming, and spoofing attacks. The communication layer is contained of both in-vehicle and V2X communications and sybil attacks. The sensing and communication layers are targeted by the attacks and affect the functionality and can further compromise the control layer security [12]. Author presented a Cyber Security Evaluation Framework (CSEF) for the evaluation of the security in-vehicle ECUs evaluation. The proposed technique is applied to On-Bord Unit (OBU) for providing a use case. The proposed CSEFis shown is to figure out assets, threats, and vulnerabilities of OBU, playing to conduct security evaluation. Moreover, for the cyber security evaluation, the CSEF can be extended such as the Telematic Box and the gateway [13].

Within the domain of high-speed mobility of vehicles, a noteworthy progressions have been made, with later considering the application of encryption algorithms in Internet of Vehicles (IoVs) to address security challenges. For instance, Wang et al. [3,4] proposed an enhancement strategy based on large traveler vulnerability to moderate the impacts of expansive traveler stream on high-speed mobility of vehicle frameworks. Xu, B et al. [6] further explored the application of encryption technology in e-commerce which is further combined with the study of encryption in aviation by Liu, J. et al. [7] Whereas these studies the effectiveness of improving safety and effectiveness, they don't broadly address the security concerns for IoVs. Later progressions in vehicular communication security have presented modern viewpoints on encryption algorithms. Quantum computing and physical layer security have developed as promising domain for supporting the security of vehicular communication systems. Quantum-resistant encryption algorithms offer security against potential dangers postured by quantum computing, ensuring the secrecy and judgment of transmitted information. Also, physical layer security strategies leverage the characteristics of remote communication channels to set up secure communication links, relieving and capturing attempted dangers. In light of these improvements, the proposed crossover encryption algorithm contributes essentially to upgrading the security of information transmission in perilous chemical logistics inside IoVs. By combining the qualities of customary encryption strategies with quantum-resistant procedures and physical layer security components, the crossover encryption algorithm offers vigorous security against different cyber dangers and guarantees the secrecy, integrity, and genuineness of delicate information transmitted over IoVs systems. Moreover, the integration of progressed encryption algorithms in IoVs frameworks upgrades the general strength and dependability of urban rail travel foundation, defending against potential security breaches and guaranteeing the secure and secure transportation of dangerous chemicals. In general, the proposed hybrid encryption algorithm provides a significant advancement in tending to security challenges in IoVs and contributes to the consistent advancement of security and security measures in urban rail travel frameworks.

The author in this paper details the emergence of the Internet that has provided a new direction for the development and upgrading of automobiles. The Internet technology and information technology are combined with existing vehicles for realizing the automobiles intelligent advancement. The developing smart cars goal is to achieve driverless driving and the problems in vehicle information security are increasing gradually. This paper studies the intelligent networking automotive technique and vehicle information security issues based on CAN bus that contributes to the intelligent networked vehicles [14].

Research Gap. Internet of Vehicles at home and abroad With the continuous improvement of sensing technology, network communication technology, and data analysis and computing technology, in-vehicle system networks, cyber-physical systems and automotive Internet of Things are also gradually developing. But for the security risks brought by intelligent connected cars, it is still a major problem that scholars need to break through.

3. Research Methodology.

Research status of the Internet of Vehicles at home and abroad. With the continuous improvement of sensing technology, network communication technology, and data analysis and computing technology, in-vehicle system networks, cyber-physical systems and automotive Internet of Things are also gradually developing. But for the security risks brought by intelligent connected cars, it is still a major problem that scholars need to break through.

Current status of foreign IoV research. Behind the convenience that cars provide people, they also threaten people's safety all the time, the convenience and safety it brings are contradictory [15]. With the increasing number of Internet of Vehicles applications, assisted driving technology, and the birth of the corresponding self-driving cars, have made the existing contradictions even more serious. In 2015, talented American hackers Charlie Miller and Chis hllmlk tested a car for a cyber attack, during the test, the Ucomect in-vehicle system was installed on the vehicle and the car was driven normally, they use remote commands to hack into the vehicle system and try to control the car, after intrusion, they use remote commands to overturn the car, such security threats will seriously affect the safety of the occupants of the vehicle. In October 2016, NHTSA (U.S. Department of Transportation Road Traffic Safety Administration) released the "Best Practices for Cybersecurity in Hyundai Vehicles", this best practice shows that the development of the Internet of Vehicles requires network security, and has good guidance for the development of the Internet of Vehicles enterprises, it clearly proposes to formulate automotive safety standards including cybersecurity, it is used to regulate individuals and organizations such as automobile manufacturing, automobile system or software design, and suppliers, aiming to improve the security of modern vehicle networks, and to guide on-board systems how to prevent and defend against cyber attacks [16]. At the same time, NHTSA also requires Internet of Vehicles companies to conduct network security assessments on in-vehicle assistance systems. Equipment that does not meet security standards cannot be installed on vehicles, and measures should be taken to deal with network threats and network vulnerabilities.

Research status of domestic Internet of Vehicles. My country introduced the concept of Internet of Vehicles in 2010, my country's automobile manufacturers, as well as various automobile technology research institutions, have established the In-Vehicle Information Service Industry Application Alliance (TIAA), committed to the research and application of Internet of Vehicles technology, and promote the development of Internet of Vehicles in my country. The research content of the Internet of Vehicles in the field of transportation is mainly divided into two aspects: highway and urban road traffic safety. Among them, the construction projects of expressways include the national network of expressway ETC, which solves the congestion problem of expressway toll stations in my country and is a major engineering project in China. Implementation of the project, there is no need to manually manage expressway tolls, which improves the operating efficiency of vehicles on expressways, the management department can also check the safety status of the vehicle operation on the highway in real time. In the construction of urban road safety, in order to alleviate traffic congestion and improve road traffic conditions, each city has established a vehicle management system to centrally control and dispatch all vehicles [17]. My country's smart car industry is in a stage of rapid development. In order to promote the development of its smart cars, the goal of the smart vehicle innovation development strategy proposed by the state, it is the six major systems that will eventually establish the Internet of Vehicles industry, and it also attaches great importance to the core technology of smart cars, my country is still in the follow-up stage and needs to establish its own car networking system, build five basic platforms for smart cars, and promote the construction of innovative capacity for smart connected cars [18]. In July 2017, in order to integrate superior resources, promote the development of the standard "Information Security Technology Automotive Electronic System Network Security Guidelines", network security research institutions represented by China Electronics Standardization Institute and University of Electronic Science and Technology of China, combined with China FAW, Shanghai Automobile, and other automobile manufacturers to form a standard preparation working group, jointly promote the first national standard in the field of automotive electronic network security in my country.

The methodology for this study involves a systematic approach to analyzing the characteristics of hazardous chemical information and conducting subsequent encryption algorithm analysis to enhance methodological transparency. Initially, relevant information regarding hazardous chemicals, including their properties, handling requirements, and transportation regulations, will be collected through literature review and consultation with industry experts. This data will then be categorized and classified based on characteristics such as toxicity level, flammability, and health hazards, followed by a comprehensive risk assessment to identify potential vulnerabilities in transportation and handling processes. Subsequently, specific information security requirements related to hazardous chemical logistics will be analyzed, focusing on aspects such as data confidentiality, integrity, availability, and authentication. Based on these requirements, criteria for selecting encryption algorithms will

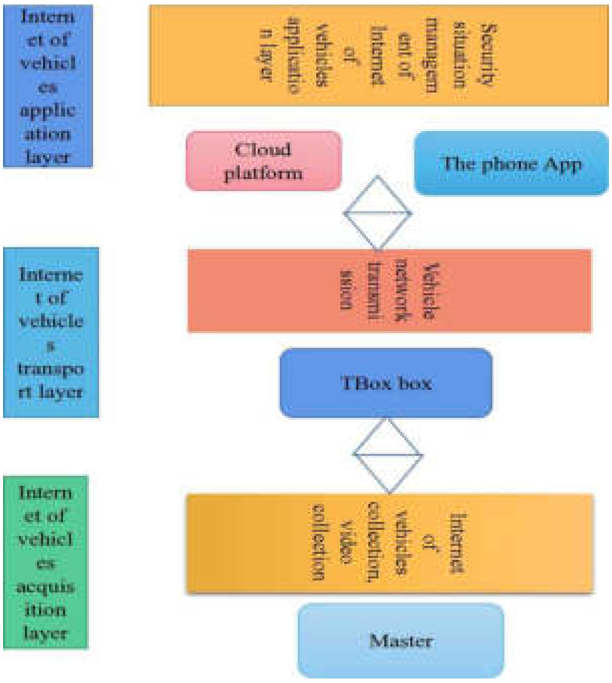

Internet of vehicles application layer
Security situation management of Internet of vehicles application layer
Cloud platform
The phone App
Internet of vehicles transport layer
Vehicle network transmission
TBox box
Internet of vehicles acquisition layer
Internet of vehicles collection, video collection
Master

Fig. 3.1: The structure of the vehicle network system

be defined, considering factors such as encryption strength, computational efficiency, and compatibility with existing systems. Different encryption algorithms will then be evaluated based on their performance against the selection criteria, including practical implementation considerations and security analysis to identify potential vulnerabilities and attack vectors. Integration of findings will involve synthesizing results from the hazardous chemical information analysis and encryption algorithm evaluation to provide recommendations for enhancing information security in hazardous chemical logistics within Internet of Vehicles (IoVs). Through these specific procedures, this methodology aims to ensure methodological transparency and rigor in addressing the research objectives.

3.1. Internet of Vehicles System. The architecture of the Internet of Vehicles is divided into application layer, transmission layer and acquisition layer, as shown in Figure 3.1.

(1) Application layer: Mainly aimed at car users, it provides users with intelligent services through mobile terminals such as mobile phones. It mainly realizes the functions of road condition analysis, vehicle status analysis and vehicle failure analysis, in this process, it achieves information sharing with the urban transportation center, provides convenient services to users, and also provides assistance to urban transportation.

(2) Transport layer: The collected data is transmitted to the application layer by means of 4G, Bluetooth and RFID.

(3) Acquisition layer: Using car sensors, video collectors and audio collectors, etc., to obtain information such as the car's own state and road conditions, and then to the application layer through the transport layer. At this stage, the function of the Internet of Vehicles is still in the monitoring stage, and the unification of people, vehicles and roads has not been effectively realized.

1. Cyber Security Assessment System. Usually like a checklist or a set of rules that specialists utilize to check how secure and secure a computer framework or a organize is. It makes a difference them get it in the event that there are any shortcomings or vulnerabilities that programmers seem abuse to take data or cause issues. By taking after this system, they can make beyond any doubt that the framework is as ensured as conceivable against cyber dangers.

2. Vehicular Communication Vulnerabilities. Envision in the event that vehicles may deliver to each other and share data like where they're going or in case there's something within the road ahead. It might to offer assistance prevent accidents and make driving more secure. But in the event that programmers were able to urge into this communication framework, they seem cause chaos. They could create the things they're not assumed to, like all of a sudden halt or swerve off the road. This might lead to mischances and put people's lives in threat. So, it's truly imperative to form any doubt that the communication between vehicles is secure and ensured from programmers.

Practical Illustrations of Encryption Algorithms in Vehicular Communication and Suggestions of NHTSA's Recommendation:

1. Illustrations of Encryption Algorithms. Encryption algorithms are like secret codes that are utilized to keep data secure when it's being sent from one put to another. For illustration, envision you're sending a message to a companion, but you do not need anybody else to be able to examine it. You may utilize an encryption algorithm to turn your message into a secret code some time recently you send it. At that point, your companion can utilize the same algorithm to turn the code back into the initial message when they get it. This way, indeed in the event that somebody tries to captured the message, they won't be able to get it it since it's all cluttered up.

2. Suggestions of NHTSA's Recommendation for V2V Communication. The National Highway Activity Security Organization (NHTSA) has proposed that modern vehicles should be prepared with innovation called Vehicle-to-Vehicle (V2V) communication. This implies that cars would be able to link each other wirelessly, sharing data about their speed, course, and area. The idea is to assist anticipate mischances by giving drivers notices in the event that there's a hazard of a collision. For illustration, in case one vehicle abruptly brakes, it seem send a flag to adjacent cars to caution them to moderate down. By utilizing encryption algorithms to secure this communication, the NHTSA points to guarantee that the data exchanged between vehicles is secured from programmers who might attempt to alter with it or cause mishaps by sending untrue signals. This would make our roads more secure for everyone.

Within the setting of unsafe chemical logistics, both AES (Advanced Encryption Standard) and RSA (Rivest-Shamir-Adleman) encryption algorithms play a critical parts.

AES Encryption Algorithm. AES may be a symmetric encryption calculation known for its productivity and security. In dangerous chemical logistics inside IoVs, AES can be utilized to encrypt delicate information transmitted between vehicles, control centers, and other partners. AES guarantees information privacy and integrity, ensuring against unauthorized access and altering of basic data such as cargo subtle elements, course plans, and crisis conventions. The strong security and effectiveness of AES make it well-suited for securing communication channels and information exchanges in dangerous chemical logistics scenarios.

RSA Encryption Algorithm. RSA is an asymmetric encryption algorithm commonly utilized for key exchange and computerized marks. In unsafe chemical logistics inside IoVs, RSA can be utilized for secure key trade instruments, empowering encrypted communication channels between vehicles and control centers. RSA gives a secure establishment for establishing trust and realness in IoVs systems, confirming the personality of members and guaranteeing the judgment of transmitted information. Whereas RSA offers solid security ensures, its computational complexity may present proficiency challenges in resource-constrained IoVs situations, requiring cautious consideration of execution trade-offs.

In this way, both AES and RSA encryption algorithms play basic parts in guaranteeing the security and security of information transmission in dangerous chemical logistics inside IoVs, defending against cyber dangers and unauthorized get to to sensitive data.

3.2. Protection method of data transmission based on encryption algorithm. For the communication transmission link of the Internet of Vehicles, the author designs a transmission link with an encryption algorithm, as shown in Figure 3.2, the T-BOX box in the Internet of Vehicles uses an encryption algorithm to encrypt the original data packets, send encrypted data to the transport channel [19]. After receiving the encrypted data packet in the backend of the Internet of Vehicles, it uses the decryption algorithm to decrypt the encrypted data packet, and performs verification processing on the data, if the data passes the verification, the corresponding instruction is executed; If the data does not pass the verification, the data is discarded.

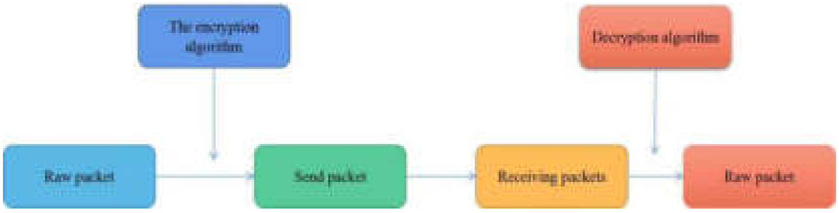


Fig. 3.2: Flowchart of adding encryption algorithm information

(1) Principle of Hybrid Encryption Algorithm. The hybrid encryption algorithm needs to consider the encryption mode, whether to use the AES algorithm or the RSA algorithm to encrypt the vehicle information, not only the characteristics of the algorithm itself, but also the characteristics of the vehicle information should be considered. Combining the previous analysis and the description of the vehicle information, the encrypted data is a series of strings. We know that the RSA algorithm takes a long time to encrypt a small amount of data, so the AES algorithm is used to encrypt plaintext data, generate AES key and ciphertext, and then encrypt the AES key with RSA, which enhances the security of the encryption system; After receiving the data, the receiver first uses the RSA key to decrypt the AES key, uses the AES key to decrypt the ciphertext, and recovers the plaintext information 20.

(2) Development Trend of Hybrid Encryption Algorithms. With the continuous advancement of positioning technology and mobile network technology, promoted the development of location-based applications, in terms of vehicle location privacy protection, schemes based on encryption techniques are commonly used. In the information security of the Internet of Vehicles, the hybrid encryption algorithm is more and more widely used in the field of network information security, and the hybrid algorithm is usually realized by a combination of hardware and software. Due to the fast encryption speed of the symmetric encryption algorithm, the asymmetric encryption algorithm has high security, so the combination of the two is a popular way today, such as DES-RSA, IDEA-RSA, DES-ELGAMAL. In today's information network communication, the hybrid encryption algorithm because of its good encryption performance, in practice, the frequency of use is extremely high. Some of ZTE's high-end products, such as routers, feature hybrid encryption. Choose a combination from symmetric encryption algorithms such as DES, 3DES, IDEA, and asymmetric encryption algorithms such as RSA, ECC, etc, the hybrid algorithm combines the advantages of the two algorithms, will be more widely used in the future.

3.3. The legal development trend of encryption algorithm. With the continuous advancement of positioning technology and mobile network technology, promoted the development of location-based applications, in terms of vehicle location privacy protection, schemes based on encryption techniques are commonly used. In the information security of the Internet of Vehicles, the hybrid encryption algorithm is more and more widely used in the field of network information security, the hybrid algorithm is usually implemented by a combination of hardware and software [21]. Due to the fast encryption speed of the symmetric encryption algorithm, asymmetric encryption algorithms have high security, so the combination of the two is a popular way today, such as DES-RSA, IDEA-RSA, DES-ELGAMAL. In today's information network communication, the hybrid encryption algorithm because of its good encryption performance, in practice, the frequency of use is extremely high. Some of ZTE's high-end products, such as routers, feature hybrid encryption. Choose a combination from symmetric encryption algorithms such as DES, 3DES, IDEA, and asymmetric encryption algorithms such as RSA, ECC, etc. the hybrid algorithm combines the advantages of the two algorithms and will be more widely used in the future [22].

4. Experiments and Research. To validate the adequacy of the proposed encryption strategy, a comprehensive exploratory setup and information examination strategies are vital. The test setup and information examination strategies, besides the insights into the suggestions of encryption algorithm determination on the in general security and effectiveness of Internet of Vehicles (IoVs) in unsafe chemical logistics.

1. Experimental Setup. Create a dataset representing different scenarios in perilous chemical logistics inside IoVs. This dataset ought to incorporate data such as vehicle courses, cargo points of interest, communication messages, and potential security dangers. Execute the proposed encryption strategy, consolidating chosen encryption algorithms (e.g., AES, RSA) into the IoVs communication system. Guarantee that encryption is connected to delicate information transmissions between vehicles, control centers, and other partners.

2. Simulation Environment. Create an environment imitating real-world IoVs scenarios, including vehicle development, communication conventions, and potential security vulnerabilities. Utilize simulation apparatuses such as ns-3 or OMNeT++ for reasonable experimentation. Then design different scenarios speaking to distinctive security dangers and attack scenarios, such as listening stealthily, altering, or information capture attempts. Control parameters to simulate diverse levels of risk escalated and encryption algorithm adequacy. For data analysis, the encryption algorithm, assess the performance of the encryption method by measuring the parameters like encryption/decryption speed, computational overhead, and resource utilization. Then comparing the performance of different encryption algorithms (e.g., AES, RSA) under various scenarios. For evaluating the security effectiveness of the encryption method by analyzing its resilience against common cyber threats and attack vectors. Generally, the determination of encryption algorithms such as AES or RSA in IoVs for dangerous chemical logistics includes trade-offs between security, productivity, and computational overhead. Cautious consideration of these components is basic to guarantee the viability and unwavering quality of encryption strategies in shielding the delicate information transmissions and ensuring against cyber dangers in IoVs situations.

4.1. Process Design of Hazardous Chemical Vehicle Encryption System. The overall process of the remote encryption monitoring system for hazardous chemicals vehicles. First initialize the system, the CPU enters the state of preparing to execute the task, load the AES initial key to the AES-EN port for key expansion, so that the encryption operation can be performed directly when there is vehicle information. Then the Beidou module transmits the obtained positioning data to the built-in CPU of the FPGA through the serial interface, and after analysis and processing, send the vehicle location information to the encryption unit AESand RSA-Core to encrypt the vehicle information to form ciphertext, and send the ciphertext to the communication module through the serial interface, the communication module sends the ciphertext to the measurement and control center through the wireless transmission network, and uses the software to decrypt the ciphertext, the vehicle position information is recovered and stored in the database, and the measurement and control center calls the vehicle information in the database to display on the terminal computer [23].

(1) Research on encryption algorithms. At present, with the expansion of the operation scope of the logistics industry, the scale of the network of vehicles for hazardous chemicals logistics is gradually expanding, and there is a lot of real-time status information of vehicles. Due to the transparency of the information transmission of the Internet of Vehicles, there is a problem with the safety of vehicle driving. Based on the actual situation and specific needs of the project, the author encrypts the information of the hazardous chemicals logistics vehicle to ensure the safe operation of the vehicle. The second chapter elaborates the encrypted monitoring system for hazardous chemicals vehicles [24]. The positioning data received from the Beidou satellite is analyzed by the CPU to extract the vehicle's driving date, longitude, latitude, altitude and other information data, these vehicle information formats are all ASCII character formats, and the corresponding message message format is designed in combination with the type and characteristics of the data.

(2) Encryption algorithm. After the "Prism" incident in the United States, my country actively promotes domestic encryption algorithms, and proposes to replace foreign encryption algorithms with domestic encryption algorithms to encrypt data, making it difficult for NSA to crack. The block cipher algorithm is a symmetric cipher algorithm, which is mainly used to realize the encryption and decryption of data information. The plaintext, key and ciphertext of the encryption algorithm are all 128 bits, and the same key is used for encryption and decryption [25]. Both the encryption algorithm and the key expansion algorithm are implemented by a

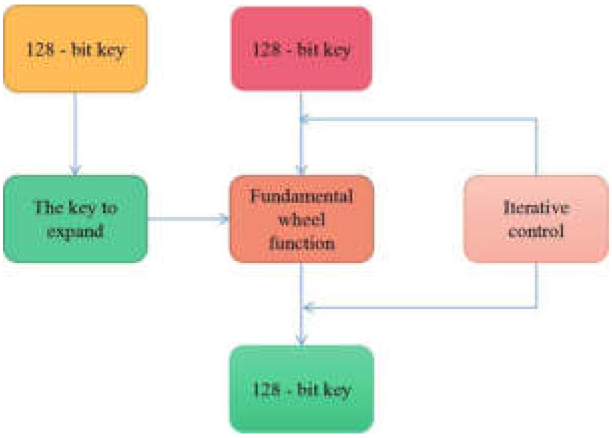


Fig. 4.1: The overall structure of the encryption algorithm

non-linear iterative round function with 32 cycles. The core of the data encryption part is the round function, which combines linear and nonlinear.The basic process is to first divide the 128-bit key into 4 groups according to a 32-bit group, and then generate 32 groups of 32-bit keys according to the key expansion algorithm; Then, the input 128-bit data is also divided into 4 groups according to 32-bit group for circular operation. The overall structure of the encryption algorithm is shown in Figure 4.1.

The plaintext X is 128 bits, rk0 rk32 are 32 sets of round keys, and the synthetic permutation T forms the round function F.

In the formula, L is a linear transformation; is a nonlinear transformation.The round key rk is generated by the key expansion algorithm. Known encryption key MK= (MK0,MK1,MK2,MK3), system parameter is FK= (FK0,FK1,FK2,FK3), fixed parameter CK= (CK0,...CK1,...CK31).

The output of the encryption transformation is: The decryption transformation of the encryption algorithm has the same structure as the encryption transformation, and the only difference is that the round key is used in the reverse order.

4.2. Performance Analysis of Encryption Algorithms. Through the theoretical analysis of the AES and RSA algorithms, the following is a comparative analysis of the performance of the hybrid encryption algorithm and the single algorithm from the aspects of encryption speed, security and key management [26].

(1) Encryption speed. For the RSA algorithm, to ensure security, its modulus n needs to be at least 1024 bits, then a large number of large integer multiplication and modulo operations are required, the time required for the RSA algorithm to encrypt and decrypt 1M (vehicle information in the Internet of Vehicles) files has been greater than 100s, the following mainly simulates the DES algorithm, the AES algorithm and the AES and RSA hybrid algorithm on MATLAB [27]. The results are shown in Figure 4.2.

As can be seen from Figure 4.2, the encryption speed of AES algorithm is obviously better than that of DES and hybrid encryption algorithm, and the performance of hybrid encryption algorithm and DES algorithm is close, however, in view of the security issues of DES and AES algorithms, hybrid algorithms are widely used in practice, therefore, comprehensive comparison and analysis, the performance of each algorithm, in this paper, AES and RSA hybrid encryption algorithm is selected to encrypt vehicle information [28].

(2) Security. This is the level of secrecy scientists have assigned to their data for a long time. From the previous analysis of DES algorithm, AES algorithm and RSA algorithm, it can be seen that a single encryption algorithm can no longer meet our security requirements, so a hybrid algorithm is proposed.

(3) Key management. From the theoretical analysis of the previous algorithm, it can be seen that RSA belongs to the public encryption system, and the encryption key is distributed in the public form, so the update of the encryption key is very easy, and for different communication objects, only need to keep their own decryption key secret. The AES algorithm belongs to the symmetric cryptosystem, for different communication

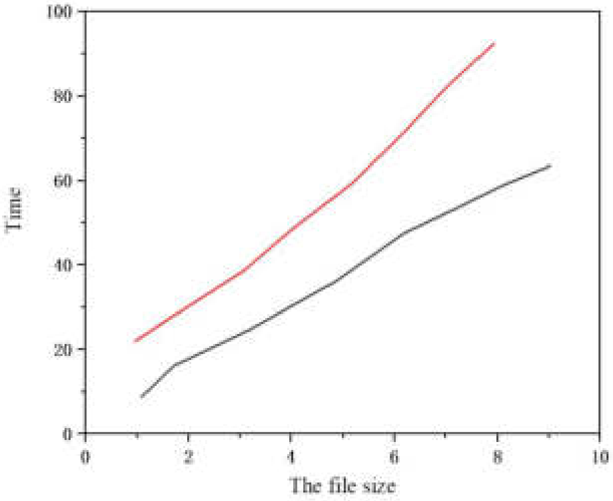


Fig. 4.2: Relationship between encryption and decryption time and file size

objects, AES needs to generate and store different passwords, the key management system has a large overhead, the key must be distributed before the communication, and the key needs to be replaced for different data, the replacement of the key is very difficult. In the hybrid encryption algorithm, the RSA algorithm only needs to encrypt the 128-bit key once, and does not need to generate a key pair, then there is no key management problem, it can be considered that the hybrid encryption algorithm solves the problem of AES and RSA key management. From this, it can be seen that, the AES encryption speed is faster than the RSA algorithm, especially for a large amount of information encryption, it is more advantageous to use the AES algorithm to encrypt. However, the AES algorithm has serious security problems in key management. There is no key transmission problem in the RSA algorithm, because the public key itself is public, and according to the RSA algorithm, it is difficult for a third party to solve the private key. Therefore, the author combines the advantages of the two algorithms, and combines the characteristics of the hazardous chemicals logistics vehicle information, and adopts the scheme of AES encryption of vehicle information and RSA encryption of AES keys.

4.3. Design of Encryption Algorithm Level Table. According to the mixed criticality of automotive electronic systems, information security is related to functional safety, different functions require different information security levels, the encryption algorithm is related to the information security level, different information security levels require different security encryption algorithms. The encryption algorithm level table reflects the corresponding relationship between key functions, information security levels and encryption algorithms. The security of encryption algorithms is quite different. High-level algorithms should ensure high encryption strength, at the same time, it meets the real-time requirements of the system, for example, once the car brake control function is cracked and exploited by an attacker, it is very likely to cause car crashes, so we must focus on protection, you can choose high-level encryption algorithms such as AES algorithm, ensure message security; While satisfying a certain encryption strength, low-level algorithms should occupy as little system resources as possible, save encryption and decryption time, and improve system efficiency, for example, if the window control function is cracked by an attacker, it will not pose a fatal risk to personnel. Lightweight encryption algorithms such as TEA are suitable choices. According to the ASCII level and the risk classification of automobile information security, the encryption algorithm level table is designed in combination with the security of the encryption algorithm, as shown in Table 4.1.

Table 4.2, from top to bottom, corresponds to the information security level, encryption algorithm, and

Table 4.1: The relationship between the security level and each algorithm

Security level	Symmetric dense length	key length	Confidentiality period
90	9	1036	2011
100	113	2038	2020
170	189	3312	2030

Table 4.2: Parameter settings

enter	din[127:0]=1 8720563580201 60101 1023456276731 key[127:0]=000102030405060708090a0b0c0d0e0f
output	dout[I 97-01-3 38hac077 fa 2 hfdla u489.70h0adh

the corresponding algorithm identification and ID range from low to high according to functional security requirements. In the CAN protocol, the frame ID is used to identify the priority of the CAN data frame. The smaller the frame ID value, the higher the frame priority. The higher the criticality of the vehicle function, the more timely the message needs to be transmitted. In order to transmit the message of the high critical function in time, the priority of the message ID should be assigned higher. According to the standard frame ID value range 0x000-0x7FF, the author assumes the ID range in Table 4.1 according to the security level. The design also combines encryption algorithms of different security levels to form an algorithm library, and assigns an algorithm identifier to each encryption algorithm, which is used as a unique characteristic value to realize the dynamic scheduling of encryption algorithms.

4.4. Analysis of AES Simulation Results. First, initialize the parameters of the designed BD2 encryption system. Combined with the project requirements, the input data is part of the vehicle information. The PLL generates the clock signal for each module in the system to work normally, the input clock signal is 50MHz, and the output clock signal c. For the frequency of 75MHz, the phase of 600 drives the SDRAM chip to work; q is a 75MHz drive encryption module; c2 is a 75MHz drive control module, such as CPU and peripheral modules. The pin UART-rxd is connected to the BD2 receiving module, and the UART-txd sends the cipher text to GPRS, and sends the cipher text to the monitoring center through the wireless transmission network, the monitoring terminal uses software to decrypt the ciphertext to recover the vehicle location information, for the relevant personnel to dispatch and manage vehicles. Table 5.1 is the setting of related parameters. This system uses Modelsim to simulate, the pins dout-a and dout-r[of the AES and RSA encryption modules have outputs, and the CPU is connected to the pins such as enc and clk, controls when encryption and key entry start. The positioning data of BD2 is input from UART-rxd, through the encryption module to the dout pin to the control module, here, the ciphertext is sent to the GPRS module via the UART-txd pin. According to the requirements of the project team, we only encrypt and decrypt the vehicle's terminal ID, positioning time, longitude and latitude. The file names are all named AES. as shown in Table 4.2.

First, use the AES encryption module alone, input din and key, and output the data doout through the AES encryption module, it is sent to the monitoring center through the wireless transmission network and decrypted by software.

5. Conclusion. To advance encryption algorithms for Internet of Vehicles (IoVs) in hazardous chemical logistics, several promising avenues warrant exploration. Firstly, the development of quantum-resistant cryptography is essential to withstand potential future threats posed by quantum computers, ensuring the longevity of IoV security. Additionally, investigating the application of homomorphic encryption could enable secure computation on encrypted data, enhancing privacy without compromising data analysis capabilities. Dynamic key management techniques tailored to the dynamic nature of IoV environments could bolster security by facilitating real-time distribution and updating of encryption keys. Moreover, staying abreast of post-quantum cryptography standards and adopting emerging techniques resilient to quantum attacks is paramount for IoV security. Furthermore, research into secure multiparty computation methods and their integration with IoV systems could enable secure collaboration and data sharing among vehicles and infrastructure. Thus, exploring the integration of blockchain technology with encryption algorithms has the potential to enhance transparency, accountability, and data integrity in hazardous chemical logistics within IoVs. The author deeply analyzes the status quo of the vehicle networking of hazardous chemicals logistics vehicles, aiming at the privacy protection of the vehicle's location, this paper proposes and designs a hybrid encryption algorithm for the vehicle networking of hazardous chemicals logistics vehicles, and completes the design and implementation of each module in the hybrid encryption system. Based on the system architecture of the Internet of Vehicles, the author proposes a protection method based on an encryption algorithm for the low security of the data of the Internet of Vehicles. Finally, the performance of the algorithm and the characteristics of vehicle information are comprehensively analyzed, and a hybrid encryption algorithm scheme is proposed, that is, AES encrypts vehicle information, technical scheme for encrypting AES keys with Rivest-Shamir-Adleman encryption (RSA). The benefit of encryption algorithm is analyzed from the security and realization of encryption algorithm. Finally, the experimental results show that the use of encryption algorithms can effectively protect the transmission data and increase the protection capability of the Internet of Vehicles information transmission. By advancing encryption algorithms in these directions, IoV systems can be fortified against evolving security threats, ensuring the robust protection of sensitive data and critical operations.

REFERENCES

[1] Yan, R., Lin, C., Zhang, W. F., Chen, L. W., Peng, K. N. , *Research on information security of users' electricity data including electric vehicle based on elliptic curve encryption*, International Journal of Distributed Sensor Networks, 16(11), 155014772096845, 2020.

[2] Sun, Y., Li, X., Lv, F., Hu, B. , *Research on logistics information blockchain data query algorithm based on searchable encryption*, IEEE Access, PP(99), 1-1, 2021.

[3] Long, N. T. , *Research on innovating and applying cryptography algorithms for security routing in service based routing*, Internet of Things and Cloud Computing, 3(3), 33-41, 2015.

[4] Yu-Mei, Y. I. , *Study on location-transportation optimization for hazardous material logistics network*, China Safety Science Journal, 21(6), 135-140, 2011.

[5] Xin-Gang, J. U., Guo, H. O., Liu, Y. , *Research of the security of iris recognition based on composite chaos encryption*, Journal of Henan Normal University(Natural Science), 37(3), 68-70, 2009.

[6] Xu, B., Zhang, L. H., Tan, X. P. , *Two-level emergency centers location model based on the hazardous chemicals' accidents*, Systems Engineering-Theory Practice, 35(3), 728-735, 2015.

[7] Wu, X., Oh, H. C., Akarimi, I., Goh, M., Souza, R. D. , *Tops: advanced decision support system for port and maritime chemical logistics*, The Asian Journal of Shipping and Logistics, 27(1), 143-156, 2011.

[8] Sang-Il, Jo, Jaesung, Lee , *Vehicle detection algorithm for vds by using décalcomanie matching based on histogram*, The Journal of Korean Institute of Communications and Information Sciences, 42(6), 1225-1232, 2017.

[9] Liu, C., Shao, Y., Cai, Z., Li, Y. , *Unmanned aerial vehicle positioning algorithm based on the secant slope characteristics of transmission lines*,IEEE Access, PP(99), 1-1.

[10] Ahn, N. Y., Lee, D. H. *Physical Layer Security of Autonomous Driving: Secure Vehicle-to-Vehicle Communication in A Security Cluster.* arXiv preprint arXiv:1912.06527, 2019.

[11] Hildebrand, B., Baza, M., Salman, T., Amsaad, F., Razaqu, A., Alourani, A. *A Comprehensive Review on Blockchains for Internet of Vehicles: Challenges and Directions.* arXiv preprint arXiv:2203.10708, 2022.

[12] El-Rewini, Z., Sadatsharan, K., Selvaraj, D. F., Plathottam, S. J., Ranganathan, P. *Cybersecurity challenges in vehicular communications.* Vehicular Communications, 23, 100214, 2020.

[13] Zhang, H., Pan, Y., Lu, Z., Wang, J., Liu, Z. *A Cyber Security Evaluation Framework for In-Vehicle Electrical Control Units.* IEEE Access, 9, 149690-149706, 2021.

[14] Chen, C., Zhang, B., Liu, M., Wei, S., Zhang, J., Shen, L. *Research on Intelligent Networking Automotive Technology and Information Security Based on CAN Bus.* In IOP Conference Series: Materials Science and Engineering (Vol. 688, No. 4, p. 044058). IOP Publishing, 2019.

[15] Zhang, H. , *An analysis on vehicular ad-hoc networks: research issues, challenges and applications*, International journal of computational intelligence research, 14(8), 641-655, 2018.

[16] Sari, A., Onursal, O., Akkaya, M. , *AReview of the security issues in vehicular ad hoc networks (vanet)*, International Journal of Communications, Network and System Sciences, 8(13), 552-566, 2015.

[17] Wan, J., Yan, H., Hui, S., Fang, L. , *Advances in cyber-physical systems research*, Ksii Transactions on Internet Information Systems, 5(11), 1891-1908, 2011.

[18] Ae Ger, A., Bimeyer, N., H Stübing, Huss, S. A. , *A novel framework for efficient mobility data verification in vehicular ad-hoc networks*, International Journal of Intelligent Transportation Systems Research, 10(1), 11-21, 2012.

[19] Fishman, S., Ph., D. , *Studies of the upper-extremity amputee* , Artificial Limbs, 5(1), 88, 1958.

[20] Kumar, N. S., Rajakumar, K. , *A study on security for adaptive periodic threshold sensitive energy efficient protocol based on elliptic curve cryptology in wireless sensor network*, International journal of computing information technology, 11(2), 137-147, 2019.

[21] Kumar, N. S., Rajakumar, K. , *A study on security for adaptive periodic threshold sensitive energy efficient protocol based on elliptic curve cryptology in wireless sensor network*, International journal of computing information technology, 11(2), 137-147, 2019.

[22] Council, B. N. , *SThird international symposium on intelligent information technology and security informatics*, Mis Quarterly, 33(4), 1, 2010.

[23] Kai, L. , *Research on adaptive target tracking in vehicle sensor networks*,Journal of Network Computer Applications, 36(5), 1316-1323, 2013.

[24] Qureshi, K. N., Alhudhaif, A., Shah, A. A., Majeed, S., Jeon, G. , *Trust and priority-based drone assisted routing and mobility and service-oriented solution for the internet of vehicles networks*, Journal of Information Security and Applications, 59(5), 102864.

[25] Goel, S., Yuan, Y. , *Emerging research in connected vehicles [guest editorial*, IEEE Intelligent Transportation Systems Magazine, 7(2), 6-9, 2015.

[26] B Ji, X Zhang, S Mumtaz, C Han, C Li, H Wen, *Survey on the internet of vehicles: network architectures and applications*, IEEE Communications Standards Magazine, 4(1), 34-41, 2022.

[27] Schumacher, H. J., Ghosh, S. , *A fundamental framework for network security*, Journal of Network Computer Applications, 20(3), 305-322, 1997.

[28] Nopmongcol, U., Griffin, W. M., Yarwood, G., Dunker, A. M., Maclean, H. L., Mansell, G. , *Impact of dedicated e85 vehicle use on ozone and particulate matter in the us*, Atmospheric environment, 45(39), p.7330-7340, 2011.

Edited by: Pradeep Kumar Singh
Special issue on: Intelligent Cloud Technologies Enabled Solutions for Next Generation Smart Cities
Received: Nov 9, 2022
Accepted: Mar 1, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1301–1310, DOI 10.12694/scpe.v25i3.2065

SECURITY SITUATION AWARENESS SYSTEM BASED ON ARTIFICIAL INTELLIGENCE

HAO WU*

Abstract. There is rapid growth of the security threats faced by enterprises and the security attacks technology is also established at a higher level.Enterprises are facing an escalating threat landscape marked by sophisticated security attacks. To address the structural and technical challenges of information security situational awareness, a method for designing an artificial intelligence-driven system is proposed. In order to solve the system structure and key technical problems of information security situational awareness technology of artificial intelligence, a method of designing information security situational awareness system is proposed, and experiments are carried out through the method. By analyzing the data sources that the system needs to collect, including network traffic mirror data, log data, security intelligence and support data, this paper verifies the feasibility of the system method of information security situational awareness technology of artificial intelligence technology. The proposed core competence platform has the characteristics of low delay and robust real-time capabilities. By presetting the event processing topology, we can quickly build the event processing process, and build the corresponding event processing topology model according to different processing requirements to meet the business requirements. The AI-powered information security situational awareness system substantially enhances security awareness and prediction accuracy.

Key words: Network security; Situational awareness; Network defense; Network defense; Real time; Event processing topology; Artificial intelligence

1. Introduction. With the continuous evolution of Internet hacker technology, the security of information network is constantly challenged, and the potential threats are becoming greater and greater. The traditional passive defense system can not meet people's needs for network security [1]. The emergence of network security situational awareness technology opens up a new way to ensure the security of information network. With the advent of the Internet era, people's lifestyles have undergone earth shaking changes. All walks of life are constantly upgrading and transforming under the impact of the Internet. At the same time, all kinds of network security threats in the Internet era are also increasing [2]. In recent years, cyber attacks with national and organizational backgrounds are increasing. The special roles of the government, military, finance and large enterprises often face more external attack threats. Therefore, the research on information security situational awareness system is necessary [3].

The rapid development of internet has caused a quick growth of network data. While it brings expediency to work and life of people, the large data also brings great risks of security to the network. Therefore, the security situation awareness method is developed, such as the Bayesian method based network security situation awareness model and the improved G-K algorithm based multi-node network security situation prediction awareness model. The security events in the network are detected by these two models but the security situation awareness model classification model is not good. New cascaded network security situational awareness model based on fusion decision tree algorithm is built. With rapid growth of energy internet and AI, big data are playing important roles in the management mode and value function construction. The large power data research is being paid more attention and data processing complexity is getting higher and higher, which brings challenge to the traditional security transmission management. The economical and highly scalable IT services are provided by the cloud computing for the remote computer users. The cloud-based data transmission technique is the transmission of big data from a cloud storage point to a destination storage point according to the cloud storage. This technique is advantageous as it can get rid of the hardware resource limitations. The data overflow will occur if there is limited cloud space capacity. The device itself, including trusted computing,

*Department of Information Engineering, ShiJia Zhuang University of Applied Technology, Shijiazhuang, Hebei 050081, China (haowu10088@outlook.com).

network equipment, security protection equipment, databases, etc. generates all the information that needs to be collected. After summary processing, it is submitted to the network security management platform on the main station and the plant side.

1.1. Contribution.

1. The article proposes a method for designing an information security situational awareness system to tackle the structural and technical hurdles associated with artificial intelligence technology.
2. Through experimentation, the method is tested, validating its efficacy in analyzing necessary data sources for the system.
3. The contribution underscores the feasibility of the proposed system in leveraging artificial intelligence to enhance information security situational awareness.

Rest of paper is organized as follows: Exhaustive literature survey is detailed in section 2 followed by the research methods in section 3. Results are discussions are presented in section 4 and the section 5 concludes the paper.

2. Literature Review. Xu, R. and others found that network security situational awareness comes from situational awareness [4]. Cohen, R. S. and others believes the recognizing and understanding environmental factors within a certain space-time range [5]. Chan, J. L. and others found that the bass functional model of network security situational awareness has been widely recognized. [6]. Starting from bass functional model, people's research mainly focuses on the following contents: First, the related technology of data collection. Korolyov, V. and others found that due to the diversity of network sensors and the complexity of network structure, how to effectively select sensors and fuse data plays a vital role in the subsequent situation analysis [7]. Khairy, D. and others pointed out that data collection is the most important part of the whole situation analysis cycle, and divided the data sources into complete content data, session data, statistical data, packet string data and alarm data. In addition, they also proposed an application collection framework (ACF) to reduce the complexity of data collection [8].

Second, the related technology of object extraction. Azar, R. and others believe that the original data has the characteristics of large amount of data and more redundant information. The object extraction process is the process of detecting the collected data in the network security situational awareness system. It takes the original data as the input and obtains high-level objects based on the original data, such as abnormal events and alarms [9]. Munir, A. and others proposed an alarm aggregation algorithm based on a commercial product. The algorithm aims to eliminate the interference of redundant alarms and obtain high-value aggregated alarms. The aggregated alarms here are the extracted objects [10]. Third, the related technology of situation extraction. The objects extracted from the original data are stored in the object library. The objects in the object library are the basis of situation extraction. The discovery of attack scene is a typical situation extraction process. Elia, G. studied the alarm correlation method based on alarm aggregation for the attack scenarios [11].

Fourth, the related technologies of threat assessment. In the field of network security situation awareness, threat assessment belongs to the category of situation assessment. At present, the mainstream situation assessment methods include knowledge-based reasoning method, statistical analysis method and so on. The methods of knowledge-based reasoning include Bayesian network, D-S evidence theory and so on. The methods based on statistical analysis include weight analysis method and analytic hierarchy process. In general, Lappin, Y. and others found that the current research on network security situational awareness is still in the preliminary exploratory stage, and the network security situational awareness technology is not mature and needs further research [12]. For many years, the information security of enterprises has been in the passive cycle of "defense discovery repair". The common practice is to find the loopholes or risks in the network and information system as early as possible and repair them in time through penetration test or risk assessment. At the same time, when the attack behavior is found, the attack behavior is determined by analyzing the relevant security device logs and network traffic, and the attack is blocked as soon as possible. In this passive defense infosec life cycle as shown in Figure 2.1, the vast majority of enterprises focus 95% on defense and 5% on discovering attacks. Basically, repair is based on the passive repair of patches issued by the original manufacturer of products / devices, and the second is to continuously optimize and improve the defense strategy and improve the defense ability. With the deepening of enterprise information security construction, the defense means of information security are also gradually strengthened. Most enterprises have built security systems such as terminal man-

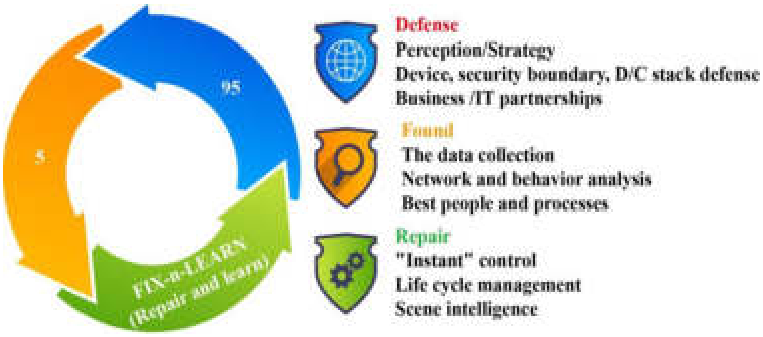


Fig. 2.1: Infosec life cycle

agement, network anti-virus, access control, security audit and vulnerability discovery, which ensure the safe operation of business to a certain extent. However, each system has its own way and is independent of each other, so it is impossible to achieve unified management, unified early warning, unified tracing and traceability. On the other hand, due to the large-scale network of large enterprises, there are a large number of logs in different formats, such as Syslog log, web service log, firewall log, NetFlow log, etc. these logs come from various business system servers and many security devices and network devices, which are widely distributed and large in number. These log data are often not effectively managed and fully utilized, and can not give full play to the analysis role of logs, especially without high-speed collection, normalized storage and correlation analysis of all logs. Shibuya, Y. and others found that in recent years, the more advanced and advanced the technology is, the more attacks the enterprise network faces. Moreover, with the continuous application of new technologies, the means and methods of attack are becoming more and more hidden and difficult to find [13]. Advanced persistent threat (APT) is a complex and covert attack means that can bypass various traditional security detection and protection measures and realize fixed-point attack through careful camouflage, long-term latency and continuous penetration. From the current research on enterprise information security situational awareness system and the current situation of information security protection in China, enterprise information security has been in a passive cycle of discovery and repair. Enterprises generally install corresponding defense systems in combination with their own work characteristics and production nature, find hidden dangers and risk problems in the network system as soon as possible through risk assessment, penetration test and other methods, and take targeted measures to solve them. After discovering the offensive behavior, the system will comprehensively investigate and analyze the whole security equipment log and network traffic, so as to determine the specific degree of behavior, and solve these problems as much as possible. In the enterprise passive circulation information security defense system, the vast majority of enterprises pay more attention to the defense process, but ignore the determination and analysis of the cause of the attack.

Verizon Data Breach Report is shown in Figure 2.2. The current attacks can lead to the leakage of enterprise data and even system paralysis in a few minutes or hours, while it takes weeks or even months for enterprises to find these attacks and effectively stop them. This makes the enterprise's network, system and data in a dangerous state for a long time, and after the old vulnerabilities are repaired, the attacker will find and exploit new vulnerabilities, resulting in the information security personnel struggling to cope with [14].

Author details the system hardware configuration optimization and the AI synchronous operation mechanism. The information security situation inference algorithm is detailed and improved on the basis of the data support vector. The universal data security features are extracting and the information security situation awareness are set and designed the system software structure. It is observed by the author that the information security situation awareness system based on big data and AI has improved the efficiency significantly and high accuracy as compared to the existing techniques [15]. Authors in this paper provide a comprehensive

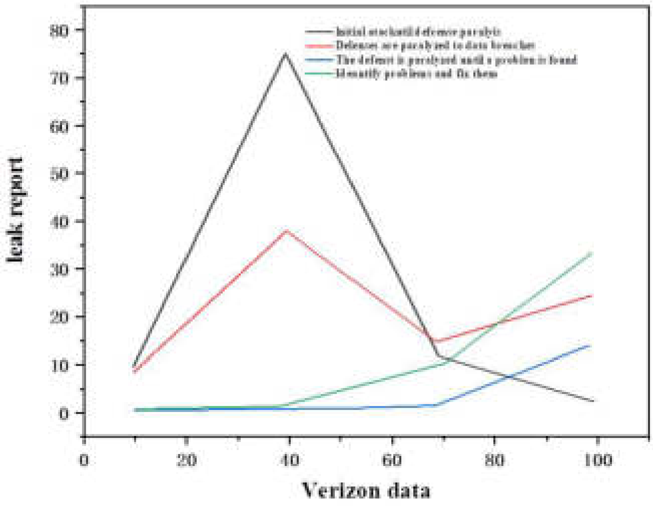


Fig. 2.2: Verizon Data Breach Report

study on the existing literature in the cyber SA for discussing the key design principles, classifications, and analysis of the techniques, and evaluation techniques [16]. Author in this paper details the security situation awareness technology which has become a new research topic in network security. A new cascaded network security situational awareness model is designed based on the fusion algorithms. An induction algorithm is also introduced for the decision tree generation on the pre-processed data for data classification. A new network security situation awareness model is shown by the results [17]. Author in this paper proposed a Power Grid Information Security Perceptual System based on AI technology. The encryption and decryption calculation method are combined and the credible risk assessment theory of dynamic cycle is established. The passive defense of power grid information security problem is solved by it and the power data risk is strengthened and the reliability of information security system of power grid is enhanced [18]. Author in this paper utilized the data mining techniques to study the power control system network security situation awareness technology. The wavelet neural network analysis method is utilized by combining the operational data collection and integrated processing. Finally, calculate the network security status through deep learning and it is concluded by the author that the AI algorithm based on wavelet NN can be utilized for power control system network security situation awareness [19]. Authors in this paper discussed the special issue of six papers on situation awareness in human machine interactive systems in teams of collaborating humans and AI [20].

The author in this paper using a big data-related technologies to analyze, filter, merge, and identify known and unknown security threats and builds a new cascaded network security situational awareness model on the basis of traditional and fusion decision tree algorithms [21]. A decision tree is generated by using the induction algorithm on the preprocessed data for the data classification according to the decision rules. A new network security situation awareness model is constructed by using decision tree calculations. Author in this paper constructed a network security situation awareness framework suitable for big data. A gate recurrent unit (GRU) model is established to extract features from the situation dataset through the deep learning algorithm [22]. This method has a good awareness effect on network threats by the experimental results and has strong representation ability. It effectively perceives the network threat situation which verifies the effectiveness of paper which improves the accuracy of security situation awareness. Author in this paper presents a NSSA that can bridge the current research status and future large-scale application and discuss the classic use cases of NSSA [23]. Finally, various challenges and potential research directions related to NSSA is detailed by the survey. In this paper, author summarizes the artificial intelligence and network security situational awareness classic models to provide artificial intelligence overview [24]. Starting from the machine learning, it introduces the neural-network-based network security situational awareness. Finally, summarizes the future development trends of network security situational awareness. In this paper, author presents the information security situation inference algorithm. By extracting the security features of the data source, the system software structure is designed [25]. The steps of comparison and security feature parameters are added to the information security situation awareness process. Finally, the optimal design of the information security situation awareness system is designed optimally. It is observed from the results that the information security situation awareness system on the basis of big data and artificial intelligence has improved significantly.

2.1. Research Gaps. There is rapid growth of the security threats faced by enterprises and the security attacks technology is also established at a higher level. With the advent of the Internet era, people's lifestyles have undergone earth shaking changes. All walks of life are constantly upgrading and transforming under the impact of the Internet. At the same time, all kinds of network security threats in the Internet era are also increasing. In recent years, cyber attacks with national and organizational backgrounds are increasing. The special roles of the government, military, finance and large enterprises often face more external attack threats. Therefore, the research on information security situational awareness system is necessary.

3. Methods. To combat diverse security threats, the methodology devised an information security situational awareness system after extensive research, finalizing its model. It identifies security data sources from enterprise intranet and the internet, including equipment alarms, logs, and threat intelligence. Conducting correlation analysis of internal and external data, the system detects and verifies security attacks and assesses risks using asset vulnerability dimensions. Results are displayed for monitoring. For advanced persistent threats (APTs), the system employs big data threat intelligence to search historical intranet data, visually presenting APT events for analysis and action, thus providing a comprehensive solution to security challenges.

3.1. Overall Design. In order to deal with various information security threats faced by enterprises, Power China launched the research on information security situational awareness system. After a lot of research and demonstration, the overall model of the system is finally determined. The security data sources of enterprise intranet mainly include security equipment alarm, equipment log (network equipment, server, application, etc.), intranet security evaluation data, network traffic data at the boundary of important areas of the network, etc. Internet security data sources mainly include commercial and open source threat intelligence data from the Internet, Internet security public opinion and vulnerability monitoring data. The security situational awareness system conducts correlation analysis of internal and external security data, determines security attacks and verifies them. At the same time, combined with the dimensions of asset vulnerability, it uses the risk assessment model for comprehensive risk assessment, and finally sends the risk assessment results to the threat situation display module for display. For advanced persistent attack (APT), it mainly relies on the threat intelligence of big data, searches the historical data stored in the enterprise intranet, finds the possible unknown threats in the intranet, and visually displays the found apt attack events in the situation display module.

3.2. Proposed System platform design. The working principle of information security situation awareness system is to analyze and perceive the relevant information security situation as shown in Figure 3.1. To this end, we have built a core platform for distributed computing based on distributed storage and big data processing technology. The core technologies of the platform mainly include:

(1) Hadoop distributed file system (HDFS). We use Hadoop distributed file system as the file system of the system. HDFS file system has the characteristics of high fault tolerance and can be deployed on low-cost PC servers. HDFS relaxed the requirements for POSIX, so that we can access the data in the file system in the form of stream. On the other hand, HDFS also supports large-scale cluster deployment, so that the demand for high-throughput concurrent data access can be solved through unlimited expansion of nodes.

(2) HBase – Hadoop database. We use NoSQL database running on HDFS - HBase as the database of the system. HBase has the characteristics of high reliability, high performance, column oriented and scalability [30].

Row key: each row has a unique row key. The row key has no data type. The row key is a byte array.

Column cluster: data is organized into column clusters in rows. Each row has the same column cluster, but between rows, the same column cluster does not need to have the same column modifier. In the

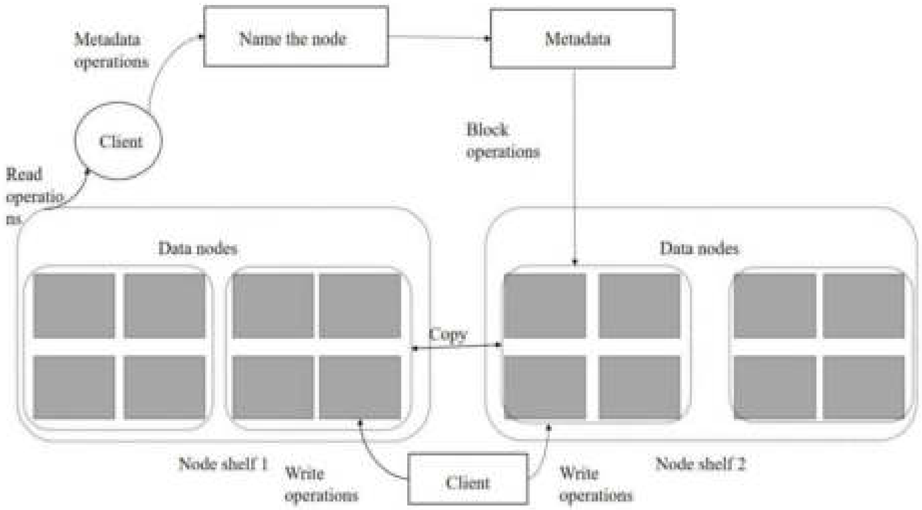


Fig. 3.1: HDFS Distributed File System

database engine, HBase stores column clusters in its own data files, which are defined in advance.

Column modifier: a column cluster defines a real column, which is called a column modifier. The column modifier is the column itself.

Version: each column can have multiple configurable versions. HBase obtains data through the version specified by the column modifier.

HBase, a data definition, storage and use mode based on columns rather than rows, is very suitable for dynamically adding data attributes. Through HBase, a large table can be created, and the attributes of this table can be dynamically added according to needs, especially suitable for the processing of unstructured data.

4. Results and Analysis. The data sources that the system needs to collect include: network traffic image data, log data, security intelligence and support data. Among them, the log data is relatively standard. You can export syslog log, web service log, firewall log, Net-Flow log, etc. by configuring the logs of relevant devices and servers [33-35]. At present, there is no unified standard for security intelligence and support data. We normalize the security intelligence data into the intelligence data that the system can identify and use. At the same time, we regularly update the intelligence base by synchronizing the cloud server or upgrade package, and store all kinds of intelligence and support data in the system for system processing and analysis.

The large-scale data acquisition and processing platform must have the ability of multi-point data acquisition and fault tolerance, especially for the large-scale data acquisition and processing center. The system pre-processes the original image traffic, uses multi-core parallel processing means to analyze, restore and analyze the original network data with large traffic, and then forms a unified traffic log format and uploads it to the big data platform for storage. The architecture of flow acquisition probe is shown in Figure 4.1.

One of the characteristics of the data layer of the attack characteristic event map is that a single attack characteristic event is a weakly connected branch of the whole attack characteristic event map. If E represents the attack characteristic event map and G represents a single attack characteristic event.

From the perspective of set, E represents the complete set, and G is a division of the complete set E. This design can facilitate the system to traverse each weakly connected branch of the attack characteristic event map when discovering the attack behavior in the later stage. The traffic collection probe is mainly divided into two modules. The basic traffic processing module is responsible for preprocessing the original traffic, including basic packet reorganization and traffic reorganization, and can analyze the information of traffic transmission layer and network layer; The high-level protocol processing module is also divided into abnormal behavior discovery,

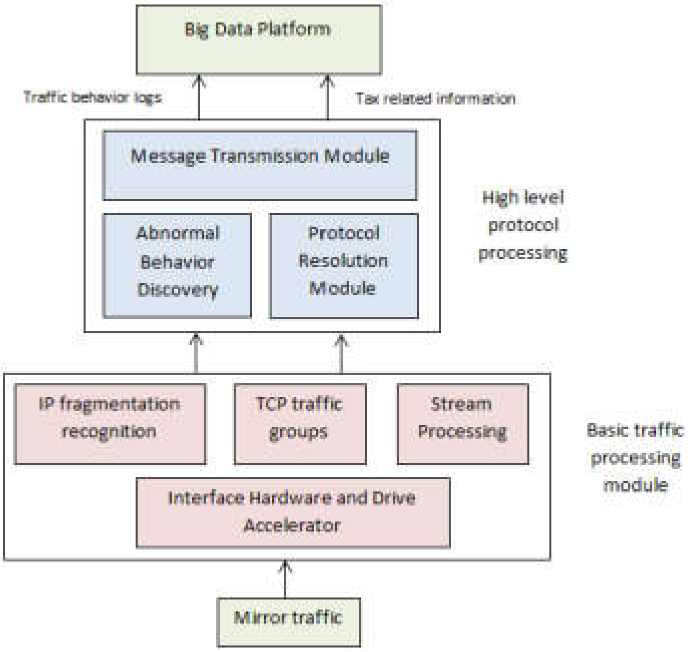

Big Data Platform
Traffic behavior logs
Tax related information
Message Transmission Module
Abnormal Behavior Discovery
Protocol Resolution Module
High level protocol processing
IP fragmentation recognition
TCP traffic groups
Stream Processing
Interface Hardware and Drive Accelerator
Basic traffic processing module
Mirror traffic

Fig. 4.1: Traffic collection architecture diagram

protocol resolution and message transmission modules. The protocol resolution module is responsible for in-depth resolution of application layer protocols, analyzing the information of application layer protocols such as HTTP, DNS and SMTP, and extracting key information to the message transmission module. At the same time, restore the files contained in HTTP, SMTP and other protocols, and send the restored information to the big data platform for saving.

This technology mainly adopts the optical splitter image or network port image technology to export the traffic in the network, and then input it to the analysis platform for correlation analysis. Traffic restoration and data analysis can perform high-performance analysis on mainstream protocols such as HTTP and SMTP / POP3 in IPv4 / IPv6 network environment, and restore the files transmitted by mainstream P2SP software through fragment file detection and P2SP reorganization.

(1) Port matching. In the process of network protocol development, a series of standard protocol specifications have been formed, which stipulate the ports used by different protocols. Although some other widely used applications do not have standardized ports, they have formed defacto standard ports. Port matching is to use TCP / UDP ports to identify behaviors according to the corresponding relationship between standards or factual standards. This method has the advantages of high detection efficiency, but it is easy to be forged. Therefore, on the basis of port detection, it is necessary to add the judgment and analysis of feature detection to further analyze the data.

(2) Traffic feature detection. There are two kinds of traffic feature detection. One is the identification of standard protocol traffic. The standard protocol stipulates a unique message, command and state migration mechanism. These traffic can be accurately and reliably identified by analyzing the proprietary fields and states of the application layer in the traffic packet; The other is the identification of undisclosed protocol traffic. Generally, it is necessary to analyze the protocol mechanism through reverse engineering and identify the communication traffic directly or through the characteristic field of message flow after decryption.

(3) Automatic connection and association. With the development of Internet applications, more and more data are transmitted on the Internet.

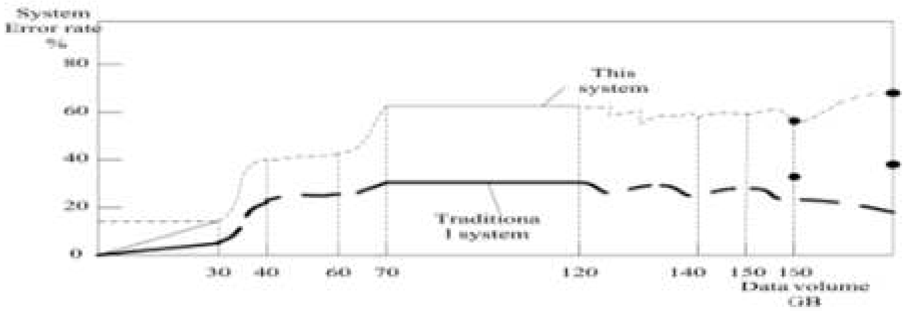


Fig. 4.2: Efficiency comparison result of system operation

(4) Behavior characteristic analysis. For some data flows that are not easy to restore, we use the method of behavior characteristics for analysis, that is, the system does not try to analyze the data on the link, but uses the statistical characteristics of the link, such as the number of connections, the connection mode of a single IP, the proportion of upstream and downstream traffic, packet transmission frequency and other indicators to distinguish the data flow. Because our core platform adopts the stream framework based on big data technology, we can stream all kinds of data according to the predetermined process to ensure the accuracy of all kinds of data processing. Stream framework is a distributed structure that supports horizontal expansion. By adding cluster nodes, the concurrent processing ability of the cluster can be improved. Stream framework also has automatic fault tolerance mechanism, which can automatically handle process, machine and network exceptions to ensure the stable operation of event processing process. When processing data, the data is not written to the disk and cached in the memory of each node. Our core competence platform has the characteristics of low delay and strong real-time. By presetting the event processing topology, we can quickly build the event processing process, and build the corresponding event processing topology model according to different processing requirements to meet the business requirements. Efficiency comparison result of system operation is shown in Figure 4.2. The data security processing accuracy was tested and compared with the security situation awareness system and the result was as follows:

The big data and AI technology information based security situation awareness system has improved the security awareness and prediction accuracy significantly. The System operation accuracy contrast detection is shown in Figure 4.3.

The information security situation awareness based on the big data and AI has improved effectively. The prediction accuracy of the massive data is also improved.

5. Conclusion. In the realm of networking, the volume of data and the accuracy of awareness have emerged as pivotal concerns across various sectors. Traditional security perception systems often suffer from inadequate perception, defense accuracy, and operational efficiency. To address these issues, optimization of information security situational awareness systems leveraging big data backgrounds has been undertaken, with AI technologies ensuring enhanced accuracy and system efficiency, thereby bolstering network information environment security. Presently, the structure of information network security situational awareness systems grounded in artificial intelligence primarily encompasses stages such as information extraction, pre-processing, fusion, situational awareness, and assessment. Key performance indicators during system operation include basic operation, network vulnerability, and threat indicators, which underpin the functioning of situational

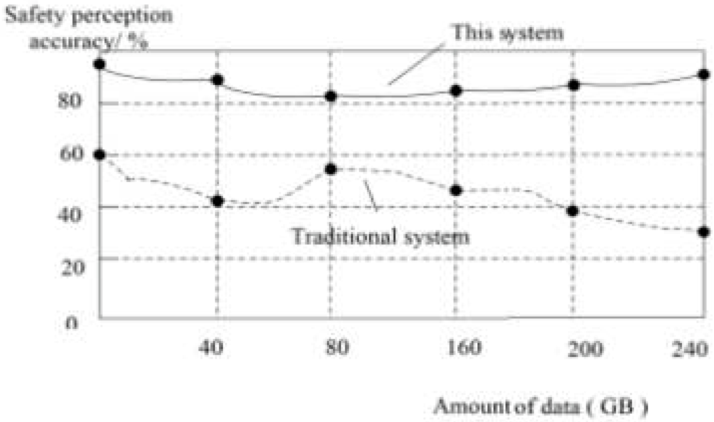


Fig. 4.3: System operation accuracy contrast detection

awareness systems. This technology advancement not only fortifies information network security but also integrates various technologies like data mining, fusion, and pattern recognition, effectively addressing early-stage security issues. Its significance extends to fostering the safe and reliable operation of power systems, thereby supporting societal production and daily life activities. Given the indispensable role of network information technology across diverse sectors, enterprises must prioritize innovation in information security technology and adeptly apply information management skills to ensure orderly progress and enterprise development. Looking ahead, optimizing decision tree algorithms will be considered to bolster models, overcoming local optimization limitations and enhancing efficiency.

REFERENCES

[1] Yu, K., Ming, F., Chen, X., Srivastava, G., *Secure and resilient artificial intelligence of things: a honeynet approach for threat detection and situational awareness*, IEEE Consumer Electronics Magazine, 1, 2021.

[2] Choi, H. T., Yoon, K. J., Kim, H., Park, S. T., Kim, J., *Design and preliminary results of novel situational awareness system for autonomous ship based on artificial intelligence techniques*, Journal of Institute of Control, 27 (8), 556-564, 2021.

[3] Kou, G., Wang, S., Tang, G. , *Research on key technologies of network security situational awareness for attack tracking prediction*, Chinese Journal of Electronics, 28(01), 166-175, 2019.

[4] Xu, R., Nagothu, D., Chen, Y. , *Decentralized video input authentication as an edge service for smart cities*, IEEE Consumer Electronics Magazine, 99, 2021.

[5] Cohen, R. S., *Fast-forward with 5g.* , Air Force Magazine, 102(6), 41-45, 2019.

[6] Chan, J. L., Purohit, H., *Challenges to transforming unconventional social media data into actionable knowledge for public health systems during disasters*, Disaster Medicine and Public Health Preparedness, 14(3), 352-359, 2020.

[7] Korolyov, V., Ogurtsov, M., Khodzinsky, A., *Statement of the problem of complete set of uav group on the basis of models of granular calculations and fuzzy logic*, Cybernetics and Computer Technologies(2), 25-38, 2021.

[8] Khairy, D., Abougalala, R. A., Areed, M. F., Atawy, S. M., Amasha, M. A. ., *Educational robotics based on artificial intelligence and context-awareness technology: a framework*, Journal of Theoretical and Applied Information Technology, 98(1817-3195), 2227-2239, 2020.

[9] Azar, R., *Substations: transformations and improvements* , IEEE Power and Energy Magazine, 17(4), 108-105, 2019.

[10] Elsheikh, A., Alzamili, H. H., Al-Zayadi, S. K., Alboo-Hassan, A. S. Munir, A., Kwon, J., Lee, J. H., Kong, J., Muhammad, K. , *Fogsurv: a fog-assisted architecture for urban surveillance using artificial intelligence and data fusion*, IEEE Access, PP(99), 1-1, 2021.

[11] Elia, G., Margherita, A. , *A conceptual framework for the cognitive enterprise: pillars, maturity, value drivers*, Technology Analysis and Strategic Management(4), 1-13, 2021.

[12] Lappin, Y., *Israel's carmel future afv programme unveiled*, Jane's Defence Weekly, 56(33), 19-19, 2019.

[13] Shibuya, Y., Tanaka, H. , *Using social media to detect socio-economic disaster recovery*, IEEE Intelligent Systems, 34(3),

29-37, 2019.

[14] MS Zitouni, Sluzek, A., Bhaskar, H., *Visual analysis of socio-cognitive crowd behaviors for surveillance: a survey and categorization of trends and methods*, Engineering Applications of Artificial Intelligence, 82(JUN.), 294-312, 2019.

[15] Little, B. D., Frueh, C. E., *Space situational awareness sensor tasking: comparison of machine learning with classical optimization methods*, . Journal of Guidance, Control, and Dynamics, 43(5), 1-12, 2019.

[16] Bao, H., He, H., Liu, Z., Liu, Z. *Research on information security situation awareness system based on big data and artificial intelligence technology.* In 2019 International conference on robots intelligent system (ICRIS) (pp. 318-322). IEEE, (2019, June).

[17] Alavizadeh, H., Jang-Jaccard, J., Enoch, S. Y., Al-Sahaf, H., Welch, I., Camtepe, S. A., Kim, D. S. *A Survey on Threat Situation Awareness Systems: Framework, Techniques, and Insights.* arXiv preprint arXiv:2110.15747, 2021.

[18] Yao, F. *Information Security Situation Awareness Based on Big Data and Artificial Intelligence Technology.* Wireless Communications and Mobile Computing, 2021.

[19] Xie, M., Chen, Z. *A Situation Awareness System for the Information Security of Power Grid.* Journal of Computers, 31(1), 192-198, 2020.

[20] Zhao, J., Li, X., Cao, Y., Liu, J., Yan, J., Li, C. *Analysis and Application of intelligent Power Control System Cyber Security Situation Awareness Based on Wavelet Neural Network.* In Journal of Physics: Conference Series (Vol. 2078, No. 1, p. 012067). IOP Publishing, (2021, November).

[21] Yao, F. Information Security Situation Awareness Based on Big Data and Artificial Intelligence Technology Wireless Communications and Mobile Computing, 2021, 1-6, 2021.

[22] Wen, Z., Zhang, L., Wu, Q., Deng, W. A Network Security Situation Awareness Method Based on GRU in Big Data Environment. International Journal of Pattern Recognition and Artificial Intelligence, 37(01), 2251018, 2023.

[23] Zhang, J., Feng, H., Liu, B., Zhao, D. Survey of Technology in Network Security Situation Awareness Sensors, 23(5), 2608, 2023.

[24] Wang, M., Song, G., Yu, Y., Zhang, B. The Current Research Status of AI-Based Network Security Situational Awareness Electronics, 12(10), 2309, 2023.

[25] Bao, H., He, H., Liu, Z., Liu, Z. Research on information security situation awareness system based on big data and artificial intelligence technology. In 2019 International conference on robots and intelligent system (ICRIS) (pp. 318-322). IEEE, 2019.

Edited by: Pradeep Kumar Singh
Special issue on: Intelligent Cloud Technologies Enabled Solutions for Next Generation Smart Cities
Received: Nov 12, 2022
Accepted: Mar 1, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1311–1320, DOI 10.12694/scpe.v25i3.2068

NETWORK-BASED MECHANICAL VIBRATION FAULT DIAGNOSIS SYSTEM

QIUCHEN ZHANG [*]AND XIAOXIA JIN [†]

Abstract. Nowadays researchers are investing in electrical machines fault diagnosis area. The users and manufacturers are strong for containing diagnostic features for reliability and scalability improvement. The regular monitoring enables machine faults early detection and hence helpful for automation by providing process control. The fault detection performance and machine-learning algorithms classification are highly dependent on features involved. The aim of this paper is to solve the network mechanical vibration problem and for that a research on mechanical vibration fault diagnosis is proposed. The first is to use the vibration signal receiving device to record the vibration signal of the target device. In the process of receiving the signal, the measurement point is related to the accuracy of the received signal, so it is necessary to prepare the measurement point. Secondly, the principle of fault detection based on vibration detection is introduced. The main purpose of this method is to identify the fault characteristics, simulate the fault with MATLAB, and obtain the error time-frequency diagram behavior. The feature vector dimension obtained by the idling confirmation example is the same as that of the rotor, which is 14, including 8 relative wavelet packet intensity entropy feature indexes and higher values, minimum value, peak-to-peak value, mean value, mean square error and variance. Finally, the deficiencies of the detected vibration faults are identified and similar improvements are proposed. Improvements only reduce signal vibration, disrupting feature isolation and identifying patterns.The observed percentage accuracy for classification of faults through proposed approach is 98.2%.

Key words: computer network; teaching management; JSP technology; system design

1. Introduction. The equipment in the network operation is accompanied by vibration, and the potentially defective parts will also vibrate when moving, especially the defects of many mechanical parts; Equipment fault diagnosis mainly includes oil analysis, vibration signal analysis, infrared thermal imaging and other methods, at present, vibration signal analysis is the most widely used method, vibration signal analysis is to use sensors to detect mechanical vibration in the form of electricity, after amplifying and filtering the input to the analysis processor, and then analyzing it, a series of processes of artificially extracting the fault characteristic signal, therefore, through vibration signal analysis, we can find out the problems that are occurring in the equipment, at the same time, the comparison of the vibration data and signal energy of the periodic test and the measuring point is used to find the deterioration trend of the equipment, and to provide a basis for the annual maintenance of the equipment [1]. In order to obtain a good frequency response range, the sensor and the measuring point of the device are installed by magnetic suction (rubidium magnet). The test frequency range can reach 10000 Hz, the equipment is tested in this way, indicating the effectiveness of the vibration analysis method for fault diagnosis, as shown in Figure 1.1.

The acquisition of fault information is the first step to realize fault diagnosis of mechanical equipment, and it is an important basis for fault diagnosis work. The acquisition of fault information is a technology of signal detection and quantification of the working parameters, performance indicators, related physical quantities and other information of the mechanical equipment itself, the sensor is a device that obtains various information and converts it into an electrical signal, and is the key and main means of obtaining fault information. The fault characteristics of machinery are often reflected in the vibration condition, the use of vibration signals to diagnose equipment is the most effective and commonly used method in fault diagnosis. The fault diagnosis and its protection background are as earliest as the machineries themselves. Initially, the machineries manufacturers and their users relied on modest safety such as protection from over voltage. The reliability and safety operation are ensured by this precaution. However, the machinery becomes more complex with increase in number of tasks. Therefore, the diagnosing faults improvement is the requirement and the fault diagnosis becomes very

[*]Xinxiang Vocational and Technical CollegeXinxiang, Henan, 453006, China (qiuchen- zhang223344@gmail.com).
[†]Xinxiang Vocational and Technical College Xinxiang, Henan, 453006, China

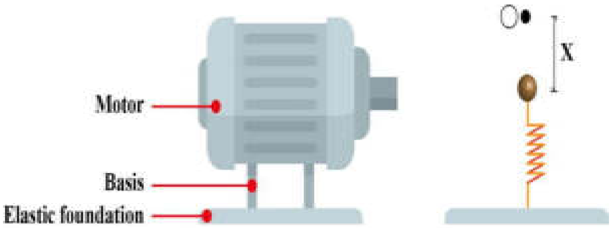


Fig. 1.1: Fault diagnosis of network mechanical vibration [2]

important because the machine unwanted downtime can cause loses.

A flexible system for communication and transmitting the information is provided by the industrial wireless sensor. The wireless sensor networks (WSN) need more attention in service parameters quality such as energy consumption, information and cost reliable transmission for obtaining good performances. The wireless sensor networks and IoT based application is challenging for managing large amount of real time data. The Fault detection and its diagnosis are important for the efficiency of machinery maintenance. The small and large rotating machinery necessity in industrial systems enforces monitoring, maintenance, and reparation. The condition monitoring necessity is rotating machines to provide machines condition knowledge at each moment without production stopping. Common techniques like vibration monitoring is one of the best condition monitoring for detection, location and distinguish faults before they become critical and dangerous. The most essential mechanical of rotating machinery elements is bearing. The rotating shafts are supported by them and on the other side the mechanical faults in rotating machinery are shown by several studies. Therefore, the level of production and equipment is influenced by the bearings fault as well as having an unsafe environment. The condition monitoring, early fault detection and fault diagnosis of these bearings is main fundamental axes of industrial research.

Contribution.

1. In order to solve the problem of network mechanical vibration, a research on mechanical vibration fault diagnosis system is proposed.
2. The first is to use the vibration signal receiving device to record the vibration signal of the target device. In the process of receiving the signal, the preparation of the measurement point is related to the accuracy of the received signal, so it is necessary to prepare the measurement point.
3. Secondly, the principle of fault detection based on vibration detection is introduced. The main purpose of this method is to identify the fault characteristics, simulate the fault with MATLAB, and obtain the error time-frequency diagram behavior.
   The rest of the paper is organized as follows. Section II provides an overview of the exhaustive literature survey followed by a methodology adopted in section III. Proposed method is detailed in section IV and the obtained results are discussed in section V. Finally, concluding remarks are provided in Section VI.

2. Literature Review. The safe and stable operation of machinery and equip- ment is a prerequisite for ensuring normal production. Therefore, it is of great signif- icance for the smooth progress of production to accurately grasp the operating status of machinery and equipment, and to detect and eliminate equipment failures in time. At present, mechanical vibration signal analysis has become one of the main methods for judging the operating status of mechanical equipment [1]. Moreover, with the increasing maturity of network technology, the research on network-based equipment condition monitoring system continues to deepen, it has outstanding advantages in resource sharing and remote monitoring. To this end, the author designs a network-based mechanical vibration signal analysis system for judging the running state of mechanical equipment. In 1984, Gai, J et al. applied the HICLAS fault early diag- nosis device developed by the Japan Construction Machinery Co., Ltd. to carry out early diagnosis of the oil pump, the wear condition inside the oil pump can be directly detected from the oil pump indication in a short period of time, and the device can judge the working life of the oil pump to continue running, make failure prevention possible [2]. In 1992, Glowacz, A used the detection of the vibration velocity of the free end bearing of the centrifuge and the comprehensive trend diagram, and analyzed the spectral characteristics, a centrifuge outer bowl imbalance fault was diagnosed.

Neural network has the ability to deal with complex multi-pattern matching and is an effective fault diagnosis method [3]. In 1996, Kim, H. E., Hwang, S et al. used BP neural network to effectively identify inner ring defective bearings, outer ring defective bearings, roller defective bearings and some comprehensive fault characteristics, which can improve the efficiency of fault diagnosis [4]. In 2004, Shan, P. et al. used the fast ICA algorithm to separate the vibration signal of the bearing, and then compared its power spectrum with the spectrum of the original vibration signal, the results show that ICA is easier to achieve early diagnosis of faults [5]. Same year, Yang, J. et al. proposed a new time-frequency analysis method with adaptive characteristics of local wave time-frequency distribution, the rubbing, misalignment and early fault signals that are common in rotating machinery are analyzed, and the frequency spectrum analysis is compared [6].

In 2006, Sun, Y. et al. combined wavelet filtering and cy- clostationarity analysis method, and first carried out Morlet wavelet defect bearing failure on the original vibration signal, moreover, the severity of the fault can be identified within a certain error range [7]. On the basis of the current research, a research on mechanical vibration fault diagnosis system is proposed. Through actual data acquisition, the accuracy of the system acquisition module is verified. Through the analysis of unbalance fault signal, oil film whirl fault signal, vibration signal of rotor speed up and down process and rolling bearing fault signal, the correctness of the vibration signal analysis module is verified. The unbalanced rotor is dynami- cally balanced by the system dynamic balance module, and a good balance effect is obtained. The system has good expansibility, and can form a vibration signal acquisi- tion network through combination. The author's research has good application value in remote acquisition and analysis of vibration signals of mechanical equipment and rotor dynamic balance. Author presented the scattering transform utilizing machine learning for translational, rotational and deformation extraction for the first time from vibration signals found from rolling element bearings (REBs). The scattering transform core idea lies in scattering network construction which is formed from a signal processing layers. The association of a linear filter bank associate each layer and utilizes wavelet filter bank, modulus rectifiers and averaging operators cascading for deep convolution network and multi-scale co-occurrence coefficients are computing. Features are extracted as scattering transform coefficients from vibration signal prognosis data repository then input to a support vector machine (SVM) classifier. To obtain distinguishing features, test results analysis and solutions are utilized [8]. Author discussed that due to the potential advantages machine fault diagnostic and prognostic techniques have been the considerable for condition-based maintenance system from reducing downtime and increasing machine availability. Research on ma- chine fault diagnosis and prognosis for the past few years has been developing quickly.

Author summarizes and recent published techniques classification of rotating machinery in diagnosis and prognosis. Furthermore, opportunities as well as the challenges are also discussed for conducting the research machine prognosis field [9]. Author in this paper presents the relevant features extraction based on oriented sport vector machine (FO-SVM) and it is capable for extracting the most relevant feature set. The most relevant features extraction before classification process in higher classification accuracy. As observed the presented technique consumes less time for cloud. The presented approach provides prediction of fault accurately based on cloud platform by utilizing the industrial wireless sensor networks. In this paper, author discussed the bearing vibration frequency features for motor fault diagnosis. This paper presents an approach using NN and time/frequency-domain for vibration analysis. The Vibration simulation is utilized in the design of various motor rolling bearing fault diagnoses. The result obtained from the presented technique indicates that NNs can be efficient agents in the various motor bearing faults diagnosis through the measurement of motor bearing vibration [10]. Author presented and implements the identification, diagnosis and common fault remedy techniques utilizing vibration analysis and summaries important techniques utilizing for rotating systems condition monitoring such as fast Fourier transform, frequency domain decomposition method and deep learning [11]. Author in this paper presented a convolutional neural network (CNN) to learn features directly from frequency data of vibration and testing the feature learning performance from raw data, frequency spectrum and combined time-frequency data [12]. The time domain, frequency domain and wavelet domain are used for comparison purpose. The presented method is validated by using gearbox challenge data and a planetary gearbox test rig. This presented method is able to learn features from frequency data adaptively and achieve higher diagnosis accuracy. This paper presents a deep CNN-based transfer learning approach and it consists of two parts; the first part is constructed with a pretrained DNN that extract the features automatically from the input, and the second part is connected stage for the feature extraction that needs to be trained by using gear fault data [13]. Case analyses by utilizing the experimental data from a gear system indicate that the presented approach not only entertains preprocessing adaptive feature extractions, but also requires training data. Author in this paper proposed a selective kernel convolution deep residual network based on the channel-spatial attention mechanism and feature fusion for mechanical fault diagnosis. The model effectively extracts fault features from the vibration signal as compared to conventional deep learning methods, and the fault recognition effectiveness is improved [14]. As compared to other algorithms, the presented method has higher fault identification ability, therefore demonstrating the channel-spatial attention mechanism network advantages and accuracy and the robustness of the model were verified.

2.1. Research Gap. After exhaustive literature survey it is found that there is a problem of network mechanical vibration and existing techniques are unable for early detection of machine faults providing process control. The feature extraction techniques are also not effective and the fault detection performance is highly depend on features involved.

3. Methodology.

3.1. Research trends at home and abroad. The state analysis and fault diagnosis technology of mechanical equipment is a new discipline developed in the middle and late 1960s. In the theoretical and applied research of state analysis and fault diagnosis, some developed countries in the United States, Japan and Europe are at the forefront of the world. Condition analysis and fault diagnosis originated in the United States, and have been widely used in aviation, military, energy, machinery and other departments, and are in a leading position in the world. In the 1970s, the U.S. Department of Defense began research on reliability-centered condition analysis techniques and applied them to aircraft, ships, and vehicles. At present, in terms of rotating machinery state analysis system, the M800A system of SKF Bearing Com- pany in the United States, the Tranmaster2000 system of Bently Nevada Company in the United States, etc., are all representative rotating machinery state analysis systems". Japanese state analysis and fault diagnosis technology began in the 1970s. In 1971, Japan began to develop its own total production maintenance (TPM), and learned about the research work of diagnostic technology from Europe and the United States, which basically reached the practical stage in 1976 [15]. Among the national research institutions, the Institute of Mechanical Technology and the Institute of Ship Technology focus on the diagnostic technology of mechanical basic parts.

Research status of fault diagnosis methods for mechanical equipment The beginning of fault diagnosis technology is the analytical redundancy method proposed by Beard of the Massachusetts Institute of Technology. [16]. After more than 20 years of research and development, although there are many research achievements in mechanical fault diagnosis in my country, fault diagnosis technology is a comprehensive discipline, with the development of fuzzy set theory, genetic algorithm, support vector machine, expert system, neural network technology and wavelet analysis theory. Due to the differences in the feature description and decision-making methods adopted by the system, the current fault diagnosis technology, different diagnoses have been formed [17].

3.1.1. The basic process, principle and method of fault diagnosis of mechanical equipment. During the degradation process of mechanical vibration, it basically follows the well-known "bathtub curve" law, the whole process includes: The running-in period, normal trial period, and wear-out period are shown in Figure 3.1.

Through the necessary measurement and fault diagnosis of mechanical vibration, it is possible to find out which phase the equipment is in at a certain stage in time, so as to prevent the equipment from entering the wear and tear period in advance. Mechanical fault diagnosis technology (Mechanical Fault Diagnosis) refers to in a certain working environment, use the detection device to detect the state information of the mechanical equipment in operation or under relatively static conditions, by analyzing the operating status information of the mechanical equipment to determine whether the mechanical equipment is in a normal operating state, combined with the failure mechanism and historical operating status of the diagnostic object, in order to qualitatively and quantitatively determine the real-time operating status of mechanical equipment and its components, and according to the corresponding fault characteristics to determine the possible faults and fault locations of the

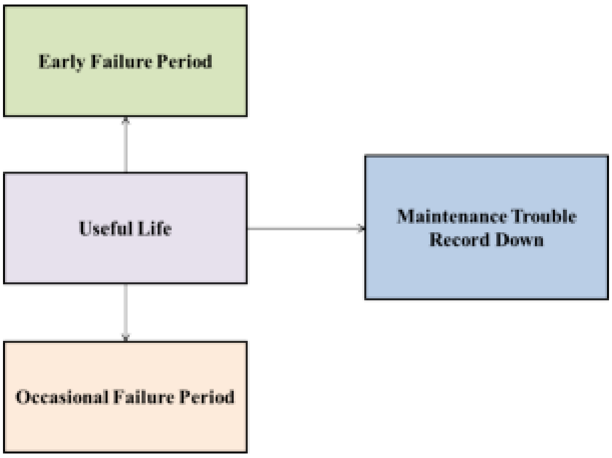


Fig. 3.1: Bathtub Curve

mechanical equipment, predict the operating trend and remaining life of related failures to determine targeted equipment management, maintenance and repair countermeasures.

The purpose of fault diagnosis is to find faults in time and minimize losses. The basic process of mechanical equipment fault diagnosis is shown in Figure 3.2, there are three main steps in the diagnosis process: The first step is to obtain characteristic signals of mechanical equipment status, such as vibration, noise, temperature, pres- sure and other signals; The second step is to extract fault features from the measured characteristic signals, and the extraction of fault features; The third step is the core of the whole diagnosis process, namely: Judging the specific fault of the equipment and forming a maintenance decision.

4. Proposed method. Fault diagnosis mainly includes the following aspects: Research on fault mechanism and fault symptom; Research on fault information acquisition method; Research on signal processing and fault feature extraction methods; Research on diagnostic reasoning methods; Research on the development of fault diagnosis systems.

(1) The program is not good, and the research is not good. A criminal investi- gation is the basis for a misdiagnosis. Defects or failures of equipment are usually caused by signal events during operation. The failure process has studied the causes of failures and the relationship between failures and symptoms, and found general laws through theoretical calculations or experimental studies. As the basis of fault diagnosis technology, only by studying the faults of the detected products can the primary and secondary causes of the faults be distinguished, and a reliable basis for judging and diagnosing faults can be provided. Many experts and scholars at home and abroad have done a lot of theoretical and experimental research on ma- chine tool failures, and made many important decisions, which are conducive to the inspection and testing of defective products. In 1968, American scientist John Sohre gave a general description of the symptoms and causes of machine dysfunction in the form of a table, and clearly and concisely divided criminal behaviors into 9 categories and 37 categories, research results. It is widely used in practice.

(2) Research on the method of fault information acquisition. The acquisition of fault information is the first step to realize fault diagnosis of mechanical equipment, and it is an important basis for fault diagnosis work. The acquisition of fault information is a technology of signal detection and quantification of the working parameters, performance indicators, related physical quantities and other information of the mechanical equipment itself, the sensor is a device that obtains various information and converts it into an electrical signal, and is the key and main means of obtaining fault information. The main physical quantities involved in the detection of mechanical equipment information include vibration, force, sound, rotational speed, temperature, and flow rate. Since the fault characteristics of machinery are often reflected in the vibration condition, the use of vibration signals to diagnose equipment is the

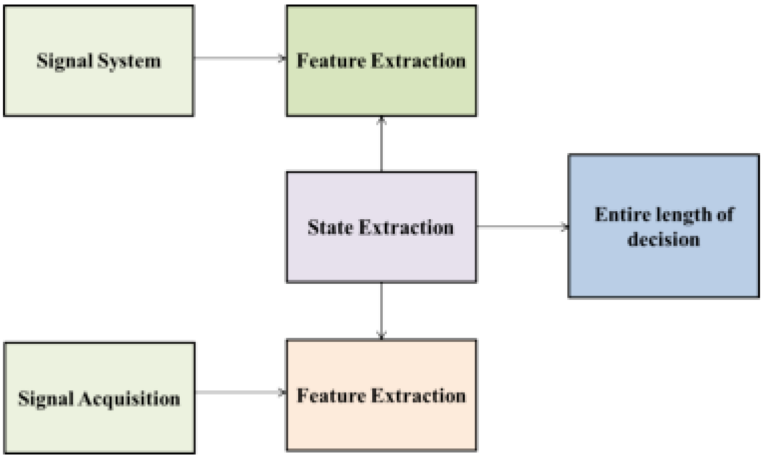


Fig. 4.1: Fault diagram for fault diagnosis

most effective and commonly used method in fault diagnosis.

(3) Research on signal processing and fault isolation. In the development of investigative technology, the most important and most important problem is the elimi- nation of illegal features, which directly affects the accuracy of the investigation, the guilt and the reliability of predicting the crime. To solve the key problem of data leak- age, people can only rely on signal processing, especially the theoretical and technical means of processing signal processing.

(4) Development and research of misdiagnosis. Regardless of craftsmanship and technological output, they will eventually become practical. Fault diagnosis technology is a practical technology. Therefore, while paying attention to theoretical innovation, attention should be paid to the development of a fault detection system with certain practical value. At present, the improvement of the criminal justice system mainly includes the following two aspects: physical examination and online monitoring of criminal investigation.

4.1. Principle of fault diagnosis based on vibration detection. In the fault detection based on vibration detection, the first vibration signal collection device is used to collect the vibration signal of the target device. In the process of receiving the signal, the preparation of the measurement point is related to the accuracy of the received signal, so it is necessary to study the preparation of the measurement point. Different functions of the device will cause changes in the vibration signal, that is, when the device has some abnormality or failure, the vibration signal will change accordingly, and the change in guilt is the sin. It is based on this principle that vibration detection and vibration signal analysis can be used to diagnose the rotor and bearing faults of centrifugal fuel pumps. It can be seen from its principle that the most important part of error detection as vibration detection is to identify fault features, and through different features, detect irregularities. The author will focus on the simple vibration signal receiving system, the time-frequency characteristics of the fuel pump failure theory, and the characteristics of the fuel pump summarized in the gas station.

5. Experiments and Research.

5.1. System Composition. The network-based mechanical vibration signal analysis system relies on the network to realize the acquisition, analysis and rotor dynamic balance of mechanical vibration signals. The system is mainly composed of sensors, data acquisition cards, and vibration signal analysis systems. Among them, the vibration signal analysis system is a software system rooted in the computer, including a vibration signal acquisition module, a vibration signal analysis module, and a rotor dynamic balance module. The block diagram of the network-based me- chanical vibration signal analysis system is shown in Figure 5.1.

In Figure 5.1, the sensor first converts the measured physical quantity into an output analog signal. Subsequently, the data acquisition card preprocesses the analog signal output by the sensor, and performs AID

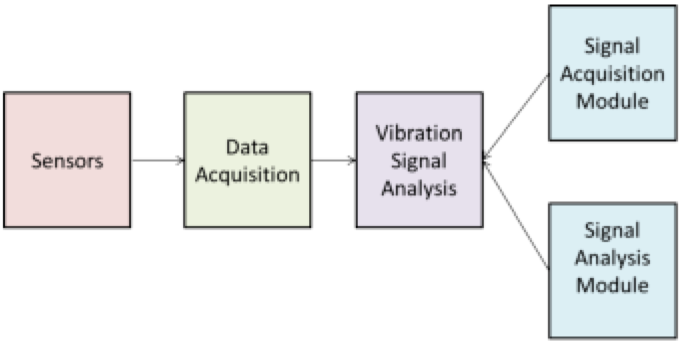


Fig. 5.1: Block diagram of network-based mechanical vibration signal analysis system

conversion on the preprocessed analog signal to convert it into a digital signal. The vibration signal analysis system completes the analysis and processing of digital signals. Among them, the signal acquisition module mainly realizes the acquisition of vibration signals; The signal analysis module can complete the offline analysis of the collected vibration signals; The rotor dynamic balancing module can dynamically balance the unbalanced rotor.

Design of vibration signal analysis module Vibration occurs natu- rally when equipment is in operation, the severity of the vibration is often a precursor to a crash, and the characteristics are usually obvious. Therefore, from the perspec- tive of vibration signal analysis and diagnosis, it is also the main means of maintaining and controlling equipment at present. Time domain analysis and frequency domain analysis are the types of vibration signal analysis, especially frequency domain anal- ysis. is one of the best ways to detect vibration faults. The vibration signal analysis module developed by the author only recognizes offline analysis of stored data, includ- ing: time domain analysis, frequency measurement instrument. Frequency anal- ysis includes: amplitude spectrum, starting power spectrum, cross power spectrum, ZOOM-FFT, envelope spectrum, cepstrum, order spectrum, waterfall plot, Bode plot. 1Time domain analysis The time domain analysis functions of vibration signals in this system mainly include: Waveform display, filtering, probability density analysis, au- tocorrelation analysis, cross-correlation analysis, bar graph analysis, recursive graph analysis, time domain indicator analysis, data playback. 1. Waveform display The waveform display can reproduce the waveform of the collected data of each channel. By observing the time-domain waveform of the vibration signal, the operating state of the mechanical equipment can be estimated. 2. Filtering The actual collected vibration signal usually contains a lot of noise, the superposition of the noise and the useful signal will distort the waveform of the useful signal, it is not conducive to the analysis of useful signals. Filtering the vibration signal can remove the noise in the signal and make the characteristics of the useful signal more obvious, which is beneficial to the waveform analysis of the useful signal.

5.2. Examples of Failure Mode Recognition. Similar to the example of fault pattern recognition of centrifugal oil pump rotor based on BP neural network, when identifying the bearing structure of oil centrifugal pump, the neural network structure of bearing fault structure should be designed first. Similarly, suppose the number of nodes in the input process is n, the number of nodes in the output process is m, and the number of nodes in the input process is o, then designing a neural network includes the following steps: 1. Determine the number n of the input process: the number of input nodes is the size of the eigenvector, the length of the eigenvector used in the fault type identification system is the same as the rotor example, a total of 14, including 8 relative wavelet packet power characteristics features and 6-time fill features such as max, min, high out, mean, squared error, and variance. 2.Determine the number of output layer nodes m: The purpose of determining the output vector is to make for each input sample, there are different output vectors corresponding to the pattern recognition. In this example, the determination principle of the output vector is as follows: When the bearing is set to the normal state, let the network output be y1=[1 000]T; When the bearing is set to the failure state of the rotating body, let the network output be y2=[0 1 00]T; When the bearing is set to the failure state of

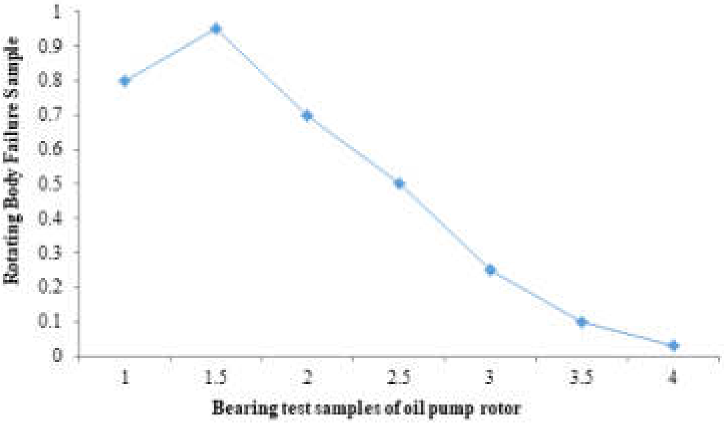


Fig. 5.2: State identification diagram of bearing test samples

the inner ring, let the network output be y3=[00 1 0]T; When the bearing is set to the failure state of the outer ring, let the network output be y4=[0 001]T. Since the fault may be one type of fault or multiple faults occur at the same time, with the deepening of research, it is necessary to continuously adjust the output layer nodes. 3. Set the number of hidden layers o: Since the three-layer BP neural network can identify the mapping from 1 n-wide to 1 m-dimensional space, the author chooses the three-layer BP neural network for fault identification. Types of centrifugal oil pump rotors. The number of latches is 2n + 1, or 29, as shown in Figure 5.2.

5.3. Key Technologies of Fault Diagnosis. Globus is a research and development project of Argonne National Laboratory in the United States, and 12 universities across the United States participated in the project. Globus researches key communication concepts such as resource management, security, data services and data management, develops network tools (Toolkit) that can run on multiple platforms, helps plan and build large-scale projects, develops large-scale network operations at the scale required application. Toolkit is the most important feature of Globus, its first version was launched in 1999. Toolkit is open source, and anyone can download the code from its website. Currently, Globus technology is used in eight applications, including the National Aeronautics and Space Administration Grid (NASA IPG), the European Data Grid (Data Grid) and the US National Technology Grid (NTG). Generally speaking, network computing focuses on large-scale projects, which, according to Globus, require collaboration between multiple organizations, who create ”virtual entities” and are done by Chinese equipment that all organizations participate in virtual organization cooperative.

5.4. Comparative analysis on the basis of performance metrics. The performance of the proposed technique is compared with the existing technique as shown in Figure 5.3 and Figure 5.4 . The observed percentage accuracy for classification of faults through proposed approach is 98.2%. The sensitivity and specificity of the existing SVM technique and the proposed technique are compared and it is obtained that the higher values of specificity and sensitivity are obtained by the proposed technique as compared to the existing technique. The precision, recall and accuracy values are also compared and the improvement is shown by the proposed technique.

6. Conclusion. Through the research on the principle of vibration detection fault diagnosis method, the author finds that its core lies in the identification of fault features. Through theoretical research and the help of MATLAB simulation, the time-frequency characteristics of common faults of centrifugal oil pump are summarized and studied. In addition, due to the gap between the theory and the present, it is theoretically possible to diagnose the corresponding fault according to the time frequency characteristics, however, in the actual operation of the oil station, other characteristics must be integrated in order to have a better identification of the fault. Even so, there are still some deficiencies in the commonly used vibration detection and fault

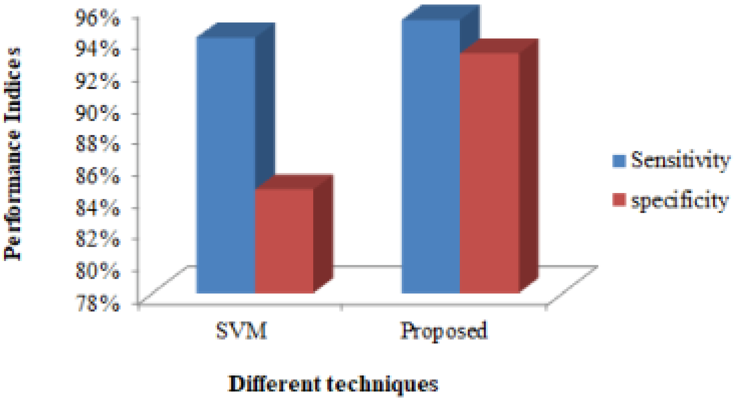


Fig. 5.3: Performance indices: Sensitivity and specificity

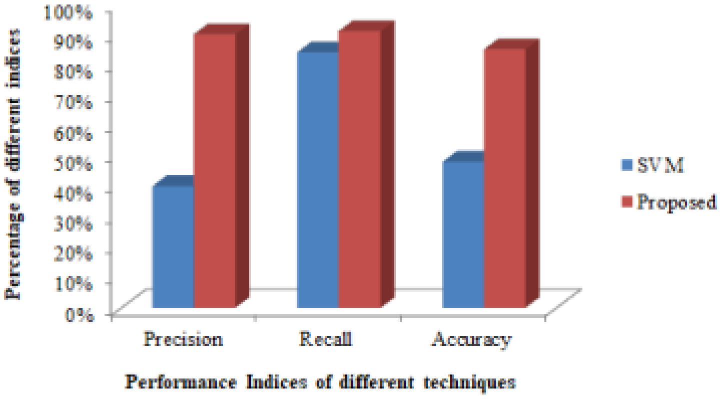


Fig. 5.4: Performance indices: Precision, recall and accuracy

diagnosis methods, therefore, it is proposed that the next step will be to focus on the noise reduction of vibration signals, the feature extraction of wavelet packet energy entropy and the use of BP neural network to identify faults.

REFERENCES

[1] Zhao, W., *Torsional vibration calculation of rolling mill based on topology network*, Chinese Journal of Mechanical Engineering, 42(7), 51-55, 2006.

[2] HU, Y., Li, H. , Zhang, H. , Zhao, Y., *Research on fault diagnosis based on singular value decomposition and fuzzy neural network*, Shock and vibration, 2018(pt.3), 1-7, 2018.

[3] GLOWACZ, A., *Diagnostics of synchronous motor based on analysis of acoustic signals with the use of line spectral frequencies and k-nearest neighbor classifier*, Archives of Acoustics, 39(2), 189-194, 2014.

[4] KIM, H. E. , HWANG, S. S. , TAN, A. , MATHEW, J. , CHOI, B. K., *Integrated approach for diagnostics and prognostics of hp lng pump based on health state probability estimation*, Journal of Mechanical Science Technology, 26(11), 3571-3585, 2012.

[5] SHAN, P. , LV, H. , YU, L. , GE, H. , LI, Y. , GU, L., *A multisensor data fusion method for ball screw fault diagnosis based on convolutional neural network with selected channels*, IEEE Sensors Journal, 20(14), 7896-7905, 2020.

[6] YANG, J. , GAO, T. , JIANG, S. , LI, S. , TANG, Q. , *Fault diagnosis of rotating machinery based on one-dimensional deep residual shrinkage network with a wide convolution layer*, Shock and Vibration, 2020(4), 1-12, 2020.

[7] Sun, Y. , Feng, T. , Jin, Z., *Review on vibration signal analysis of rotating machinery based on deep learning*, Journal of Physics: Conference Series, 1820(1), 012034 (6pp), 2021.

[8] Ambika, P. S., Rajendra kumar, P. K., Ramchand, R., *Vibration signal based condition monitoring of mechanical equipment with scattering transform*, Journal of Mechanical Science and Technology, 33(7), 3095-3103, 2019.

[9] Van Tung, T., Yang, B. S., *Machine fault diagnosis and prognosis: The state of the art*, International Journal of Fluid Machinery and Systems, 2(1), 61-71, 2009.

[10] Zhang, X., Rane, K. P., Kakaravada, I., Shabaz, M, *Research on vibration monitoring and fault diagnosis of rotating machinery based on internet of things technology* Nonlinear Engineering, 10(1), 245-254, 2021.

[11] Li, B., Chow, M. Y., Tipsuwan, Y., Hung, J. C , *Neural-network-based motor rolling bearing fault diagnosis*, IEEE transactions on industrial electronics, 47(5), 1060-1069, 2000.

[12] Jing, L., Zhao, M., Li, P., Xu, X., *A convolutional neural network based feature learning and fault diagnosis method for the condition monitoring of gearbox*, Measurement, 111, 1-10.

[13] Cao, P., Zhang, S., Tang, J., *Preprocessing-free gear fault diagnosis using small datasets with deep convolutional neural network-based transfer learning*, Ieee Access, 6, 26241-26253, 2018.

[14] Zhang, S., Liu, Z., Chen, Y., Jin, Y., Bai, G., *Selective kernel convolution deep residual network based on channel-spatial attention mechanism and feature fusion for mechanical fault diagnosis*, ISA transactions, 133, 369-383, 2023.

[15] ABDULBARY, M. B., EMBABY, A. G., GOMAA, F. R. , *Fault Diagnosis in Rotating System Based on Vibration Analysis*,Engineering Research Journal, 44(3), 285-294, 2021.

[16] Miao, Y., Li, C., Shi, H., Han, T., *Deep network-based maximum correlated kurtosis deconvolution: A novel deep deconvolution for bearing fault diagnosis* Mechanical Systems and Signal Processing, 189, 110110, 2023.

[17] Li, C., Chen, J., Yang, C., Yang, J., Liu, Z., Davari, P., *Convolutional Neural Network-Based Transformer Fault Diagnosis Using Vibration Signals* Sensors, 23(10), 4781, 2023.

Edited by: Pradeep Kumar Singh
Special issue on: Intelligent Cloud Technologies Enabled Solutions for Next Generation Smart Cities
Received: Nov 13, 2022
Accepted: Feb 29, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1321–1331, DOI 10.12694/scpe.v25i3.2628

MINIMIZING OVERHEAD THROUGH BLOCKCHAIN FOR ESTABLISHING A SECURE SMART CITY WITH IOT MODEL

ZHIXIONG XIAO*

Abstract. Conventional safety measures are inconsistent with inexpensive technologies like the Internet of Things (IoT) due to their significant storage traces, which are prohibitive to their utilization. The blockchain (BC) framework maintains the five essential security primitives: genuineness, credibility, secrecy, accessibility, and non-renunciation. Most IoT gadgets have limited resources, so a traditional blockchain deployment is inappropriate. Traditional deployment of blockchain computing in the Internet of Things leads to significant power consumption, delay, and computational inefficiency. The proposed solution improves the blockchain's conception to serve IoT technologies better. This article proposes a blockchain-based intelligent city design for the IoT that keeps all encryption safety precautions in place. Adding blockchain to an IoT platform does not add much extra labour. After comparing all safety requirements to existing literature, it is clear that the proposed method achieves satisfactory safety effectiveness.

Key words: Internet of Things, Blockchain Computing, Smart City, Encryption, Cyber Security.

1. Introduction.

The Internet of Things (IoT) might provide high-quality, low-overhead, and human-free answers to many problems in many fields. Developing ”smart communities,” which integrate various IoT-enabled activities such as intelligent conveyance, intelligent garbage administration, smart accommodation, and smart water, is an essential use of the technology. Such a wide range of offerings gives developers of smart city collaboration apps much flexibility [1].

The concept of a ”smart community” within a ”smart city” refers to creating information technology to provide comprehensive regional collaboration services based on electronic records and technology, with the ultimate goal of enhancing the standard of life for city citizens. When multiple companies must work together to get a job done, keeping private data safe and secure cannot be easy [2]. Information security and privacy must be prioritized, and citizens and policymakers require reliable data access. The online security framework for software-defined networks (SDN) and smart contract-enabled governmental smart towns relies on authorization and validation for usage in limited environments. This layout was developed for times when funds are tight. The proposed security framework for shared service delivery is now being piloted on the distributed ledger Blockchain networks [3].

The shared task of designing a smart city, a fresh take, is provided for using intelligent agreements in numerous blockchains to protect sensitive information at every stage [4]. The safety measure uses the adaptable nature of intelligent contracts for the confidentiality and integrity of all transactions and interactions between diverse IoT networks. The authors built and ran a use case involving collaborative services inside an SDN-enabled Internet of Things framework to test the viability of the proposed service safety framework. As the global population increases and the idea of smart cities becomes a reality, developing novel approaches to environmental tracking and management, citizen well-being, and government effectiveness will become more crucial. This research's proposed design aids the communication framework of disparate Internet of Things (IoT) networks, letting them link up and cooperate on various tasks. The strategy suggested by the authors will utilize novel safety techniques [5].

The recommended structure for a smart city explains how the IoT gadgets on various networks should register with one another, exchange data, and carry out adaptive application security measures. The researchers found that the proposed method scales well, even when the number of queries made throughout the length of

* Urban Construction College, Fuzhou Technology and Business University, Fuzhou, Fujian, 350715, China (Corresponding author's e-mail: zhixiongxiao7@126.com)

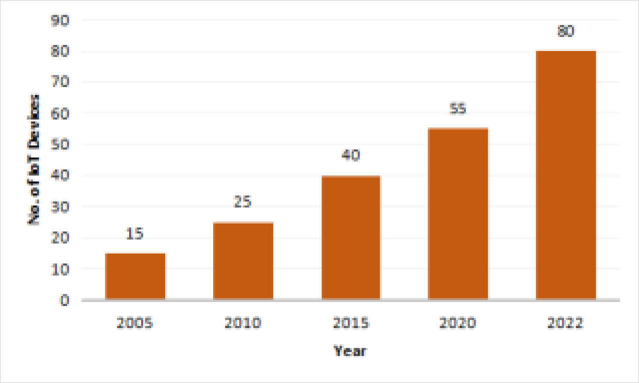


Fig. 1.1: Survey report for the blockchain-based IoT in smart city

an interaction between two separate IoT networks during work together. A high degree of authorization, compatibility, and the transfer of health data are all made possible through the application of blockchain-based technologies for drug supply management (DSM). Academics have shown a surprisingly high interest in IoT-based urban planning during the past several years [6,7].

By providing access to various high-tech services, ”smart cities” (SC) aim to raise the living standards of their residents. Smart cities, the Fourth Industrial Revolution, and innovative banking are just a few examples of these uses. SC may offer a higher level of security by implementing blockchain technology (BCT). For this goal, events are recorded in an immutable, encrypted, autonomous online database open to public inspection. The study's overarching goal is to thoroughly explore the present state of research using state-of-the-art technologies like BCT and the IoT in DSM and SC. Figure 1 shows the blockchain in the IoT model under a secure smart city.

2. Literature Review. The development and layout of DSM and SC programs that use BCT and IoT are focused on the first group. The second category includes a wide range of research into BCT and IoT applications in DSM and SC settings. The third type of contribution is reviewing papers on incorporating BCT and IoT into DSM and SC-based systems. The author provides an overview of the many benefits of employing BCT and IoT in DSM and SC, as well as suggestions for overcoming some of the challenges that have been identified. The new work adds to the corpus of information by analysing all potential avenues in-depth and pinpointing gaps in the understanding. The relevance of BCT and its execution are thoroughly discussed, giving academics a thrilling opportunity to develop more decentralized DSM and SC applications and resulting from extensive dialogue on the usefulness of BCT and the steps needed to put it into practice [8].

This research analyses the chosen literature to determine how BCT is used for IoT and how it enhances the organization of data processes. This study is an in-depth examination and classification of blockchain technology with the IoT and other SC and DSM applications. This analysis and classification reveal many recurring themes in the literature on the topic. BCT can handle several types of big data, use secure information in both digital and physical environments, and not depend on a single point of failure, all contributing to its rising appeal as a solution for handling data. Furthermore, it can decode encrypted data even when disconnected from the internet. Several studies have looked into different aspects of data management to see if these objectives have been met. These aspects include data collection, processing, security, distribution, retrieval, and storage [9].

The BCT-based solutions raise the standard in IoT and secure communication areas. Enhanced identification features like data collecting may allow BCT-based computer systems to produce more. One application of this idea is the usage of public keys inside encryption methods. The improved authentication capabilities provided by BCT-based systems also facilitate other data management activities, such as data collection [10].

A safer cryptographic method provides the features of data collection. Various authentication procedures, from biological processes to public-key encryption techniques, have allegedly been used. Fingerprinting is one method in this category. Similarly, a growth in the popularity of electronic contracts is an advantage brought about by using BCT in data processing. Although smart contracts were around long before BCT became popular, they were rarely employed as a data processing tool until BCT became mainstream. However, blockchain technology is used in various ways by platforms built on it, including handling data automatically. Transmission and recovery professionals in the field of data management should be noticed. Despite this, several writers provide a clear and concise account of the distribution strategies they employed in their deployments. However, the authors found various factors significantly impact data storage. The designers find ways around these restrictions by establishing a data lake, registering a file catalogue, and storing only file locations. Developers constantly adjust their code to adhere to regulations [11, 12].

BCT and IoT devices have immense potential to revolutionize the healthcare industry, smart cities, and other sectors like agriculture, transportation, and manufacturing. Because the Internet of Things (IoT) uses such a wide variety of recent technological advancements, it is not feasible to construct a single suggested architecture that could be used as a master plan to accommodate all potential requirements. There are specific potential applications of the IoT that have yet to be investigated or need sufficient knowledge on how to approach them [13]. It demonstrates the need for further study in this challenging field to discover new and possibly significant societal advantages. Although smart cities give inhabitants and suppliers of capital a variety of benefits, there are many ways in which breaches might endanger people's health and safety. As a direct consequence of this, the IoT may accommodate several distinct suggestion schemes at the same time. This study investigates the relationship between technology and morality related to the safety of IoT-enabled technologies in modern urban construction. Therefore, it offers a secure IoT network architecture for smart cities that combines blockchain technology and deep intelligence to protect users' privacy and trustworthiness [14].

The structure was developed by combining blockchain technology with advanced intelligence. This system uses the blockchain network for risk assessment and mitigation in the context of intelligent city facilities. A neural network model and an optimization approach are both included in this structure. The optimization algorithm ensures that the smart city infrastructure optimizes its resources. In this study, a secured smart city infrastructure employing a blockchain and deep intelligence architecture is built. This infrastructure aims to guarantee that IoT connectivity in smart cities is trustworthy and protects users' privacy. According to the prior discussion, sophisticated deep learning powered by blockchain mechanisms might be merged to handle computational intelligence and security challenges on the IoT-enabled intelligent urban infrastructure. The operational insights made possible by fog and edge cloud apps increase the ability to transform massive amounts of data that are either stationary or in motion into activities that begin immediately. A neural network model and an optimization algorithm are both included in this structure. The optimization algorithm ensures that the smart city infrastructure optimizes its resources [15].

To succeed smart city new solutions will be required in the following six areas: ecological living and health, energy, safety and security, finance, government and schooling, and transportation. Many recent technological advances may be traced back to the exponential growth of the IoT over the past few decades, including the notion of the smart city. To improve the quality of life in healthcare, trade, farming, and conveyance, a ”smart city” is built by integrating IoT devices with technological advances in communication and information. It is crucial to build these technologies safely to prevent attackers from penetrating the existing systems, but many new privacy hazards and challenges have emerged due to this advancement [16].

Blockchain, a relatively recent innovation built on cryptographic rules, may play a crucial role in the safety of future smart cities. This research covered a wide range of blockchain applications for smart cities. The authors examined whether and how blockchain technology's openness, republic, restructuring, and safety advantages may improve smart city services. This research will allow the implementation of an intelligent contract voting system based on the Ethereum blockchain, revolutionizing electronic voting use. The authors focused on the problem of inadequate security precautions in smart cities and offered many options for improving safety based on the research. The studies have focused primarily on blockchain technology and its potential to enhance the safety and privacy of smart city services. Based on the blockchain platform, researchers' solution will facilitate the development of a trustworthy and distributed digital voting mechanism. The authors suggested a voting system that uses technology as the principal service to facilitate voting in smart towns. Voting in smart cities may be simplified with the help of our technology. After all, a distributed digital voting system may have its flaws owing to the reality that it is still an infant technology. Therefore, additional study and investigation of the technology are needed. Sybil's attack is one of the threats to a digital voting system because it uses a vulnerability that may enable a voter to create many identities on a blockchain network [17].

The Ethereum blockchain houses intelligent contracts and user funds in a wallet. On the digital currency Ethereum, accounts manage user authentication by generating encoding content, which forms the backbone of the existing construction for voting online. While the user's private key remains secure, any other peer on the internet may read the public key [18, 19]. Smart contracts are used to automate core aspects of the voting process, including voter verification and tallying. Once consensus is reached across nodes, payments are checked for accuracy before being added to a new block on the distributed ledger. The computerized voting procedure is speculated to be powered by the Ethereum blockchain. Adding a block to the blockchain causes irreversible changes to the blockchain. The update is also sent to all nodes through broadcasting. Furthermore, it guarantees that voters are legitimate, that contract events are widely broadcast and distributed, and that all network participants may access these exchanges, but only one person can unlock them [20].

Multiple safeguards are in place to secure users' private data stored in the public cloud, which is essential for the growth of smart cities. Social manipulation and hacking are two forms of deception that criminals may employ to get access to private user data. These methods may be exploited to steal users' credentials and financial data. Phishing is still the initial stage of a multiple-phase assault, although its technological sophistication has dramatically increased over the past few years. Deception kits have evolved into tools for attack that have become more intuitive, readily available, and simple to deploy over time [21].

Utilizing non-Latin symbols in the URL, typo-squatting of eminent domains, using protected symbols in redirections, and multiple chains scamming indicate a successful scamming campaign. When files containing phishing URLs are uploaded to cloud storage, hackers are offered a helping hand and a push in the right direction. Criminals' use of cloud servers for these kinds of assaults is becoming more common. Current spoofing URL blocking software does not provide enough defence against multilayered phishing. Instead, it is up to the user to take precautions, making them ultimately responsible for their safety. The indestructibility of blockchain data and the impact of avalanches further demonstrate the efficacy of these protections as prerequisite measures to implement. Altering in a method supported by blockchain technology is the most effective solution to safeguard users' cloud-based data [22].

Certain restrictions are embedded in phishing, and the time it takes to mine a block on the Ethereum network has increased due to the standard complexity level. If the CSP has access to a privately configured blockchain with the Phish Block algorithm, then the CSP may change the protocol to speed up the blockchain's block generation process. Incorporating Phish Block as a product would increase security for cloud data and users and provide value as a trust component to the cloud provider's service level contract [23].

3. Materials and Methods. The entire system may be classified into three distinct components: programmable blocks, canopies system, and cloud computing [24]. The components are explained as follows:

A. Programmable Blocks. "Smart constituents" are commonly referring to these smaller divisions. Each smart building block has various detectors, including a sensor for imaging, temperature, LDR, etc. These sensor-equipped gadgets belong to a single block administrator and may only be accessed by that individual. The many bits of data gadgets are kept on an encrypted blockchain managed by the blocking administrator. In contrast to how Bitcoin's database is managed—by a decentralized network of nodes—the local BC is managed by a central authority. The block operator will connect all activities made with or via the devices.

The block administrator is responsible for updating the ledger with new gadgets or removing existing ones. Adding device operations will operate similarly to Bitcoin's' make coin' operation. The local BC has an authority element that allows the block administrator to control every exchange throughout the local blockchain. Only with the approval of the block's operator will electronic devices be allowed to communicate and share any necessary data. Authorization for the operations, it may be possible to utilize the Diffie-Hellman algorithms key transfer mechanism to enable the collaboration of a shared key. While the entry header for each block in the blockchain is recorded in that block's header, only the latest header is used to verify transactions. The suggested encrypted blockchain does not use evidence of work or any other challenges to reduce the associated expenses.

After appending a reference to the initial transaction and duplicating the policy from the preceding block's header, the user attaches the entire block to the distributed ledger. When an activity is included in a block using the algorithm underpinning Bitcoin, it is regarded as legitimate, irrespective of the fact that the block has been processed. Keep in mind that a personal blockchain may be set up to manage not just user authorization but also collaboration between gadgets, in addition to generating and recording securely in an immutable ledger, both operation data and scenario-based IoT agreements. Every intelligent building block will have its digital storage and a set of public credentials to provide user entry to the information stored in neighbouring units.

B. Canopy Network. Lantern is a network operated by peers made up of intelligent buildings and other individuals, including law enforcement agencies and government public administration entities. Every group of the canopies network's nodes elects a Group Head (GH) by a majority vote of the cluster's participants. Each GH must keep a public blockchain operational. The GHs remember the public key pairs of the consumers and use this information to decide whether or not the requester is authorized to see the information stored in the associated intelligent blocks. The public keys of requestors of bright bricks that belong to this set and may be retrieved are likewise managed and stored by the GHs.

C. Cloud Computing. The cloud is an official associate of the Group Head. The smart block's constituent gadgets could sometimes choose to back up their information to the cloud. This information must be shared with an outside organization to make these features available on mobile platforms. Suppose a limited number of additional state or centralized institutions opt to access the information in creative blocks. In that case, they will have read-and-write access to the information preserved in the data centre.

All interactions between nearby gadgets and canopy nodes are identified with activities. The proposed platform can accommodate five distinct types of trades. Suppose the intelligent gadgets inside the block want to save their data somewhere other than in memory, such as the block administrator's file system or the cloud. In that case, this will be considered as a Write operation. A read transaction will be started if the block manager, several states, or a central institution decides to access information stored in the cloud. A monitor transactional will be generated if additional states, centralized businesses, or block managers want to request data from monitor devices directly. An inception operation is utilized to add a new device to the intelligent block, and an elimination activity is performed to remove a device that already exists from the smart block. Every transaction entering or leaving the intelligent block will be documented on the distributed local ledger to the intelligent block. Data security concerns will be addressed by using an inexpensive hashing technique.

4. Proposed Methodology. The process for this inquiry may be broken down into four main parts. The low-processing platform's requirements are defined during the setup phase. Higher productivity (HP) nodes have their underlying files, including block admin and group head, set before launch. It shows how actual data is sent across the local and canopy networks throughout the exchange of information. A smart city's whole infrastructure, comprising intelligent interference, overhead system, and online storage, is shown in Figure 4.1.

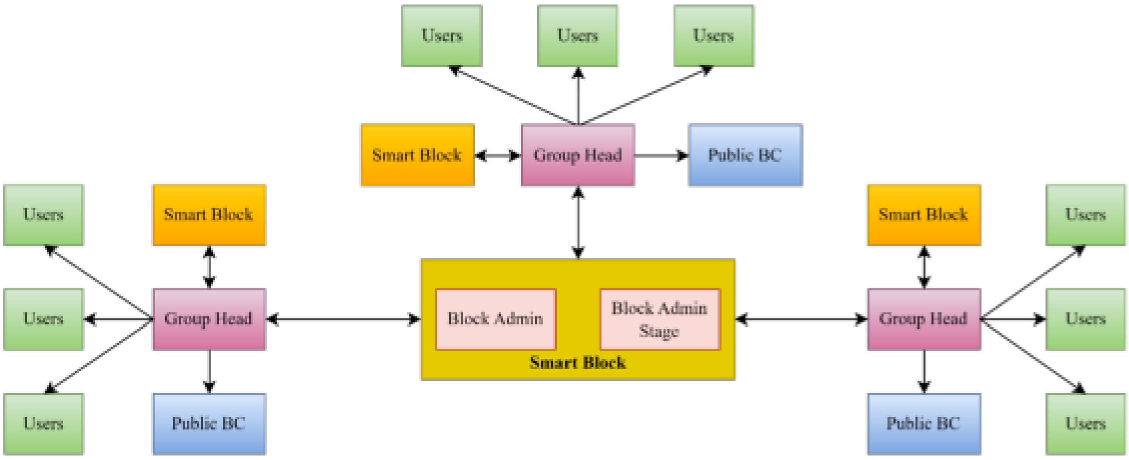


Fig. 4.1: Proposed Intelligent Smart City Infrastructure

a) Initialization. Low-processing (LP) area and high-processing (HP) block administrators and canopies networks are catered to design by two distinct startup methods. Both HP and LP have limited computational capabilities.

In step one, Low Processing (LP) Initialization has 'm' total connections. An unlimited number of 'n' sensors may be connected to each system component. Each node has its unique device ID and three keys. Three keys are required to encrypt a symmetrical key: the public, private, and actual. All nodes and block administrators get the encrypted key using an efficient critical transfer procedure. The unique identification of the gadget is the hash of its standard key. The Bitcoin account ID may also be calculated using the public key's hash value. The uniqueness of a random integer's hash might be verified using the hash attribute.

In step two, high-performance (HP) initialization, low-performance nodes use an inexpensive cryptographic approach to protect the data they collect from the sensors before sending it to the high-performance (HP) end nodes. High-level nodes that process messages will receive these packets and check their sender equipment ID, sensors ID, list of public keys, authorization header, transactional type, and hash information before continuing the procedure. All high-end processing nodes must agree that the transaction is valid before authorization, and their copy of the transactions is added to the blockchain. The organizational head in the overhead networks and the blocking administrator in the local network are examples of the high-processing units that we have encompassed into our architecture. The local blockchain maintained by the block supervisor must keep a copy of all transactions. Under certain conditions, the deal may be published on the public blockchain managed by the group's leader. The group's leader and the block administration are privy to the node's open keys and the header controlling who may access them. Validation of the access management header, receiver devices ID, sensor ID, list of publicly accessible keys, and hash information will precede every interaction from the minor processing nodes to the district admin and the block administrator to the group person.

b) Transactions. Introducing new gadgets, deleting old ones, and transferring information comprise the three activities that may be performed in the proposed design. Like the genesis operation of BC, the add device operation will add the new gadget's public key to the list of publicly accessible keys the block manager keeps. The public key will be deleted from the essential public list maintained by the blocking administrator as part of the deactivated device operation. It will keep the chain alive and well in the blockchain itself. There will be two distinct types of data transmission operations. The first tier of the structure consists of minimally processed nodes, which may transmit sensor data to the block administrators and receive data from the block admin.

Assume that video gauges are relaying, from the low analyzing node to the block manager, the number of automobiles that have passed through a given lane and that the block administrators utilize the resulting data to continually calculate the length of time that the green light must stay on for vehicles passing through that

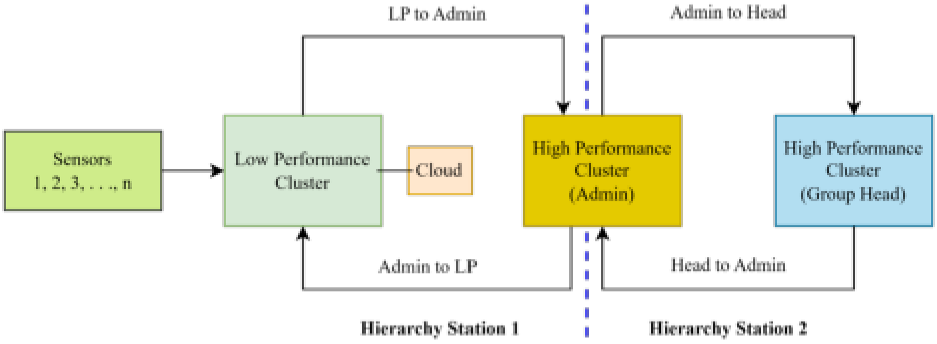


Fig. 4.2: Transactions at various Hierarchical levels

lane. The block administration can send information to the organization's group head, and the group head can send information back to the block manager at the subsequent level of the structure. Let us pretend that the state or federal government has to keep an eye on data gathered from a camera put in some intelligent block through the group head. Figure 3 depicts the two operations that involve the organizational structure.

c) Packets. Every message in this architecture travels only between the LP and HP nodes. The data within the frame is constantly generated by the LP and is used either in the HP memory or at the LP's production. Three parts make up the package that is made in LP.

1. The data collected by the LP's detectors is protected using the symmetrical key algorithm for encryption. The blockhead HP is given access to the encrypted symmetrical key using the Diffie-Hellman method.
2. The LP generates a hash of every sensor reading using the lightweight method to ensure the packet is uncorrupted.
3. An electronic signature is generated when the LP encrypts the encrypted information using a personal key and a public-key encryption algorithm.
4. Accessibility Control Preamble: The access management header stores the different storage types' read/write authorization. Each LP has its specialized sensor, and the results of these readings are often saved in either the block administrator's local database or the principal network's cloud database. Each output in an LP may access the block administrator's storage for retrieving information. According to the use case, numerous sensors can need access to different data or systems.
5. Extra Information: Time mark, biosensor ID, and gadget ID (LP ID) are also included in the previously mentioned package.

d) Process Flow. The following is the sequence in which components of the regional network's design are activated:

- All private and public keys are precomputed and maintained in LPs; connected block managers get LPs' public keys using the Diffie-Hellman technique. Due to its limited computational power, the architecture never generates private and public keys at the LP level.
- The standard symmetrical key is sent from the block administrator to the LPs via the Diffie-Hellman method.
- After various LP nodes collect information from different devices, it is scrambled and encoded using symmetrical keys before being put to use. The scrambled information is re-encrypted with the corresponding private key to create an electronic signature.
- The data is encrypted with a symmetrical key, the hash value is generated using a mapping technique, and the private key is encrypted with an asymmetrical essential cryptographic procedure, all inside the context of the LP technique. The resultant data packet is secure, and the wireless network device relays this signal to the building manager.
- All of the data streams from the different LPs are received by a PHP-based applications periphery

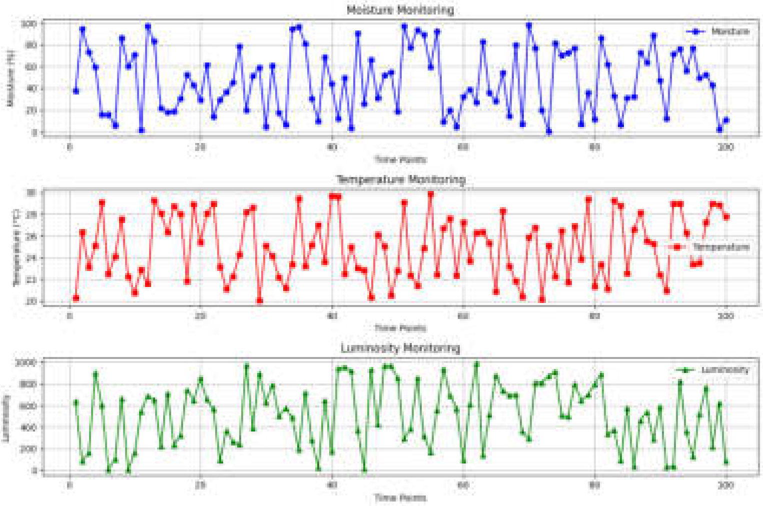


Fig. 5.1: Moisture, temperature and luminosity monitoring at various time points

interface (API). To verify the electronic authorization, the preceding API collects the device or LP ID and receives the corresponding public key from the LP. If the electronic identity checks out, the packet comes from the expected device or LP ID. The shipment will be returned to the sender if its signature does not match. The hash value will likewise be checked to ensure the packet's authenticity.

After verifying the algorithm's hash and authorization, the raw data can be accessed using the standard key and recorded in the relevant record of the block administrator's blockchain. To organize head operations, the block administration must follow the same steps.

5. Experimentation and Results. The single-chip Node MCU acts as the LP in our structural concept and is connected to monitors for measuring moisture, temperature, and luminosity. Figure 5.1 demonstrates the moisture, temperature, and luminosity monitoring at various times.

The DHT11 sensor is a combination thermometer and hygrometer, the level of light detection and the Light Detection Resistant (LDR) tool for taking measurements. The block administrator uses a standard personal computer. The LP public essential list, authorization header, and local network are all stored in a MySQL store. The API is written in PHP to get all the information from the sent packets. The Ethereum infrastructure activates the shade connection, and Figure 5.2 defines the data train Vs test.

Table 5.1 ensures that the five principles of cryptography are satisfied by the appropriate measures. The five principles are secrecy, accessibility, reliability, verifiability, and nonrepudiation. An analysis of the suggested layout of the existing Bitcoin structure and a recent application as a standard is provided in Table 5.2.

The most fundamental issue with constructing a distributed ledger is the Merging Overhead, discussed in considerable detail in the third row of Table 2. All previous activities are part of the Bitcoin blockchain, and new mining is needed to download the entire chain in its present form and the article. The proposed solution adds the public identity of each newly joined user to a separate private ledger and retrieves every previous transaction recorded in its blockchain. This blockchain stores the public keys of all permitted users, including the block admin's public key.

In this blockchain, all changes, including additions and deletions of devices, are continuously recorded by cryptography. The canopy network's leader also adopted this strategy across its system. The simulated result of blockchain technology for IoT with a mini batch size of data processing in percentage is shown in Figure 5.3.

When public keys are stored using blockchain systems, malicious devices cannot access networks or communicate with block controllers to obtain passwords. Due to the absence of the malicious device's public key in the public key blockchain, it will be unable to complete any packet transactions.

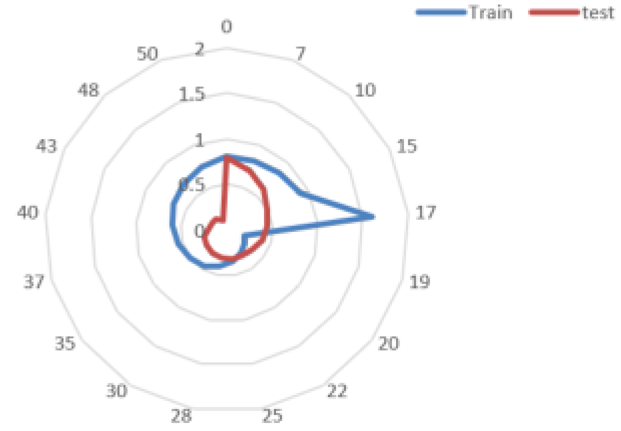


Fig. 5.2: Data Train Vs Test

Table 5.1: Principles of Cryptography

Safety Problems	Recommended Response
Secrecy	Every exchange is encrypted with a symmetric key.
Reliability	It is the hash of all transactions.
Accessibility	When a legitimate user requests an internet service, the regional and remote networking permissions management header processes the inquiry.
Validation	Using a ”connection statement” and ”displayed keys” does this.
Nonrepudiation	The agreement creator signs all regional and global operations to ensure no repudiation. Furthermore, as a result, no party can dispute their role in an arrangement.

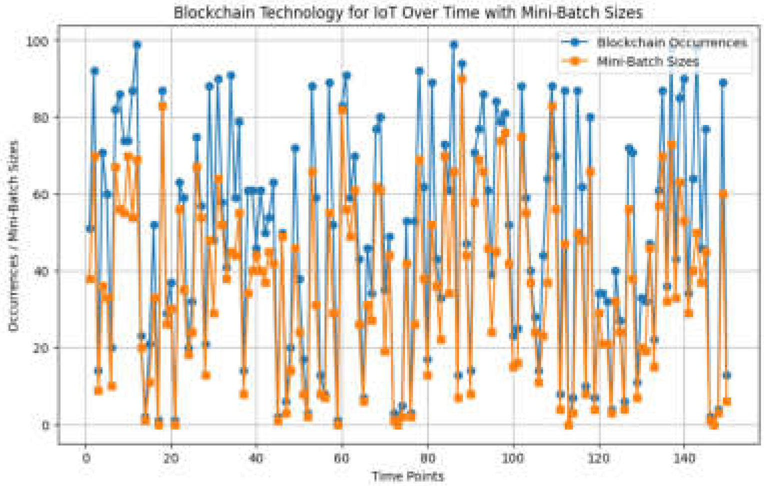


Fig. 5.3: Mini batch size of data in percentage for IoT over time

Table 5.2: Examination of New and Established Blockchains

Factor	BitCoin	Local BC	Public BC	Proposed Local BC	Proposed Public BC
Removal	PoW	Nothing	Nothing	Nothing	Nothing
BC capacity	Public	Private	Private	Private	Private
Client combination in the clouds	Download all blocks on the PC	Download all blocks on the PC	Download all blocks on the PC	Download all blocks in PC and public key	Download all blocks in PC and public key
BC manage	Nothing	Owner	Nothing	Owner	Nothing
Double spend	Not possible	Not applicable	Not applicable	Not applicable	Not applicable
Operation Type	Broadcast	Unicast	Unicast/ Multicast	Unicast	Unicast/ Multicast
Operation Parameter	Input, Coin output	Block-no, Hash data, PK time, Output, Policy rules	Output, PKs	Block-no, Hash data, PK time, Output, Policy rules	Output, PKs
Block Description	Hash puzzle	Policies	Policies	Access header	Access header
Encryption procedure	Public key cryptography	Not studied	Public key, Symmetric key	Public key, Symmetric key	Public key, Symmetric key
Forking	Not Permitted	Permitted	Permitted	Permitted	Permitted
54% attack	Double spending	Not possible	Not possible	Not possible	Not possible
Remuneration	Coins	Nothing	Not defined	Nothing	Nothing
Pool removal	Permitted	Can’t be defined	Can’t be defined	Can’t be defined	Can’t be defined
Malicious User	Permitted	Possible	Not possible	Not possible	Permitted
Miner division	Self- -selection	Owner choice	Node in group choice	Owner choice	Node in group choice

6. Conclusion. Conventional safety measures should be avoided wherever possible due to the enormous temporal and spatial requirements of IoT applications. The traditional structure of the blockchain system is altered in this study so that it may be used for IoT applications. The structure that is being suggested maintains anonymity while also ensuring genuineness, accessibility, honesty, and acceptance. This blockchain relies on the IoT program, and it can prevent various common assaults, including the denial access threat, the 53 % threat, the alteration threat, the computational threat, the person in the centre attack, the throwing threat, and others. The analysis of the suggested layout is better than the previously published material, and demonstrates that it satisfies and exceeds the expectations of several significant concerns.

REFERENCES

[1] M. E. Pamukov and V. K. Poulkov, ”Multiple negative selection algorithm: improving detection error rates in IoT intrusion detection systems,” in Proceedings of the 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), pp. 543–547, Bucharest, Romania, September 2017.

[2] A. Islam, A. Al Amin, and S. Y. Shin, ”FBI: a federated learning-based blockchain-embedded data accumulation scheme using drones for internet of things,” IEEE Wireless Communications Letters, vol. 11, no. 5, pp. 972–976, 2022.

[3] A. Dorri, S. S. Kanhere, R. Jurdak, and P. Gauravaram. Blockchain for IoT security and privacy: The case study of a smart home. In 2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops), pages 618–623, March 2017.

[4] A. Islam, T. Rahim, M. Masuduzzaman, and S. Y. Shin, ”A blockchain-based artificial intelligence-empowered contagious pandemic situation supervision scheme using internet of drone things,” IEEE Wireless Communications, vol. 28, no. 4, pp. 166–173, 2021.

[5] Rafiullah Khan, Sarmad Ullah Khan, Rifaqat Zaheer, and Shahid Khan. Future internet: The internet of things architecture,

possible applications and key challenges. In 2012 10th International Conference on Frontiers of Information Technology (FIT): Proceedings, pages 257– 260. Institute of Electrical and Electronics Engineers Inc., 2012.

[6] Arvind Narayanan, Joseph Bonneau, Edward Felten, Andrew Miller, and Steven Goldfeder. Bitcoin and Cryptocurrency Technologies: A Comprehensive Introduction. Princeton University Press, Princeton, NJ, USA, 2016.

[7] A. Khannous, A. Rghioui, F. Elouaai, and M. Bouhorma, ”MANET security: an intrusion detection system based on the combination of Negative Selection and danger theory concepts,” in Proceedings of the 2014 International Conference on Next Generation Networks and Services (NGNS), pp. 88–91, Casablanca, Morocco, May 2014.

[8] A. Mosenia and N. K. Jha. A comprehensive study of security of internet-of-things. IEEE Transactions on Emerging Topics in Computing, 5(4):586–602, Oct 2017.

[9] P. Widulinski and K. Wawryn, ”A human immunity inspired intrusion detection system to Search for infections in an operating system,” in Proceedings of the 2020 27th International Conference on Mixed Design of Integrated Circuits and System (MIXDES), pp. 187–191, Lodz, Poland, June 2020.

[10] Manik Lal Das. Privacy and security challenges in internet of things. In Raja Natarajan, Gautam Barua, and Manas Ranjan Patra, editors, Distributed Computing and Internet Technology, pages 33–48, Cham, 2015. Springer International Publishing.

[11] A. Borkar, A. Donode, and A. Kumari, ”A survey on intrusion detection system (IDS) and internal intrusion detection and protection system (IIDPS),” in Proceedings of the 2017 International Conference on Inventive Computing and Informatics (ICICI), pp. 949–953, Coimbatore, India, November 2017.

[12] M. Kumar and A. K. Singh, ”Distributed intrusion detection system using blockchain and cloud computing infrastructure,” in Proceedings of the 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184), pp. 248–252, Tirunelveli, India, June 2020.

[13] S. Ouiazzane, M. Addou, and F. Barramou, ”Toward a network intrusion detection system for geographic data,” in Proceedings of the 2020 IEEE International conference of Moroccan Geomatics (Morgeo), pp. 1–7, Casablanca, Morocco, May 2020.

[14] Alessandra Rizzardi, Luigi Alfredo Grieco, and Alberto Coen-Porisini. Security, privacy and trust in internet of things: The road ahead. Computer Networks, 76:146– 164, 2015.

[15] J. Yu, P. Tian, H. Feng, and Y. Xiao, ”Research and design of subway BAS intrusion detection expert system,” in Proceedings of the 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), pp. 152–156, Chongqing, China, October 2018.

[16] X. Zhan, H. Yuan, and X. Wang, ”Research on block chain network intrusion detection system,” in Proceedings of the 2019 International Conference on Computer Network, Electronic and Automation (ICCNEA), pp. 191–196, Xi’an, China, September 2019.

[17] L. Hong, ”Immune mechanism-based intrusion detection systems,” in Proceedings of the 2009 International Conference on Networks Security, Wireless Communications and Trusted Computing, pp. 568–571, Wuhan, China, April 2009.

[18] E. D. Alalade, ”Intrusion detection system in smart home network using artificial immune system and extreme learning machine hybrid approach,” in Proceedings of the 2020 IEEE 6th World Forum on Internet of Bings (WF-IoT), pp. 1-2, New Orleans, LA, USA, June 2020.

[19] Y. Shen, Y. Fei, L. F. Zhang, A. Ji-yao, and M. L. Zhu, ”An intrusion detection system based on system call,” in Proceedings of the 2005 1st IEEE and IFIP International Conference in Central Asia on Internet, p. 4, Bishkek, September 2005.

[20] K. A. da Costa, J. P. Papa, C. O. Lisboa, R. Munoz, and V. H. C. de Albuquerque, ”Internet of things: a survey on machine learning-based intrusion detection approaches,” Computer Networks, vol. 151, pp. 147–157, 2019.

[21] E. M. Campos, P. F. Saura, A. Gonz’alez-Vidal et al., ”Evaluating federated learning for intrusion detection in internet of things: review and challenges,” Computer Networks, vol. 203, Article ID 108661, 2022.

[22] A. Mihoub, O. B. Fredj, O. Cheikhrouhou, A. Derhab, and M. Krichen, ”Denial of service attack detection and mitigation for internet of things using looking back-enabled machine learning techniques,” Computers and Electrical Engineering, vol. 98, Article ID 107716, 2022.

[23] Z. S. Malek, B. Trivedi, and A. Shah, ”User behavior pattern -signature based intrusion detection,” in Proceedings of the 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4), pp. 549–552, London, UK, July 2020.

[24] G. Zhu, H. Yuan, Y. Zhuang, Y. Guo, X. Zhang, and S. Qiu, ”Research on network intrusion detection method of power system based on random forest algorithm,” in Proceedings of the 2021 13th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), pp. 374– 379, Beihai, China, January 2021.

Edited by: Venkatesan C
Special issue on: Next Generation Pervasive Reconfigurable Computing for High Performance Real Time Applications
Received: Sep 26, 2023
Accepted: Feb 8, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1332–1340, DOI 10.12694/scpe.v25i3.2639

TARGET IMAGE PROCESSING BASED ON SUPER-RESOLUTION RECONSTRUCTION AND MACHINE LEARNING ALGORITHM

CHUNMAO LIU*

Abstract. This article proposes a target image processing method based on super-resolution reconstruction and machine learning algorithms, which solves the low-resolution problem in medical images during imaging. This method uses nonlocal autoregressive learning based on a medical image super-resolution reconstruction method. The autoregressive model is introduced into the sparse representation-based medical image super-resolution reconstruction model by utilizing medical image data inherent nonlocal similarity characteristics. At the same time, a clustering algorithm is used to obtain a classification dictionary, improving experimental efficiency. The experimental results show that ten randomly selected CT/MR images are used as test images, and each image's peak signal-to-noise ratio and structural similarity values are calculated separately. Compared with other methods, the method proposed in this paper is significantly better and can achieve ideal results, with the highest value being 31.49. This method demonstrates the feasibility of using super-resolution reconstruction and machine learning algorithms in medical image resolution.

Key words: Medical image, Super-resolution reconstruction, Nonlocal autoregression, Classification dictionary

1. Introduction. Many areas of life have high requirements for image resolution, such as security, medical, aerospace, criminal investigation, etc., and image super-resolution reconstruction technology has gradually become a hot topic in recent years. Image super-resolution methods mainly include difference-based, reconstruction-based, and learning-based methods. The difference-based method utilizes the pixel values around a certain pixel point and their relative positional relationships to evaluate mathematical methods, mainly including nearest neighbour interpolation and bicubic interpolation [1]. Reconstruction-based methods mainly use the prior knowledge of natural expectations like smoothness. Common methods include convex set projection and the maximum posteriori probability method. Learning-based methods mainly include traditional machine and deep learning methods, which have been widely used in recent years. Machine learning-based methods generally combine matrix transformation-related methods to learn complex mapping relationships, requiring manual feature extraction for learning. Medical image refers to the use of medical devices to obtain images of internal organs or tissues of a part of the human body without invading the human body to diagnose the patient's condition or conduct medical research [2]. Deep learning is a branch of machine learning and an important means to realize artificial intelligence technology. With the widespread application of deep learning technology in image processing and computer vision, deep learning technology's auxiliary and decision-making role in clinical diagnosis has become a research hotspot in medical image analysis. Intelligent diagnosis of medical images can be roughly divided into three steps. First, obtain a large number of high-quality image data, then preprocess the image, and finally mine the image information for analysis and prediction, as shown in Figure 1.1 [3].

Among them, massive and high-quality image data is the basis of deep learning and training. Image preprocessing, such as registration and region of interest extraction, is the basic guarantee for the accuracy of subsequent analysis. Mining information and establishing prediction models are the keys to clinically intelligent decision-making [4].

2. Literature Review. Super-resolution using a convolutional neural network (SRCNN) is the pioneer of deep learning applications in the field of super-resolution and has made a breakthrough in the effect and speed of image super-resolution reconstruction [5]. Many improvements based on SRCNN, which greatly promoted image

* School of Electronics and Information Engineering, Henan Polytechnic Institute, Nan Yang, Henan, China, 473000 (Corresponding author's e-mail: chunmaoliu3@163.com)

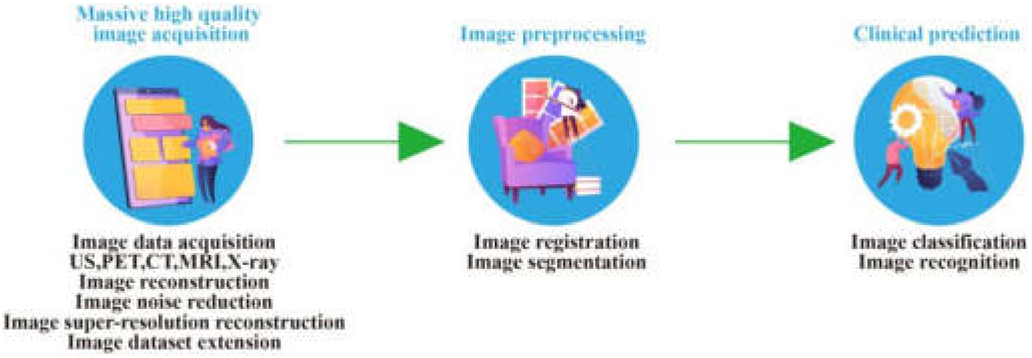


Fig. 1.1: General steps of medical image processing and analysis

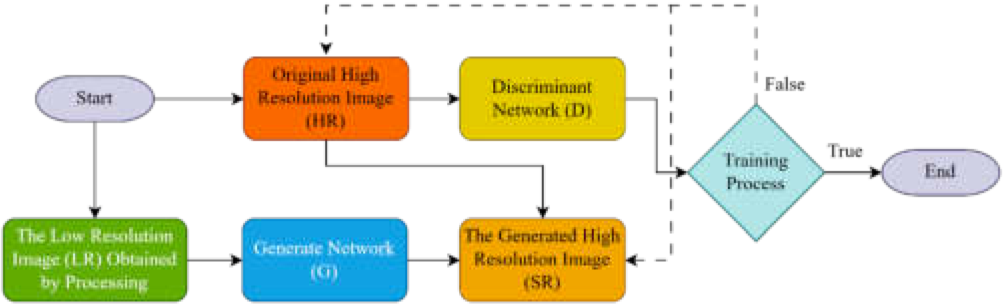

Start
Original High Resolution Image (HR)
Discriminant Network (D)
False
Training Process
True
End
The Low Resolution Image (LR) Obtained by Processing
Generate Network (G)
The Generated High Resolution Image (SR)

Fig. 2.1: SRGAN training process

super-resolution technology development, are investigated in [6]. After that, the generated countermeasure network is analyzed for image super-resolution reconstruction and the SRGAN algorithm, which makes the image have more high-frequency details and feel better after reconstruction. Still, the training of the Gan network is unstable, and the training time is longer [7]. The training process of SRGAN is shown in Figure 2.1.

The generation network used in SRGAN is a residual network, a dense connection model. Specifically, each layer will accept the previous layer's output as input and the output of all previous layers. The SR reconstruction technology processes one or more low-resolution images to improve the resolution of the original image, compensate for the lost details of LR images, and reconstruct high-resolution (HR) images [8].

An iterative interpolation algorithm is proposed based on curvature after weighing the effect and efficiency of the algorithm and using GPU to accelerate the real-time interpolation of HR images [9]. The learning-based SR algorithm needs to build a learning library by dictionary learning on many HR images to obtain the learning model from LR images to HR images. The authors were inspired by compressed sensing and randomly selected 100,000 image blocks from the training set as samples for training. They also used a sparse coding algorithm to obtain a compact dictionary with many atomic terms [10]. The authors proposed an image SR algorithm using low rank and total variational regularization, which is applied to the SR reconstruction of MR images [11]. A sample learning-based image SR algorithm divides the feature space into multiple subspaces, uses the collected samples to learn the prior information of each subspace, and generates effective mapping functions [12].

3. Research Methods. This paper uses a sliding window to divide the image into several overlapping sub-image blocks with a certain step size, as described below.

3.1. Image similarity characteristics. After image segmentation, it is usually found that there are many similar structures in these image blocks, and this phenomenon also exists in different images. This universal similarity can be applied to SR reconstruction of images as a priori information to improve the reconstruction quality of images [13]. When realizing SR reconstruction of medical images, HR images can be obtained by weighting similar image blocks. Therefore, to improve the reconstruction quality of medical images, the constraint of nonlocal similarity of image blocks can be added to the observation model of image super-resolution, $y = DHx$. After dividing the image into blocks, take an image block x, and its nonlocal similar image block. x_i^j can be weighted to obtain an image block x, that is

$$x_i \approx \sum_j \omega_i^j x_i^j \tag{3.1}$$

In the experiment, the center of the image block is used to represent the image block x_i, x_i^j is the similar image block of x_i , and ω_i^j is the weighting coefficient of the similar block. When looking for similar image blocks, first use the K-means clustering algorithm to get k clustering centers. When judging that the image blocks x_i^j and x_i are similar, calculate the difference between the current image block x_i^j and the clustering center $\hat{x}$ [14].

$$e = ||x_i - \hat{x}||_2^2 \tag{3.2}$$

The closest image centre from the difference value, and then from this closest cluster is expressed as

$$d_i^j = ||x_i - x_i^j||_2^2 \leqslant \theta \tag{3.3}$$

The first j image blocks most similar to x_i, where θ is the set threshold which is the difference between the j^{th} image block and x_i. When d_i^j is less than the threshold, it is determined that it is a similar image block. Take the first j image blocks most similar to determine ω_i^j:

$$\hat{\omega_i} = arg\,\underset{\omega_i}{min}(||x_i - X\omega_i||_2^2 + \eta||\omega_i||_2^2) \tag{3.4}$$

where x, ω_i, η is the regularization coefficient. The purpose of regularization in Equation 3.4 is to improve the stability of the least squares solution. The conjugate gradient method can solve Equation 3.4 to obtain the solution.

$$\hat{\omega_i} = (X^T X + \gamma I)^{-1} X^T x_i \tag{3.5}$$

We use ω_i to introduce it into $x = Sx + e_x$ to build a nonlocal autoregressive model of image x. e_x is the model error, while

$$S(i,j) = \begin{cases} \omega_i^j, \text{If } x_i^j \text{ is a nonlocal similar block of } x_i, \\ 0, \text{If } x_i^j \text{ is not a nonlocal similar block of } x_i \end{cases} \tag{3.6}$$

A nonlocal autoregressive model is introduced into the SR reconstruction model of sparse images as a new numerical fidelity term to constrain the sparse reconstruction process. Therefore, the SR reconstruction model can be changed to

$$\hat{a} = arg\,\underset{\alpha}{min}\{||y - DS\Psi\alpha||_2^2 + \gamma \cdot R(\alpha)\} \tag{3.7}$$

$$y = D\psi\alpha \tag{3.8}$$

where D is the down sampling of the image, S is the fuzzy kernel function, ψ is the dictionary, α is the sparse coefficient of the image block under the dictionary, γ is used to balance the data regularization term, and the data fidelity term and $R(\alpha)$ is the regularization term. In this way, when the dictionary is trained, dictionary ψ can be obtained. After entering the input LR image, the sparse coefficient α corresponding to image x can be used to reconstruct the HR image using the SR sparse representation model [15,16].

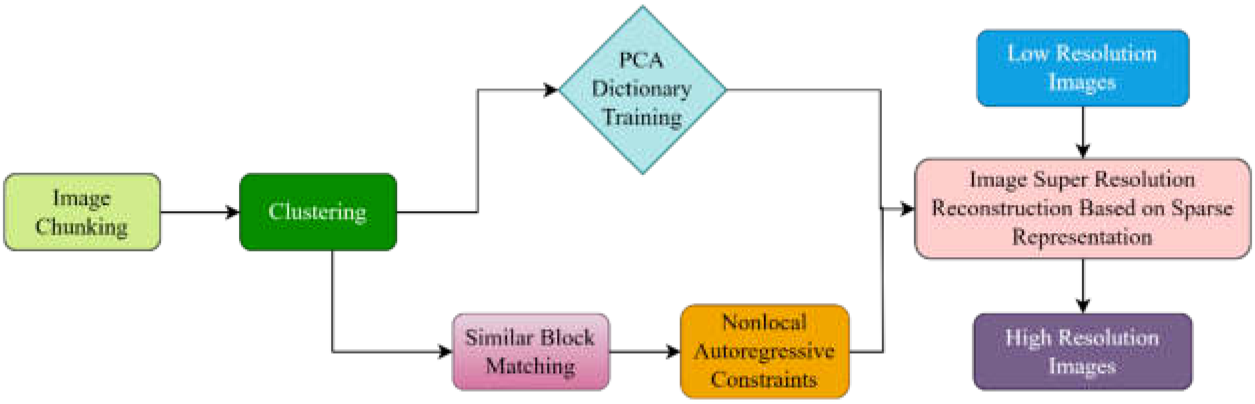


Fig. 3.1: Flow chart super-resolution reconstruction method

3.2. Medical image super-resolution reconstruction method based on nonlocal autoregressive learning. The flow chart of the medical image super-resolution reconstruction method based on nonlocal autoregressive learning is proposed and illustrated in Figure 3.1.

3.2.1. Regularization parameter solution. In the image SR reconstruction model using sparse representation, the selection of the regularization term $R(\alpha)$ has a great impact on the reconstruction effect. Generally, using l1 norm sparse regularization as a constraint, image x is divided into blocks to obtain image blocks x_i, i= 1, 2,..., N. Each image block can be encoded under dictionary Ψ. The SR model based on l_1 norm constraint can be rewritten as

$$\hat{a} = arg\,\min_{\alpha}\{||y - DS\Psi\alpha||_2^2 + \lambda\sum_{i=1}^{N}||\alpha_i||_1\} \tag{3.9}$$

$$s.t. y = D\psi\alpha \tag{3.10}$$

where α_i is the coding vector of the image block x_i, and the sparse coefficient α is composed of α_i. In similar image blocks, image block x_i can be linearly represented by a similar image block x_i^j, and the sparse coefficient is closely related to the image block x_i. For the sparse coefficient α_i^j corresponding to the similar image block x_i^j, it should also be closely related to α_i. Therefore, α_i should also be very close to the weighted average value of α_i^j, that is, $||\alpha_i - \sum_j \omega_i^j \alpha_i^j||_2$ should be very small. α_i^* is used to represent the weighted average value of α_i^j is given by

$$\alpha_i^* = \sum_j \omega_i^j \alpha_i^j \tag{3.11}$$

Similarly, the super-resolution model is expressed as

$$\hat{a} = arg\,\min_{\alpha}\{||y - DS\Psi\alpha||_2^2 + \lambda||\alpha||_1\} \tag{3.12}$$

Add the nonlocal regularization constraint of the above formula to the sparse representation model that is

$$\hat{a} = arg\,\min_{\alpha}\{||y - DS\Psi\alpha||_2^2 + \lambda\sum_{i=1}^{N}||\alpha_i||_1 + \eta\sum_{i=1}^{N}||\alpha_i - \alpha_i^*||_2^2\} \tag{3.13}$$

$$s.t. y = D\psi\alpha \tag{3.14}$$

Since the weighted representation of the normal form can be improved and the sparsity of the formula, the constraint of the normal form is added to the above formula and is given as:

$$\hat{a} = arg\,\underset{\alpha}{min}\{||y - DS\Psi\alpha||_2^2 + \sum_{i=1}^{N}\sum_{j=1}^{r}\lambda_{i,j}|\alpha_{i,j}| + \sum_{i=1}^{N}\sum_{j=1}^{r}\eta_{i,j}(\alpha_{i,j} - \alpha_{i,j}^*)^2\} \tag{3.15}$$

$$s.t.y = D\psi\alpha \tag{3.16}$$

Rewrite the above formula using Equation 3.15

$$\hat{a} = arg\,\underset{\alpha}{min}\{||y - DS\Psi\alpha||_2^2 + \sum_{i=1}^{N}||\lambda_i\alpha_i||_1 + \sum_{i=1}^{N}||\eta_i(\alpha_{i,j} - \alpha_{i,j}^*)||_2^2\} \tag{3.17}$$

$$s.t.y = D\psi\alpha \tag{3.18}$$

where λ_i and η_i are diagonal weighting matrices composed of $\lambda_{i,j}$ and $\eta_{i,j}$,

$$\lambda_{i,j} = \frac{c_1}{|\alpha_{i,j}^{(l)}| + \varepsilon} \tag{3.19}$$

$$\eta_{i,j} = \frac{c_2}{(\alpha_{i,j}^{(l)} - \alpha_{i,j}^*)^2 + \varepsilon} \tag{3.20}$$

$\alpha_{i,j}^{(l)}$ is the value of $\alpha_{i,j}$'s -th iteration, c_1 and c_2 are preset constants, ε is a small positive number used to increase the stability of the above formula.

3.2.2. Algorithm solution. The constraint minimization is solved by variable decomposition, and the objective function is mentioned as follows:

$$\begin{aligned}(\hat{x}, \{\hat{\alpha_i}\}) =& arg\,\underset{x,\{\alpha_i\}}{min}\{||y - DSx||_2^2 + \beta\sum_{i=1}^{N}||R_ix - \psi\alpha_i||_2^2 + \sum_{i=1}^{N}||\lambda_i\alpha_i||_1 \\ &+ \sum_{i=1}^{N}||\eta_i(\alpha_i - \alpha_i^*)||_2^2\}\end{aligned} \tag{3.21}$$

$$s.t.y = Dx \tag{3.22}$$

Among them, R_i is used to extract the image block x_i of the image at i. if there are enough parameters, R_ix_i is very close to ψx_i, and the objective functions 3.21 and 3.22 are close to equations 3.17 and 3.18.

This paper first clusters the image blocks into K clusters and learns the PCA Dictionary of each cluster. By introducing ψ_k into equations 3.21 and 3.22, using an adaptive dictionary, the objective function can be written as

$$\begin{aligned}(\hat{x}, \{\hat{\alpha_i}\}) =& arg\,\underset{x,\{\alpha_i\}}{min}\{||y - DSx||_2^2 + \beta\sum_{k=1}^{K}\sum_{i\in C_k}||R_ix - \psi_k\alpha_i||_2^2 + \sum_{i=1}^{N}||\lambda_i\alpha_i||_1 \\ &+ \sum_{i=1}^{N}||\eta_i(\alpha_i - \alpha_i^*)||_2^2\}\end{aligned} \tag{3.23}$$

$$s.t.y = Dx \tag{3.24}$$

Where C_k is the index set of image blocks in cluster k. Fixed sparsity coefficient $\{\alpha_i\}$, x can get the optimal result by minimizing the following Equation:

$$\hat{x} = arg\min_{x}\{||y - DSx||_2^2 + \beta \sum_{k=1}^{K} \sum_{i \in C_k} ||R_i x - \psi_k \alpha_i||_2^2\} \tag{3.25}$$

$$s.t.y = Dx \tag{3.26}$$

$$\begin{aligned}\{\hat{\alpha_i}\} =& arg\min_{\{\alpha_i\}}\{|\beta \sum_{k=1}^{K} \sum_{i \in S_k} ||R_i x - \psi_k \alpha_i||_2^2 \\ &+ \sum_{i=1}^{N} ||\lambda_i \alpha_i||_1 + \eta \sum_{i=1}^{N} ||\eta_i(\alpha_i - \alpha_i^*)||_2^2\}\end{aligned} \tag{3.27}$$

The sparse coefficient $\{\alpha_i\}$ is obtained.

The above optimization process is iterated until convergence. In the iterative process, the β Makes equations 3.21 and 3.22 better approach equations 3.17 and 3.18. Then, equations 3.25 and 3.26 are solved by using the enhanced Lagrange multiplier (ELM), and equations 3.25 and 3.26 are converted into

$$\begin{aligned}L(x, Z, \mu) =& ||y - DSx||_2^2 + \beta \sum_{k=1}^{K} \sum_{i \in C_k} ||R_i x - \psi_k y \alpha_i||_2^2 + \langle z, y - Dx\rangle \\ &+ \mu||y - Dx||_2^2\end{aligned} \tag{3.28}$$

where $\langle z, y - Dx\rangle$ represents the inner product of z and y-Dx, and Z represents the Lagrange multiplier, μ represents a positive scalar. The optimization problems of equations 3.25 and 3.26 can be solved by ALM, which consists of the following iterations:

$$x^{(l+1)} = arg\min_{x} L(x, Z^{(l)}, \mu^{(l)}) \tag{3.29}$$

$$Z^{(l+1)} = Z^{(l)} + \mu^{(l)}(y - Dx^{(l+1)}) \tag{3.30}$$

$$\mu^{(l+1)} = \tau\mu^{(l)} \tag{3.31}$$

where $\tau$is a constant greater than 1. Fix $Z^{(l)}$ and $\mu^{(l)}$. In the above formula, by making $\partial L(x, Z^{(l)}, \mu^{(l)})/\partial x = 0$, we can derive

$$\begin{aligned}\hat{x}^{(l+1)} =& \left[(DS)^T DS + \beta \sum_{i=1}^{N} R_i^T R_i + \mu^{(l)} D^T D\right]^{-1} \\ &\cdot \left[(DS)^T y + \beta \sum_{i=1}^{N} R_i^T R_i(\psi_k \alpha_i) + \frac{D^T Z^{(l)}}{2 + \mu^{(l)} D^T y}\right]\end{aligned} \tag{3.32}$$

Because the right inversion matrix of the above formula is large, the conjugate gradient algorithm (CG) is used to calculate X. With the updated estimation of X, it is easy to update Z and μ. The process can be iterated until convergence [17].

For the given x, Equation 3.27 is a typical sparse coding problem based on image blocks. For each image block, the sparse coding problem is as follows:

$$\hat{\alpha_i} = arg\min_{\alpha_i}\{\beta||R_i x - \psi_k \alpha_i||_2^2 + ||\lambda_i \alpha_i||_1 + ||\eta_i(\alpha_i - \alpha_i^*)||_2^2\} \tag{3.33}$$

Table 4.1: PSNR comparison of four methods

Test image	Bicubic Interpolation [18]	Sparse Coding [19]	Quick Method [20]	Proposed Method
No. 1	26.75	28.79	30.67	31.21
No. 2	24.94	28.85	30.81	31.37
No. 3	26.52	27.38	31.35	31.43
No. 4	26.41	30.25	29.11	31.49
No. 5	25.12	28.84	29.59	30.43
No. 6	25.62	27.72	30.77	31.31
No. 7	26.47	29.58	30.63	31.36
No. 8	26.13	27.61	31.09	31.39
No. 9	25.35	28.50	28.47	29.89
No. 10	25.98	28.50	30.38	31.12

To solve the nonlocal regularized sparse coding problem, the iterative shrinkage method is extended from dealing with one l_1 norm constraint to mixed l_1 and l_2 norm constraints. The closed shrinkage function is derived as follows:

$$\hat{\alpha_{i,j}} = \begin{cases} 0, |v_{i,j} \leqslant \frac{\tau_{1,j}}{2\tau_{2,j}+1} \\ v_{i,j} - sign(v_{i,j})\frac{\tau_{1,j}}{2\tau_{2,j}+1}, other \end{cases} \tag{3.34}$$

where

$$\begin{aligned} v_{i,j} = \frac{\gamma_{i,j}}{2\tau_{2,j}+1}, \gamma_i = \frac{2\eta_i \alpha_i^*}{\beta} + \psi_k^T R_i x_i \\ \tau_{1,j} = \frac{\lambda_{i,j}}{\beta}, \tau_{2,j} = \frac{\eta_{i,j}}{\beta} \end{aligned} \tag{3.35}$$

4. Results and Discussion. To verify the effect of this method in improving the resolution of medical images, the proposed method is compared with the bicubic interpolation method, sparse coding-based method (SC), and fast direct super-resolution method. This paper uses the real CT/MR images provided by TCIA and LIDC in the experiment, and the test images are randomly selected.

Some parameters need to be present in the actual experiment. The image block size is set to 5 × 5. The step size is 2. Number of clusters $K = 60$, $\gamma = 42000, \mu^{(0)} = 1.4, \tau = 1.2, \beta^{(0)} = 0.1, \rho = 2$, λ and η balance sparse regularization and data fidelity terms. Initialization, when calculating the adaptive regularization parameters λ_i and η_i, it is necessary to set the values of c_1 and c_2. Through repeated experiments, this paper sets, $c_1 = 0.25$ and $c_2 = 3.6$ to balance the operation speed and visual effect.

In the data set of this paper, 10 CT/MR images are randomly selected as test images. Each image's peak signal-to-noise ratio (PSNR) value and structural similarity (SSIM) value are calculated and compared with other existing methods. Table 4.1 shows the comparison of PSNR with other conventional methods. Figure 4.1 and Figure 4.2 show the comparison results of corresponding PSNR and SSIM values, respectively.

It can be seen from the chart that the PSNR value and SSIM value of the proposed method are higher, which are more similar to the original HR image. The proposed method has achieved satisfactory results in SR reconstruction of medical images.

5. Conclusion. This paper proposes a target image processing based on super-resolution reconstruction and a machine learning algorithm. Starting from the sparse representation, this paper introduces the existence of similar image structures in medical images. The self-similarity characteristics of medical images are added to the sparse representation theory of images, and the PCA dictionary learning is used to complete the SR reconstruction process of medical images. Experimental analysis is conducted through subjective visual judgment and objective data comparison. From the experimental results, it can be seen that the reconstruction effect

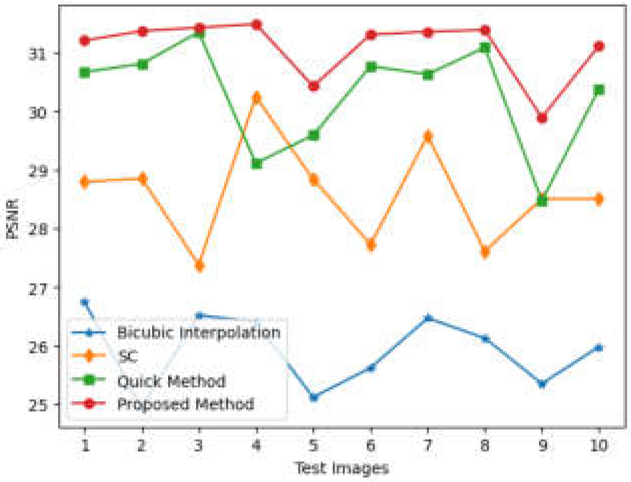


Fig. 4.1: PSNR Comparison of three models with the proposed method

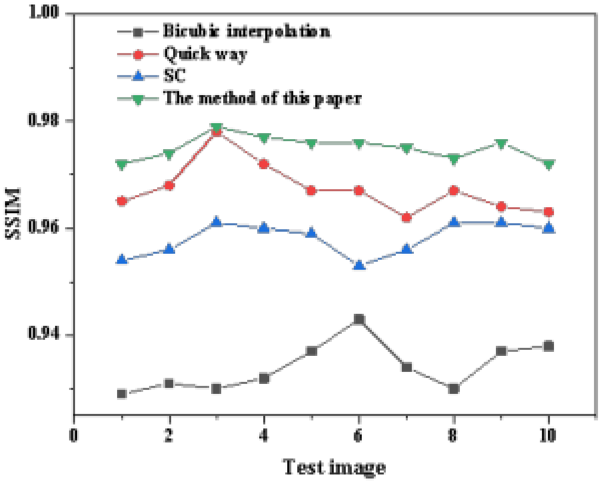


Fig. 4.2: SSIM Comparison of three models with the proposed method

of this method is good. In addition, this paper first clusters the images and significantly reduces the time of querying similar image blocks. The proposed method exhibits a long running time if the number of datasets is huge.

REFERENCES

[1] Junrui, X., Yutan, W., Aili, Q., Jiaxin, Z., Zhenwei, X., Haiyan, W., & Haowei, S. (2021). Image segmentation method for Lingwu long jujubes based on improved FCN-8s [J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 37(5), 191-197.

[2] Liu, B. , & Chen, J. . (2021). A super resolution algorithm based on attention mechanism and SRGAN network. IEEE Access, PP(99), 1-1.

[3] McAuliffe, M. J., Lalonde, F. M., McGarry, D., Gandler, W., Csaky, K., & Trus, B. L. (2001, July). Medical image processing, analysis and visualization in clinical research. In Proceedings 14th IEEE Symposium on Computer-Based Medical Systems. CBMS 2001 (pp. 381-386). IEEE.

[4] Jin, W., Li, X., Fatehi, M., & Hamarneh, G. (2023). Guidelines and evaluation of clinical explainable AI in medical image analysis. Medical Image Analysis, 84, 102684.

[5] Li, K., Yang, S., Dong, R., Wang, X., & Huang, J. (2020). Survey of single image super-resolution reconstruction. IET Image Processing, 14(11), 2273-2290.

[6] Zhang, S., Liang, G., Pan, S., & Zheng, L. (2018). A fast medical image super resolution method based on deep learning network. IEEE Access, 7, 12319-12327.

[7] Zhu, Y., Zhou, Z., Liao, G., & Yuan, K. (2020, April). Csrgan: medical image super-resolution using a generative adversarial network. In 2020 IEEE 17th International Symposium on Biomedical Imaging Workshops (ISBI Workshops) (pp. 1-4). IEEE.

[8] Zhang, K., Tao, D., Gao, X., Li, X., & Xiong, Z. (2015). Learning multiple linear mappings for efficient single image super-resolution. IEEE Transactions on Image Processing, 24(3), 846-861.

[9] Bhatti, U. A. , Yu, Z. , Yuan, L. , Nawaz, S. A. , & Wen, L. . (2020). Geometric algebra applications in geospatial artificial intelligence and remote sensing image processing. IEEE Access, PP(99), 1-1.

[10] Bao, C., Ji, H., Quan, Y., & Shen, Z. (2015). Dictionary learning for sparse coding: Algorithms and convergence analysis. IEEE transactions on pattern analysis and machine intelligence, 38(7), 1356-1369.

[11] Xiao, Z., Du, N., Liu, J., & Zhang, W. (2021). SR-Net: a sequence offset fusion net and refine net for undersampled multislice MR image reconstruction. Computer Methods and Programs in Biomedicine, 202, 105997.

[12] Zhang, H., Zhang, L., & Shen, H. (2012). A super-resolution reconstruction algorithm for hyperspectral images. Signal Processing, 92(9), 2082-2096.

[13] Niu, X. (2018, December). An overview of image super-resolution reconstruction algorithm. In 2018 11th International Symposium on Computational Intelligence and Design (ISCID) (Vol. 2, pp. 16-18). IEEE.

[14] Katkar, J., Baraskar, T., & Mankar, V. R. (2015, October). A novel approach for medical image segmentation using PCA and K-means clustering. In 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT) (pp. 430-435). IEEE.

[15] Zhao, H. , Liu, Y. , Huang, C. , & Wang, T. . (2020). Hybrid-weighted total variation and nonlocal low-rank based image compressed sensing reconstruction. IEEE Access, PP(99), 1-1.

[16] Fayez, M., Safwat, S., & Hassanein, E. (2016, July). Comparative study of clustering medical images. In 2016 SAI Computing Conference (SAI) (pp. 312-318). IEEE.

[17] Zha, Z. , Yuan, X. , Zhou, J. , Zhu, C. , & Wen, B. . (2020). Image restoration via simultaneous nonlocal self-similarity priors. IEEE Transactions on Image Processing, PP(99), 1-1.

[18] Triwijoyoa, B. K., & Adila, A. (2021). Analysis of medical image resizing using bicubic interpolation algorithm. J. Ilmu Komput, 14(2), 20-29.

[19] Zhang, R., Shen, J., Wei, F., Li, X., & Sangaiah, A. K. (2017). Medical image classification based on multi-scale non-negative sparse coding. Artificial intelligence in medicine, 83, 44-51.

[20] Zhang, S., Liang, G., Pan, S., & Zheng, L. (2018). A fast medical image super resolution method based on deep learning network. IEEE Access, 7, 12319-12327.

Edited by: Venkatesan C
Special issue on: Next Generation Pervasive Reconfigurable Computing for High Performance Real Time Applications
Received: Sep 27, 2023
Accepted: Feb 8, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1341–1349, DOI 10.12694/scpe.v25i3.2641

PERFORMANCE EVALUATION OF MICRO AUTOMATIC PRESSURE MEASUREMENT SENSOR FOR ENHANCED ACCURACY

SHUIQUAN ZHU*

Abstract. The major objective of this research is to design sensitive components, conversion components, and various sensor circuits to achieve the miniaturization design for more accurate measurements. This article conducts performance testing on the designed miniaturized pressure sensor to determine whether it meets the qualified standards. The response time of designed sensor results increases with the increase of pressure under experimental conditions of different pressure application values (5MPa to 50MPa). The detection accuracy of the micro automatic pressure measurement sensor designed in this paper can reach 0.0452%; the average pressure is 0.00364%, and the insulation resistance is 68.44 megohms, which meets reliability requirements. The sensitivity is 0.0582%; the nonlinearity is 0.0741%; the hysteresis is 0.0266%; The repeatability of 0.0625% meets the qualification standard for this instrument. Still, compared with traditional sensors, the sensor reduces the response time of results by about 60%. However, the author conducted the detection in an ideal environment. The actual working environment of sensors is relatively good. Therefore, the detection results obtained in this article may have some errors compared to the actual situation, and further analysis and testing are needed to optimize the performance of the designed sensor.

Key words: Micro-mechanical electronic technology, automatic measurement sensor, sensitive diaphragm, conversion element

1. Introduction.

Sensors are one of the most representative achievements in modern science and technology development and have the same status as communication technology and computer technology used for inertial navigation and space attitude determination [1]. Biosensors, fluid sensors, etc., are used in biomedicine for clinical measurements and pathological diagnosis [2]. In environmental monitoring, temperature, humidity, and gas sensors are used to monitor changes in the surrounding environment to determine whether pollution or adverse weather has occurred [3]. To meet the requirements of more refined monitoring work, sensors are gradually developing toward more intelligence, automation and miniaturization. Micro-detectors are microsensors, the most commonly used field in miniaturization development. Based on the different uses and types of microsensors, sensors are divided into capacitive MEMS differential pressure sensors, magnetic field measurement microsensors, Hall-type magnetic liquid micro differential pressure sensors, etc. [4,5]. Based on this, the author selects a commonly used automatic pressure measurement sensor as the object to research the design of automatic measurement sensors for micro-mechanical electronic technology. The design is divided into two parts: theoretical design and application testing. It includes several modules, such as sensitive diaphragm design, conversion element design, signal conversion circuit design, sensor interface circuit design and working program design [6,7].

Figure 1.1 shows a reference for producing and manufacturing microsensors and promotes the application and development of micromechanical electronic technology.

2. Literature Review.

In recent years, issues such as medical diagnosis, environmental hygiene, and food safety have occurred frequently in China. This places higher demands on the rapid and real-time detection of viruses, antigens, and other substances, also known as point-of-care testing [8]. In 2014, the globally renowned market research company Persistence Market Research (PMR) released an authoritative survey report on the future development of biosensors. The report indicates that the global biosensor market will grow rapidly in the next six years. In 2014, the market value of biosensors was $12.9 billion, and by 2020, the market value will increase to $22.5 billion, with a compound annual growth rate of 9.7% [9,10]. Among them, biosensors have the most applications in the POCT field, and the demand is also the largest and has been increasing continuously [11]. Therefore, it is urgent to research point-of-care detection technology, which has great economic and

*Tianjin Electronic Information College, Tianjin, 300350, China (Corresponding author's e-mail: shuiquanzhu@126.com)

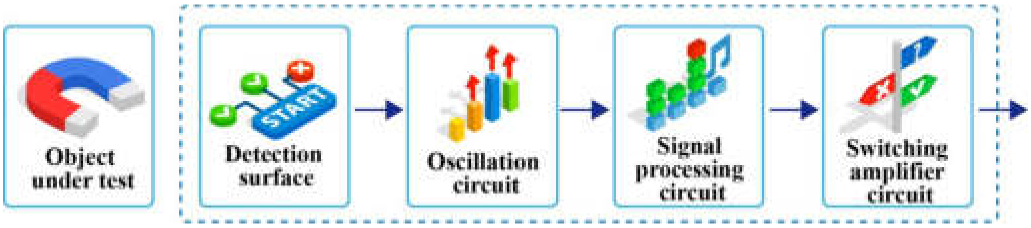


Fig. 1.1: Measurement process for automation control

social significance [12].

The basic principle of biosensors is that microelectronic control systems control the flow of sensitive membranes through the microfluidic environment of the tested object (antibodies, viruses, bacteria, etc.). The substance to be tested is adsorbed on the surface of the sensitive membrane, and the transducer converts the biological signals on the sensitive membrane into electrical signals. Then, it is measured through a dedicated signal detection system. Finally, the concentration and quality are achieved by measuring various characteristics of the sensors [13, 14]. In addition, driven by microelectromechanical systems (as well as microelectronics technology), piezoelectric biosensor systems are also developing towards automation, intelligence, and miniaturization. This lays a solid foundation for applying piezoelectric sensors in point-of-care testing (POCT) [15]. The researchers observed that an anisotropic crystal generates charges on the surface under external mechanical pressure. When the external force is removed, the surface charge disappears. This phenomenon is called the Piezoelectric effect [16]. The piezoelectric effect is divided into the positive and inverse piezoelectric effects. Specifically, the positive piezoelectric effect means that when an external force acts on the piezoelectric crystal in a certain direction, electric polarization will occur inside the crystal, at the same time, positive and negative charges are generated on the two surfaces of the crystal; When the crystal loses the external force, it is in an uncharged state; when an external force is applied in the other direction of the crystal, the charge polarity will be changed [17, 18]. Since the dielectric constant, elastic constant and piezoelectric constant of piezoelectric crystals tested with different boundary conditions have little difference in general, therefore, generally do not distinguish between the elastic constant of the circuit and the elastic constant of the short circuit, the free permittivity and the clamping permittivity, at this point, the relationship between the stress T and the electric field E in the piezoelectric crystal can be established, which is also the basis for this effect to be applied to automated measurement sensors [19].

3. Research Methods.

3.1. Microsensors for MEMS. The types of microsensors in the MEMS are different; the author uses the pressure sensor to conduct a design study. The pressure sensor is based on the piezoresistive effect combined with Ohm's law to estimate the pressure value. When a certain pressure is applied to the semiconductor material (sensitive diaphragm) from the outside, the material will undergo corresponding strain, and the strain will drive the resistance to change at the same time, at this time, the bridge will lose its balance and the corresponding voltage value will be output, then the actual pressure value can be converted by using Ohm's law [20, 21].

The mathematical Equation of resistance change is expressed as:

$$\frac{\Delta R}{R} \approx \pi_1 \sigma_1 + \pi_t \sigma_t \tag{3.1}$$

In Equation 3.1, $\frac{\Delta R}{R}$ is the rate of change of the varistor; π_1 and π_t represent the transverse piezoresistive coefficient and the longitudinal piezoresistive coefficient, respectively; σ_1 and σ_t represent the transverse stress and the longitudinal stress, respectively.

When the semiconductor material (sensitive diaphragm) is not affected by the force, the four bridges are always in a balanced state, and the output voltage at this time = 0. Its Equation is expressed as

$$R_1 = R_2 = R_3 = R_4 = 0 \tag{3.2}$$

In Equation 3.2, R_1, R_2, R_3,R_4 and represent the resistance values of the four arms of the Wheatstone Bridge.

When the outside world pressures the semiconductor material (sensitive diaphragm), the bridge connected to it will be out of balance. At this time, the voltage output is expressed as

$$V_{out} = \frac{(R_1 + \Delta R)(R_3 + \Delta R) - (R_2 - \Delta R)(R_4 - \Delta R)}{(R_1 + R_2)(R_3 + R_4)} V_i \tag{3.3}$$

When R_1, R_2, R_3, and R_4 are all equal to the same value, it can be simplified to formula 3.4:

$$V_{out} = \frac{\Delta R}{R} V_i \tag{3.4}$$

In the formula, V_i is the supply voltage.

Before designing, it is necessary to understand the structure and composition of the micro pressure sensor, that is, the design of the sensitive diaphragm, the design of the conversion element, the design of the signal conditioning and conversion circuit, the design of the sensor interface circuit, and the design of the sensor working program. These design modules are analyzed in detail [22].

3.2. Sensitive Diaphragm Design. The sensitive diaphragm is the most critical component in automatic measurement sensors, which can directly sense the measured value and is the only external force contactor and voltage output source. In this article, the sensitive diaphragm design adopts the MEMS etching process in silicon micromachining technology, and the specific process is as follows [23]:

Step 1: Select the substrate of the sensitive diaphragm, that is, the bearing part of the sensitive element. Here, single-crystal silicon is selected as the substrate of the sensitive diaphragm;

Step 2: Monocrystalline silicon processing. The production of the sensitive film needs to ensure that the single crystal silicon wafer is free of any damage, needs to be ground and polished, and then needs to be cleaned to remove impurities attached to the surface. After that, wait for etching;

Step 3: Etching active marks, that is, preparing alignment marks and scribing frames;

Step 4: Depositing a masking layer, that is, covering a layer of protective film on the monocrystalline silicon wafer;

Step 5: Open the varistor and the ohmic contact area, that is, remove part of the silicon on the back of the silicon wafer to form a cavity;

Step 6: The contact area is implanted with ions to form varistor strips;

Step 7: Standard cleaning monocrystalline silicon wafer;

Step 8: Deposition-layer SiO_2, as isolation layer;

Step 9: Use magnetron sputtering technology to sputter Ti-Pt-Au on the glass sheet to form an electrode plate;

Step 10: Photolithography of electrode hole area and lead area pattern;

Step 11: Reduce the thickness of the monocrystalline silicon wafer;

Step 12: Bonding the silicon wafer and glass together;

Step 13: Packaging.

To ensure the sensitivity and accuracy of the sensitive chip, the external environment should not be corrupted. It is necessary to use an isolation diaphragm to separate the sensitive diaphragm from the measured medium to avoid contamination and damage.

3.3. Conversion Elements. The automated measurement sensor in this study is a miniature pressure sensor, and its measurement principle is based on the piezoresistive effect; the crimp resistor is the transducer element of the sensor. The varistor is an electrical signal that converts the strain force of the sensitive abdominal piece feels into a resistance value. Then, it converts into a voltage in combination with the given supply voltage, so the varistor design is very important [24].

3.4. Design of Signal Conditioning Circuit for Measurement Sensor. The function of the signal conditioning conversion circuit is to amplify, modulate and filter the signal to improve the signal quality. Figure 3.1 is a signal conditioning conversion circuit designed for an automated measurement sensor.

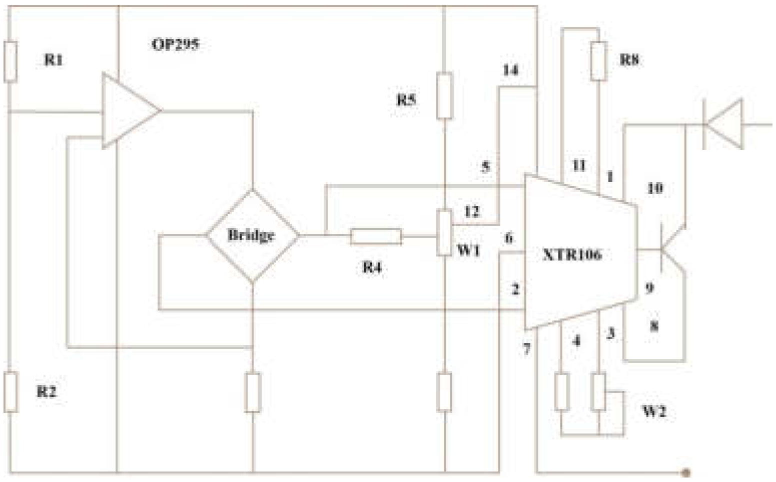


Fig. 3.1: Design of signal conditioning and conversion for the automatic measurement sensor

Table 3.1: The function table of each interface circuit pin of the sensor

Pin number	Pin symbol	Pin function
1	GND	Ground, Power negative
2	SYNC	Circuit common ground voltage
3	DVCC	Voltage into the circuit
4	VDD	The working voltage of the chip
5	NC	Empty feet for extended functions
6	V1P	Empty feet, please hang
7	V1N	The turn-off control signal of op-amp unit 1
8	V2N	Op-amp unit 2 shutdown control signal
9	V2P	Timer interrupt
10	RESET	The chip is not powered on, reset the sensor
11	AGND	Serial data, Single bus
12	GIN	Signal output
13	DGND	Digital analog
14	CLKOUT	Output a clock signal by default

3.5. Sensor Interface. The designed automatic measurement sensor is the main part of the micro-mechanical electronic system, so the sensor must be connected to other external devices. Therefore, the sensor interface design is reasonable, and the data collected by the sensor can be used directly by the MEMS system [25]. The sensor interface, also known as the pin, is the interface between the internal circuit and the peripheral device, and all the pins are the sensor's interface. There are 14 sensor pins, and each pin has different functions. Table 3.1 shows the pin functions of each interface circuit of the sensor.

3.6. Sensor Working Procedure. The automatic measurement sensor design belongs to the hardware category. In addition, it is necessary to write and design a working program to provide logical guidance. The working procedure of the sensor is the process of collecting data from the sensor. The steps of the sensor working procedure are as follows:

Step 1: Initialize the sensitive chip;

Step 2: Initialize each circuit;

Step 3: Wait for the collection order;

Step 4: Judge whether the acquisition command arrives. If it arrives, then enter the next step; otherwise,

Table 4.1: Dimensional Specifications for Microsensor Components in 3D and 2D Models

Geometry	3D model (μm)	2D model (μm)
Interdigital electrode width	5	5
Interdigital electrode length	40	-
Interdigital electrode thickness	0.3	0.3
Adjacent interdigital electrode spacing	5	5
Delay line width	40	500
Delay line length	40	-
Delay line thickness	0.3	0.3
Number of interdigitated electrode pairs	5+1 (pair)	20+20 (pair)
Waveguide layer width	200	910
Waveguide layer length	60	-
Waveguide layer thickness	2	2
Base width	200	910
Base length	60	-
Substrate thickness	40	300

enter the low power consumption state and return to the previous step;

Step 5: The sensor perceives the measured quantity;

Step 6: Utilize the conversion element to convert the measurement amount into an electrical signal;

Step 7: Utilize the adjustment conversion circuit to amplify and filter the electrical signal. After the filter sensor completes the acquisition task, it is a very important link because the initial information collected by the sensor contains a lot of noise, and the amount of useful information is covered up, so filtering is necessary, so it is essential to use the adjustment conversion circuit for filtering;

Step 8: If the collection time arrives, then enter the interrupt program; otherwise, send the collected receiver;

Step 9: If the next collection round is entered, go back to step 5. Otherwise, wait for the shutdown command to shut down the entire sensor system.

4. Result Analysis and Discussions.

4.1. Sensor Materials and Assembly. According to the theoretical design of the automatic measurement sensor, relevant materials are used to prepare a miniature pressure sensor. The dimensions of the sensor preparation materials are shown in Table 4.1.

The table's materials are assembled to prepare a finished automated measurement sensor. Due to the different content of subsequent tests, six finished samples were prepared with six copies of the same material. The expected standards of the finished automated measurement sensor prepared are as follows:

Accuracy: $\leqslant \%F \cdot S \pm 0.10$;

Nonlinear: $\leqslant \%F \cdot S \pm 0.13$;

Hysteresis: $\leqslant \%F \cdot S \pm 0.12$;

Repeatability: $\leqslant \%F \cdot S \pm 0.11$.

The functional diagram of the pressure measurement system is shown in Figure 4.1. The system mainly comprises a capacitive micro-sensor, signal processing circuit, low-pass filter, amplification circuit, A/D conversion circuit, single-chip microcomputer and its peripheral devices. Capacitive micro-sensor converts the measured pressure change into capacitance change. The signal processing circuit synchronously detects the capacitance change signal and obtains a DC signal after filtering by a low-pass filter. We must amplify the signal to ensure direct analog-to-digital conversion of the signal and improve the signal's anti-interference ability and the instrument's sensitivity. The amplified DC voltage signal is transmitted to the single-chip microcomputer system through A/D conversion and then processed and analyzed by the computer.

4.2. Sensor Performance Test . The performance detection of an automatic measurement sensor is analyzed with the help of the following test items.

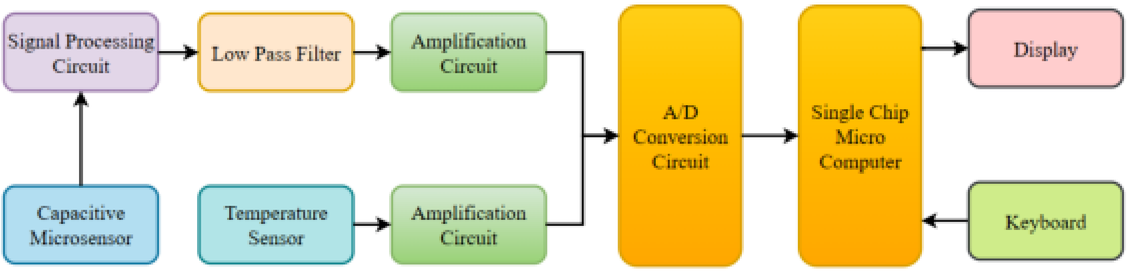


Fig. 4.1: Functional diagram of the pressure measurement system

1) Accuracy test. Accuracy is the error between the value collected by the sensor and the pressure value acting on the sensor. The test device is a universal pressure testing machine, and the designed sensor is placed directly under the lower plate of the testing machine. Then different pressure application values (5MPa to 50MPa) are set, the test is ten times, the absolute average value, and finally, the difference with the actual pressure value, and the verification difference is $\leqslant \% F \cdot S \pm 0.10$.

2) Reliability. Sensor reliability detection refers to the degree of sealing of the detection interface, whether air or electric leakage exists. The test devices for the above two phenomena are pressure leak detectors and megohmmeters. The former qualification standard is the average value $\leqslant$ 0.05% of the applied stable pressure value, and it is considered that there is no leakage problem at the interface. The latter qualification standard is the insulation resistance value $\geqslant$ 50 megohms, and it is considered that there is no leakage problem at the interface.

3) Sensitivity. Sensitivity refers to the degree to which the sensor's output changes as the input changes. The data required for the test comes from the data measured by the device. The sensitivity is calculated using Equation 4.1:

$$S = \frac{\partial C}{\partial P} = \frac{C_0}{2P}\left(\frac{1}{1+\frac{Pd}{P_m g}} - \frac{tanh^{-1}\sqrt{\frac{Pd}{P_m g}}}{\sqrt{\frac{Pd}{P_m g}}}\right) \times 100\% \tag{4.1}$$

In Equation 4.1, C represents the increment of the sensor output; g represents the distance between the initial voltage value of the sensor and the electrode; C_0 represents the initial voltage value of the sensor; P_m represents the maximum pressure when the maximum deflection of the center of the diaphragm is equal to the constant d, and P represents the pressure of the outside gas.

4) Nonlinearity. Nonlinearity means that according to the time series, the fitted deviation between the curve drawn by the output of the measured value by the automated measurement sensor and the actual pressure curve. The measurement formula is shown in Equation 4.3:

$$R_L = \frac{\Delta L_{max}}{F_{YS}} \times 100\% \tag{4.2}$$

In the Equation, ΔL_{max} represents the maximum deviation between the curve drawn by the measured value output by the automatic measurement sensor and the actual pressure curve; F_{YS} represents the full-scale output voltage value.

5) Hysteresis. Hysteresis refers to the maximum difference between the output voltage values of the two when the input value increases and decreases at the same test point within the full-scale range. The data sources required for the test are the same as above. The hysteresis can be calculated by Equation 4.4:

$$Y_H = \frac{\Delta C_{max}}{F_{YS}} \times 100\% \tag{4.3}$$

where ΔC_{max} represents the maximum hysteresis error of the pressure sensor within the test range. F_{YS} represents the full-scale output voltage value.

Table 4.2: Performance evaluation of an automatic measurement sensor

Test items		Test results	Eligibility criteria
Precision		0.0451%	Qualified
Reliability	Pressure average	0.00363%	Qualified
	Insulation resistance value	68.44 megohms	Qualified
Sensitivity		0.0581%	Qualified
Nonlinearity		0.0740%	Qualified
Hysteresis		0.0264%	Qualified
Repeatability		0.0624%	Qualified

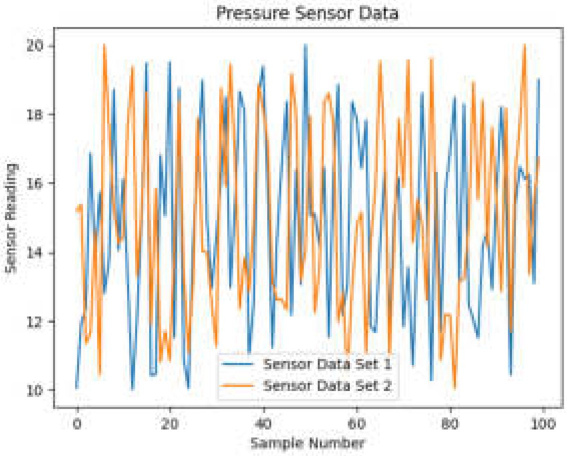


Fig. 4.2: Performance comparison of two different sensor data

6) Repeatability. The degree of agreement between the results of multiple consecutive measurements under the same test conditions required for the test is the same. The formula for finding the repeatability is given by Equation 4.5:

$$F = \frac{2a \sim 3a}{Y_{FS}} \times 100\% \tag{4.4}$$

where a represents the Bessel standard deviation.

The test results of the miniature pressure automatic measurement sensor designed by the author are shown in Table 4.2.

Table 4.2 compares the test qualification standards. The actual test results are within the range specified by the qualification standards, proving that the designed sensor meets expectations and can be used in actual pressure testing. Figure 4.2 serves as a visual aid for scrutinizing the performance of a pressure sensor over time, enabling a direct comparison between two distinct sets of sensor data.

Figure 4.3 shows that under the experimental conditions of different pressure application values (5MPa to 50MPa), the sensor's response time increases with pressure.

5. Conclusion. The sensor designed in this article showcases a broad spectrum of applications, highlighting its versatility in measuring and detecting various parameters. This adaptability significantly simplifies the data collection processes, catering to the needs of both individuals and professionals. However, among the ongoing exploration of sensor technology, a noticeable trend towards increased miniaturization becomes apparent, propelled by the pursuit of heightened precision in measurements. The author emphasizes the intricate design of sensitive elements, conversion components, and various circuits essential to the sensor's functionality.

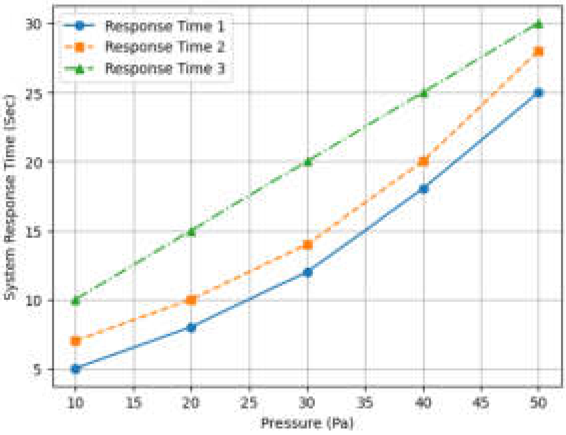


Fig. 4.3: System response time under different pressure values

Following the design phase, a meticulous performance testing protocol is implemented for the sensor products, ensuring compliance with rigorous product qualification standards. This comprehensive approach guarantees that the designed sensor meets and exceeds the required benchmarks, reinforcing reliability and efficiency across diverse applications.

REFERENCES

[1] Kurmendra, & Kumar, R. . (2021). A review on rf micro-electro-mechanical-systems (mems) switch for radio frequency applications. Microsystem Technologies, 27(7), 2525-2542.
[2] Boniface, M. , Plodinec, M. , Schlgl, R. , & Lunkenbein, T. . (2020). Quo vadis micro-electro-mechanical systems for the study of heterogeneous catalysts inside the electron microscope?. Topics in Catalysis, 63(15-18), 1623-1643.
[3] Jena, S. , & Gupta, A. . (2021). Review on pressure sensors: a perspective from mechanical to micro-electro-mechanical systems. Sensor Review, 41(3), 320-329.
[4] Song, P., Ma, Z., Ma, J., Yang, L., Wei, J., Zhao, Y., ... & Wang, X. (2020). Recent progress of miniature MEMS pressure sensors. Micromachines, 11(1), 56.
[5] Lysenko, I. E. , Tkachenko, A. V. , Ezhova, O. A. , Konoplev, B. G. , & Sherova, E. V. . (2020). The mechanical effects influencing on the design of rf mems switches. Electronics, 9(2), 207.
[6] Schmitt, P. , & Hoffmann, M. . (2020). Engineering a compliant mechanical amplifier for mems sensor applications. Journal of Microelectromechanical Systems, PP(99), 1-14.
[7] Granados-Ortiz, F. J. , & Ortega-Casanova, J. . (2020). Mechanical characterization and analysis of a passive micro heat exchanger. Micromachines, 11(7), 668.
[8] Liu, S. , Wang, K. , Wang, B. , Li, J. , & Zhang, C. . (2021). Size effect on thermo-mechanical instability of micro/nano scale organic solar cells. Meccanica, 57(1), 87-107.
[9] Lin, H. , Y Li, Lam, J. , & Wu, Z. G. . (2021). Multi-sensor optimal linear estimation with unobservable measurement losses. IEEE Transactions on Automatic Control, PP(99), 1-1.
[10] Evtikhiev, Kozlov, A. V. ,Krasnov, Rodin, V. G. , Starikov, R. S. , & Cheremkhin, P. A. . (2021). Estimation of the efficiency of digital camera photosensor noise measurement through the automatic segmentation of non-uniform target methods and the standard emva 1288. Measurement Techniques, 64(4), 296-304.
[11] Zheng, T. , Yu, Z. , Wang, J. , & Lu, G. . (2020). A new automatic foot arch index measurement method based on a flexible membrane pressure sensor. Sensors, 20(10), 2892.
[12] Dufour, D. , Noc, L. L. , Tremblay, B. , Tremblay, M. N. , & Topart, P. . (2021). A bi-spectral microbolometer sensor for wildfire measurement. Sensors, 21(11), 3690.
[13] Zhang, F. , Li, J. , Lu, J. , & Xu, C. . (2021). Optimization of circular waveguide microwave sensor for gas-solid two-phase flow parameters measurement. IEEE Sensors Journal, PP(99), 1-1.
[14] Offenzeller, C. , Knoll, M. , Voglhuber-Brunnmaier, T. , Hilber, W. , & Jakoby, B. . (2020). Screen printed sensor design for thermal flow velocity measurement with intrinsic compensation of thermal fluid conductivity. IEEE Sensors Journal, PP(99), 1-1.
[15] Guan, X. , Zou, Y. , Shang, J. , Bian, X. , & Li, Q. . (2021). A mutual inductance sensor for cryogenic radial displacement measurement. Cryogenics, 115(5), 103263.
[16] Feng, Y. , Tao, W. , Feng, Y. , Yin, X. , & Zhao, H. . (2020). High-precision width measurement method of laser profile

sensor. Sensor Review, 40(6), 699-707.

[17] Kim, S. G., Priya, S., & Kanno, I. (2012). Piezoelectric MEMS for energy harvesting. MRS bulletin, 37(11), 1039-1050.

[18] Lyu, C. , Li, P. , Wang, D. , Yang, S. , & Sui, C. . (2020). High-speed optical 3d measurement sensor for industrial application. IEEE Sensors Journal, PP(99), 1-1.

[19] Rayguru, M. M. , Elara, M. R. , Balakrishnan, R. , Muthugala, M. , & Samarakoon, S. . (2020). A path tracking strategy for car like robots with sensor unpredictability and measurement errors. Sensors, 20(11), 3077.

[20] Zhao, L. , Zhang, Y. , Chen, Y. , Chen, Y. , Yi, G. , & Wang, J. . (2020). A fiber strain sensor with high resolution and large measurement scale. IEEE Sensors Journal, 20(6), 2991-2996.

[21] Keshari, A. K. , Rao, J. P. , Murthy, A. , & Jayaraman. (2020). Design and development of instrumentation for the measurement of sensor array responses. Review of Scientific Instruments, 91(2), 024101.

[22] Kamp, J. N. , Srensen, L. L. , Hansen, M. J. , Nyord, T. , & Feilberg, A. . (2021). Low-cost fluorescence sensor for ammonia measurement in livestock houses. Sensors, 21(5), 1701.

[23] Kim, S. , & Lee, K. . (2020). Development of underground wireless saw integrated sensor measurement system using magnetic field. Transactions of the Korean Institute of Electrical Engineers, 69(6), 903-913.

[24] Cecchi, S. , Spinsante, S. , Terenzi, A. , & Orcioni, S. . (2020). A smart sensor-based measurement system for advanced bee hive monitoring. Sensors, 20(9), 2726.

[25] Li, C. , Li, B. , Huang, J. , & Li, C. . (2020). Developing an online measurement device based on resistance sensor for measurement of single grain moisture content in drying process. Sensors, 20(15), 4102.

Edited by: Venkatesan C
Special issue on: Next Generation Pervasive Reconfigurable Computing for High Performance Real Time Applications
Received: Sep 27, 2023
Accepted: Jan 22, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1350–1360, DOI 10.12694/scpe.v25i3.2360

USE OF TOPIC ANALYSIS FOR ENHANCING HEALTHCARE TECHNOLOGIES

USHA PATEL,[*] PREETI KATHIRIA,[†] CHAND SAHIL MANSURI,[‡] SHRIYA MADHVANI,[§] AND VIRANCHI PARIKH[¶]

Abstract. Nowadays, technology has played a vital role in the advancement of the healthcare sector. Various healthcare datasets are available on the web in the form of patents, scientific papers, articles, textual feedback, chatlogs, abstracts of papers, medical reports, and social media posts. It is a tedious task for the stakeholders to find hidden crucial knowledge on the discussed topic from this content, which if utilized optimally can lead to the rapid development of the healthcare sector. Topic analysis concepts are very effective in extracting meaningful topics from the data. Here, frequently applied Topic modeling methods -Latent Semantic Analysis, Probabilistic Latent Semantic Analysis, Latent Dirichlet Allocation, Correlated Topics Model, and Non-negative matrix factorization are surveyed along with their benefits and drawbacks. Insights on new innovative topic modeling techniques used in healthcare with their objective, opportunities, and challenges are provided, which can help the researchers for the enhancement of healthcare facilities.

Key words: Topic Analysis, Topic Modeling, Health Care 5.0 , Stress Analysis, Latent Semantic Analysis, Probabilistic Latent Semantic Analysis, Latent Dirichlet Allocation, Correlated Topics Model, Non-negative matrix factorization

1. Introduction. In today's connected world, technology has become vital support in every business industry. Healthcare technology is one of the most critical industries where technology plays a key role. As new and improved technology is finding its applications in the healthcare industry, stakeholders in the health sector are getting more reliant on it for saving countless lives worldwide. It helps doctors to improve their practice by giving an advanced diagnosis to enhance patient care. Healthcare technology includes IT tools or software which is designed for better hospital administration, provide new insights in the field of medicines and treatment, or enhance the complete quality of care. Recent technology development includes Artificial intelligence in healthcare like Natural Language(NLP) technology and Machine learning [28].

The topic analysis is one of the natural language processing techniques which enable us to extract meaningful words from the input text by detecting topics based on their occurrence in the textual data. The topic analysis is the process of analysis that establishes a particular topic structure. This can be of assistance in pointing out which topics are present in the input text and what significance it brings in understanding the predicament at hand. The topic analysis is found useful in diverse applications related to text processing. One such example includes the assistance of topic analysis in the formation of an automatic indexing technique to infer information. It benefits in grasping the main topics and subtopics from the complete input textual data and also provides the information as to where exactly these subtopics are used in the data. The two renowned approaches for using machine learning to achieve topic analysis are topic modeling and topic classification. Topic classification is a supervised machine learning approach, it requires labels to classify the data into different classes. In the medical field, it is difficult to always get the labelled data. Therefore, with the unlabelled data, unsupervised learning - clustering is needed [25], and further topic modeling takes place.

1.1. Importance of topic analysis in Healthcare. Most of the knowledge and information are collectively digitized and stored in the form of scientific articles, books, news, blogs, web pages, and social networks. It becomes somewhat difficult to collect the information in need efficiently and effectively. All this information is in an unstructured way, needed to convert into an organized and structured form. Computational methods

[*] Institute of Technology, Nirma University, India (ushapatel@nirmauni.ac.in)
[†] Institute of Technology, Nirma University, India (Corresponding author, preeti.kathiria@nirmauni.ac.in)
[‡] Institute of Technology, Nirma University, India (19mca016@nirmauni.ac.in)
[§] Institute of Technology, Nirma University, India (shriyu1304@gmail.com)
[¶] Institute of Technology, Nirma University, India (19mced08@nirmauni.ac.in)

and tools are needed to convert it into structured form and also to understand and search these vast amounts of information.[7]

Apps and portals are organizing online health communities which help in patient-doctor interactions and feedback of the patient regarding their overall experience with the organization and their caregivers. The electronic health record system is gaining popularity with physicians and patients. During the COVID-19 pandemic, more telephonic services were used by doctors, medical specialists, patients, and health systems. These practices provide a huge amount of structured as well as unstructured data which leaves an immense possibility for analyzing and understanding the data. The biggest challenge for health and medical data science research is to develop effective methods for finding the concealed meaning in considerable complex medical and healthcare datasets and using them to respond to the questions about that data[35].

Two popular methods utilized in analyzing medical text are bag-of-words and topic modeling. Bag-of-words technique acts on data as documents on the basis of frequency of the words like a matrix [24]. Topic modeling on the other hand is used for obtaining topics from the collection of documents.

The data for analysis not only includes the patient records but can also include various other means of information that can help in understanding the overall medical industry and help in enhancing the technological and administrative practices. The data for analysis can also include the various previous researches and patents information or current news articles or social media discussions on the current medical situations. The previous research and patients help in improving the innovation and also helps in understanding the impact of it in the current commercial market[13]. Scientific knowledge is transformed into new technology and that knowledge becomes the basis of further technological innovations[1].

1.2. Contribution of the paper. This paper discusses the importance of Topic modeling in Healthcare. Well-known Topic modeling techniques with their basic mathematical model, advantages, and disadvantages are discussed. Various forms of data used for topic modeling along with possible types of outcomes are analyzed from the different state-of-the-art publications.

Also, innovative topic models with their objectives used in healthcare are also included. At last major challenges and applications of topic modeling in healthcare are also given.

1.3. Organization of paper. Figure 1.1 reflects the structure of the survey. section - 1 gives the introduction about Topic analysis and its importance in healthcare along with the motivation and contribution of the paper. Section 2 explains Healthcare 5.0 and well-known Topic Modeling methods. Section 3 compares the various forms of data used in the healthcare domain on which topic analysis can be applied. Section 4 represents the advanced methods opted by the topic analysis in the healthcare domain. Section 5 describes the application areas, challenges, and research opportunities in this field.

2. Topic Modelling and Healthcare 5.0. The combination of Topic Modeling and Healthcare 5.0 can prove to be miraculous in the healthcare sector. For instance, we have multiple datasets of various patients and their medical records, which with the help of topic modeling can be used to identify the category of the disease of the future patients based on patterns, findings, and provide diagnosis for the current patients. This can lead the healthcare sector to advancement and help the stakeholders in the healthcare sector to improve healthcare services.

2.1. Healthcare 5.0. The use of technology is driving the health industry into a significant innovative transformation. Using technological devices regularly to monitor one's health helps in enhancing the standard of life. This stage of the utilization of cutting-edge technologies to benefit people in enhancing the functioning in the healthcare industry is referred to as Healthcare 5.0. The fundamental features of healthcare 5.0 include the use of Internet of Things (IoT), 5G communications, and Artificial Intelligence (AI). The research by Mohanta et al. [34] mainly focuses on Healthcare 5.0. It emphasizes the importance of 5G as the fundamental network infrastructure for enabling smart healthcare. Furthermore, Ambient Assisted Living (AAL) technology based on IoT provides variety of resolutions for improving people's quality of life, assisting stakeholders by providing impairments. It provides the products which helps in monitoring day to day health conditions.[43]

Nowadays, a wide variety of healthcare digital data is available in the form of sensor readings, patient detail records, social media, and news articles. Collecting all the data and finding out insights is also a challenging

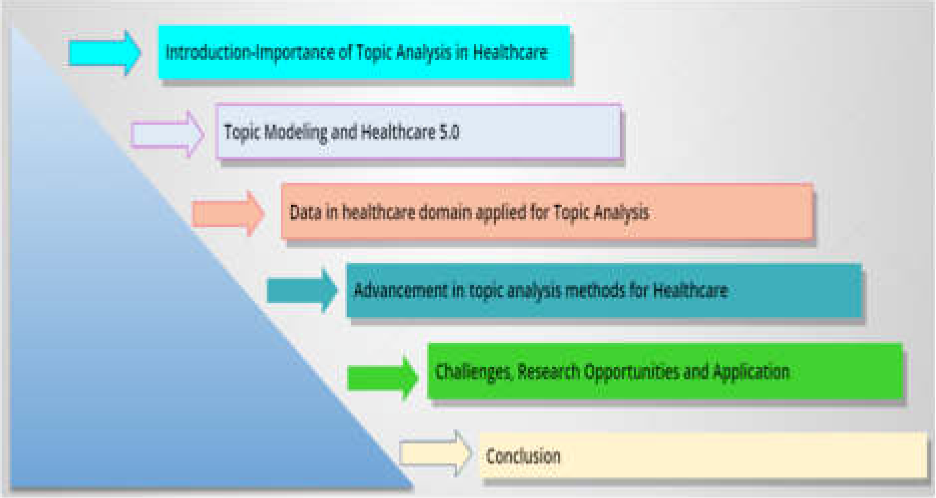

Introduction-Importance of Topic Analysis in Healthcare
Topic Modeling and Healthcare 5.0
Data in healthcare domain applied for Topic Analysis
Advancement in topic analysis methods for Healthcare
Challenges, Research Opportunities and Application
Conclusion

Fig. 1.1: Organization of the Paper

task for the big data analytics researcher. To fulfill the challenge one of the ways - Topic modeling, from health care data, may be useful.

2.2. Topic Modeling. Topic modeling is a very popular topic analysis technique which follows an unsupervised machine learning approach that scans documents to detect words and phrases patterns within them and produces the clusters of words and expressions based on their similarity. Topic modeling defines a document like a probability distribution over topics and each topic as the probability distribution over the words [23]. Topic modeling is applied in any field including software engineering, crime, medical, geographical, political, and linguistic sciences [21]. Recent examples of topic modeling in healthcare include extracting knowledge from electronic health records [3] [39] and analyzing user comments and online reviews [16]. It is applied by the authors Kathiria et. el in their research work for finding out the recent trends of topics from the abstract of the research papers[26].

The two primary approaches for topic analysis are supervised learning and unsupervised learning. Supervised learning approach contains labeled datasets and the method helps in disclosing the hidden structure from the dataset. Various Word embedding models Doc2Vec, Word2Vec, Glove, BERT can help in it [27]. Unsupervised learning approach has unlabeled data and it works in discovering the pattern from it to find some insights. The widely used supervised learning technique - Classification works to train the corpus with already available labeled dataset and based on that classify a new document accordingly [33]. The other widely used unsupervised technique - clustering assigns each document of a corpus to a respective cluster as per the similarity between the documents.

Figure 2.1 relfects the steps used for topic modeling process. First, the data needs to be collected for which the topic needs to be extracted. Since the data can be in the form of unstructured textual form it needs pre-processing. The pre-processing includes tokenization, removing stop words, words being lemmatized and stemmed. Once all the unwanted words and characters are filtered out of the data it is ready for topic modeling. Then the appropriate topic model algorithm is used. The results are then visualized by suitable means.

2.2.1. Well known Topic Models used in Literature.

LSA: Latent Semantic Analysis (also known as Latent Semantic Index, LSI) is one of the popular techniques for topic modeling. LSA makes use of the bag-of-words model which helps in generating a term-document matrix that demonstrates the occurrence of terms in the document [49]. LSA finds out latent topics by carrying out matrix decomposition over the term-document matrix using Singular Value Decomposition. In short, LSA acts as the dimension reduction approach.

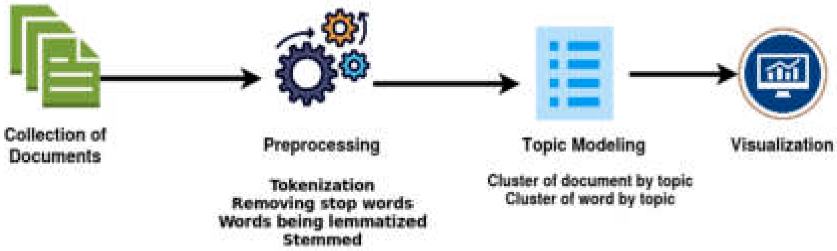


Fig. 2.1: Topic Modeling

PLSA: Probabilistic Latent Semantic Analysis is the probabilistic technique used to model the data to find the topics by making sense of the context of the text data. It evolved from Latent Semantic Analysis where the topics are hidden variables. It is used in applications that involve natural language processing, information retrieval, and filtering, applying machine learning on textual data, and in other related areas. The traditional Latent Semantic Technique makes use of linear algebra and executes Singular Value Decomposition of co-occurrence table while PLSA method makes use of the latent class model to derive mixture decomposition which includes strong statistical concepts [17].

LDA: Blei, Ng, and Jordan in 2003 [38] had proposed an unsupervised generative probabilistic model - Latent Dirichlet Allocation (LDA) which is opted to calculate the similarity between the given text files, moreover achieving their respective distributions of each document over topics. LDA is based on a three-level hierarchical Bayesian model. LDA follows a basic notion in which the documents correspond to random mixtures over latent topics, where a topic is set apart by distribution over words.

LDA can be said as a distinguished tool for latent topic distribution for a sizable corpus. Due to this, it inhibits the ability to recognize sub-topics for a technology area compiled of many patents, representing each of those patents in an array of topic distribution. Using LDA a vocabulary is generated which is then applied to discover hidden topics. There are several methods given to estimate LDA parameters, such as Gibbs sampling, variational method, and expectation propagation.

In health and medical science, LDA also serves in a variety of applications like the use of the knowledge obtained from the literature to predict protein-protein relationships [5], dig up relevant medical concepts and structures from health records of the patients [3], detecting patterns of medical events in a group of brain cancer patients [2], etc. LDA denotes more precise meaningful words as compared to LSA that's why it provides better accuracy and results [50].

CTM: Correlated Topic Model is the extension of LDA which on further evolution can be useful for creating more advanced topic models. Although LDA is the most popular topic modeling method, it has some limitations. LDA is not able to correlate the topics because it uses the Dirichlet distribution to model the unevenness amid the topic proportions. CTM makes use of logistic normal distribution to demonstrate correlation in the topic proportions 1[8].

NMF: Non-negative matrix factorization (also known as non-negative matrix approximation) represents a set of algorithms in linear algebra and multivariate analysis in which a matrix X is decomposed into two matrices W and H making sure that these are non-negative values. The problem cannot give exact values in which case approximation of numerical values is done. The application of non-negative matrix factorization can be found in fields such as document clustering, computer vision, audio signal processing, astronomy, recommender systems, missing data imputations, and other similar areas. In Table 2.1 the majorly used topic models with their advantages and disadvantages are discussed. As far as healthcare and medical topic analysis are concerned, from Table 2.1 , LDA is performing better than most of the topic modeling methods.

3. Usage of Various Forms of Data in Healthcare Domain Applied for Topic Analysis. Various forms of data present in healthcare domain can provide insights for the enhancement of the medical field in terms of diagnosis, treatment, administration, and providing other healthcare facilities. The dataset can include patents, scientific papers or articles, textual feedback, chat logs from the chat groups, abstracts of the scientific

Table 2.1: Various topic modeling techniques with pros and cons

Models	Advantages	Limitations
Latent Semantic Analysis (LSA)	- Using a single value decomposition for reducing the dimensionality of tf-idf - Statistical background is not robust - Helps in extracting synonyms of words	-Estimating the number of topics is difficult - In some cases, labeling a topic seems difficult using the words in the topic
Probabilistic Latent Semantic Analysis (PLSA)	- To some extent PLSA handles polysemy - Each word is generated from the single topic; - Different words can be generated from different topics in a document	- At the level of documents there is no probabilistic model
Latent Dirichlet Allocation (LDA)	- Can manifest nouns and adjectives in topics - Long-length documents can be handled - Provides complete generative model including distribution (i.e., multinomial) for words in dirichlet distribution over topics and other topics	-Not able to model relations amid topics
Correlated Topics Model (CTM)	-Logistic normal distribution is used for relations among opics -Helps in forming topic graphs -In other topics the appearance of the word is allowed	- Occurrences of general words in topics is allowed - Requires complex computations
Non-negative matrix Factorization (NMF)	- The use of positive values turns out easier inspection of resulting matrices	- Since it has a constraint of positive values, it can lose more information when truncating.

papers, medical reports of the patients, and social media posts.

Some of the recent papers were referred to understand what kinds of datasets are being used which is reflected in Table 3.1. From table 3.1, it can be observed that the use of social media data is more frequent in recent years. Social media data can include posts, comments, and messages. The main purpose is to explore the data and find the topics using topic modeling to understand the data correctly and find some trends, patterns, or directions for better research in the future [26]. The aim of each research can be unique and their social media data may also be unique based on its aim. But the method of solving the problem involves the exploratory analysis using the concepts of topic modeling here. One such research paper's dataset Dreaddit [20] in the next subsection is considered for the Categorical Stress Analysis on Dreaddit Dataset.

3.1. Categorical Stress Analysis on Dreaddit Dataset. Stress is omnipresent in one form or another. There can be various reasons for a person to feel stressed; it can be due to home lessness, relationship problem, domestic violence, and many others. Many surveys are conducted every year to know the actual cause of stress and provide assistance to the ones who are the victim of it. People commit illegal actions and can also lead to harming themselves if are not helped at the right time. Stress does not only affect the life of the person suffering from it, but it also has a significant impact on the people around the victim. This matter is also observable in social media as most people write multiple posts each day in which when closely observed; one can notice that the particular person is suffering from stress and even know the specific category of stress. This observation can be done with the help of topic analysis. Topic Analysis is a technique in which topics to text data automatically. Analyzation of unstructured text is done using Topic Analysis; such as social media interactions and emails.

A great study regarding stress is shown in paper [20] where the authors have presented Dreaddit; a large social media data from multiple domains for identifying stress in people. This dataset contains 1,90,000 posts which are collected from five different communities on Reddit. They additionally labelled 3500 segments (total) which were taken from 3000 posts with the help of Amazon Mechanical Turk.

Table 3.1: Types of data used in recent studies (X indictes the presence of the data type)

Author	Type of Data used						Purpose
	Scientific paper, abstracts, patents	Textual feedback	Chat logs	Patients clinical records	Social media posts	News articles	
[13]	x						Forecasting commercial viability or sustainability of healthcare innovations
[23]	x	x			x	x	Proposing a new method in topic modeling
[12]	x						Build an understanding for future reference
[47]	x						Analysis of use of “personality” and “mental health”
[19]		x					Insights for improvement of quality healthcare by analysis the patients reviews for their physicians
[48]			x				Insights for improving healthcare for pregnant women based on their social opinions
[45],[40]			x		x		Analyzing the feedbacks and patient opinions for improvement
[9]			x				Mental Health insights
[18],[44]				x			Predict clinical risk of a patient
[32]				x			Predict depression prior to clinical diagnosis
[3]				x			Case-based information retrieval of similar patients using patient’s clinical records
[11]					x	x	Insights for improvement of the public health communication strategies
[4], [11], [46], [31], [14] [22], [37]					x		Discover and understand trends
[42],[42]					x		Clustering using topic models
[30]						x	Insights for improvement of health inequality issues in Korea
[10]					x		Analysis of use of social media in healthcare research

On the same dataset using the four specific features namely: confidence, anxious, angry, sad; which show the feelings of the person who is the victim of stress, stress analysis is done. Figure 3.1 shows a bar graph which depicts the four features and their significance on different categories of stress such as ptsd, assistance, relationships, survivorsofabuse, domestic violence, anxiety, homeless, stress, almosthomeless, and food pantry. For instance, people are equally angry, anxious, and sad at the same time due to relationship problems. Also, due to anxiety a person can feel stressed as shown in the graph wherein anxiety, the anxiousness of a person increases on a large scale which can lead to stress.

There is another type of data that is most commonly used in recent studies is the scientific papers, their abstracts, patents, books, or grey literature. There is already an ample amount of research papers published. It

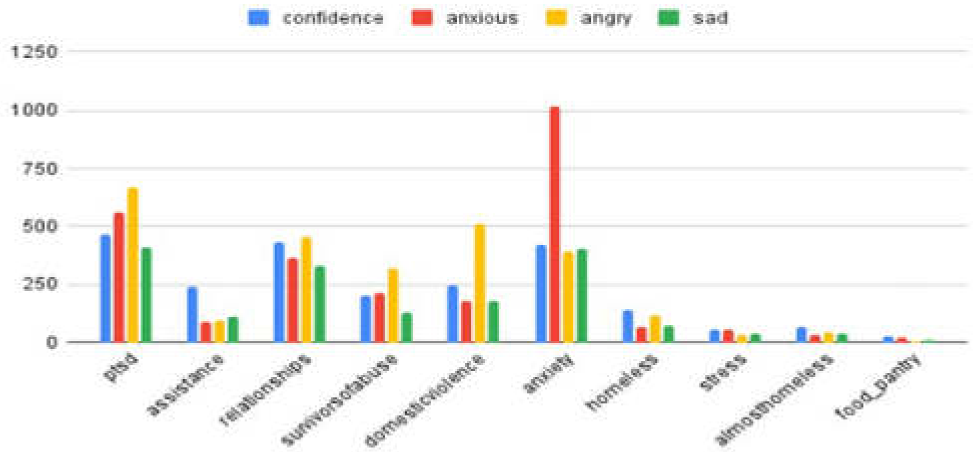


Fig. 3.1: Categorical Stress Analysis on Dreaddit Dataset

is difficult to read them all and infer knowledge from them which requires both time and effort. Topic modeling can help in analyzing the data and finding the relevant information in less time.

4. Advancement in Topic Analysis Methods for Healthcare. LDA is one of the well known topic modeling approaches among all. Apart from LDA, some research trends in topic modeling to follow and a new proposed improved framework for topic modeling are given in Table 4.1.

A study on understanding the different health topics related to social media is done by the authors' Paul et. al. [30]. A new statistical topic model was proposed by them referred to as Ailment Topic Aspect Model (abbr. ATAM). Since LDA is able to discover not only topics related to health around ailments but also other frequent topics which are not relevant as ailments and can be symptoms or something else. This noise needs to be filtered out to have just the health-related topics. ATAM was developed to explicitly label each tweet according to its ailment category; the model can also incorporate treatment and symptoms information. It is based on LDA where the ailment model will have three separate word distributions for symptoms, treatment, and general words. Thus it has a structurally different distribution.

Another study describes that the common statistical topic analysis approaches are not practical for rapidly processing the ever increasing online data. They proposed an alternative approach of automatic topic detection on the basis of document clustering for the extraction of topics related to health from online communities [30, 24]. The use of Expectation Maximization (EM) Clustering is applied in the same work. EM Clustering is a category of probability clustering method that allocates each occurrence with a probability which denotes that they belong to each cluster.

Apart from LDA, there are other techniques like correlated, dynamic, and hierarchical topic models. In recent years, another technique that is gaining popularity is Structural Topic Modeling (STM) [14]. STM encompasses metadata related to the text such as where, when, by whom the text was written, etc. For the estimation of Correlations, the Dirichlet distribution in the standard LDA can be replaced with the logistic normal distribution as it is in the Correlation Topic Models [8].

Likewise, there are different requirements to solve different problems at hand, and to do so we might need to boost the power of LDA accordingly. Thus, there are different variants possible and researchers tend to manage to improve it by combining it with other methods to form a hybrid topic model [42], to make it more effective and as per the requirements.

Table 4.1: New proposed improved framework for topic modeling

Paper	Year	Innovative Model	Objective
[6]	2023	The study examines deep learning (DL) and machine learning (ML) methods for healthcare prediction.	The paper assesses predictive analytics in healthcare, focusing on ML and DL techniques, emphasizing accurate disease prediction.
[15]	2022	Comparative Analysis and Classification of Topic Modeling Algorithms.	This study aims to provide a comprehensive overview of topic modeling in healthcare, including a comparative analysis of algorithms and their applications.
[32]	2020	Hierarchical Clinical Embeddings with Topic Modeling	Predicting depression
[46]	2020	Structural Topic Model	Study the contribution of social bots in the COVID-19 discussions on twitter
[31]	2020	Clustering Newman Algorithm	Social media based health disparity for COVID-19
[42]	2019	Visual Non-negative Matrix Factorization (VNMF), Visual Latent Dirichlet Allocation (VLDA), Visual Probabilistic Latent Schematic Indexing (VPLSI), Visual Latent Schematic Indexing (VLSI)	Using Hybrid Topics models by integrating topic models with VAT, for visualizing the health tendency and the topic clouds in the document collection.
[14]	2019	Structural Topic Model	To analyze tweets of stroke Survivors; their reactions based on their gender
[23]	2018	Fuzzy Latent Semantic Analysis (FLSA)	A better topic modeling approach compared to LDA is proposed in medical domain
[29]	2016	Application and Development of Topic Models in Bioinformatics	This paper reviews bioinformatics applications of topic models, categorizes studies, and underscores the need for tailored models to optimize biological data interpretation.
[48]	2016	Topic Interest Model	Use online healthcare chat logs to extract topics and infer user interests.
[18]	2015	Probabilistic Risk Stratification Model (PRSM)	Predict patients clinical risk to strategize the treatment accordingly
[41]	2014	Ailment Topic Aspect Model (ATAM)	Obtaining topics related to health from the tweets including its symptoms and treatment
[44]	2013	Co-occurrence Based Clustering, Dirichlet Process Mixture Model	To estimate patient disease risk

5. Challenges, Research opportunities, and Application. Various challenges, research opportunities and application of topic modeling techniques are discussed in this section.

5.1. Challenges. Figure 5.1 describes that there are many challenges in healthcare related to topic modelling. These can be categorized as technical, social and authenticity related challenges. Technical challenges involve security-related and topic modeling related challenges. Security-related challenges include problems related to advanced encryption and authentication. While, topic modeling-related challenges include problems related to used topic modeling approaches. Furthermore, Social challenges involve challenges related to mental insecurity and privacy of patient records. Moreover, authenticity related challenges include medical data related difficulties which consists of gathering labeled data and social media challenges which can be further questioned for authenticity. The limitation of the popular topic modeling approach LDA involves the difficulty in recognizing the numerical value of topics. Therefore, the general judgment of researchers can be used to find out the numerical value of topics.

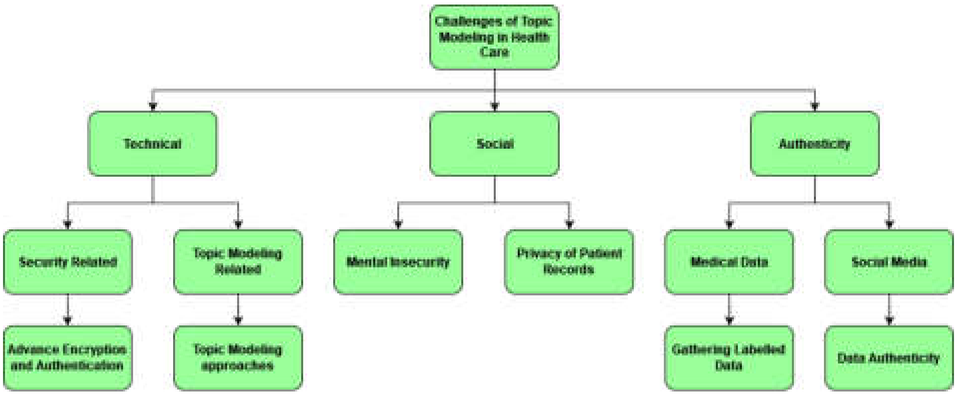


Fig. 5.1: Challenges of Topic Modelling in Healthcare

The topic analysis in healthcare often requires patient records. The protection of a patient's privacy and confidentiality is critical as well as challenging. The violations of privacy can not only hurt the patients but also the reputation of the healthcare firms. These violations can result in legal battles along with eroding the patient's trust and affect the long-term viability of healthcare firms.

5.2. Research Opportunities and Application. Research opportunities and Applications in this area of research are as follows.

The visualization of the health tendency can be found based on past data which can help for better planning of nutritious food for better health. The topic modeling applied to scientific research articles can give insight into the advancement of medical diagnosis, tools, and technologies.

The vast amount of data available on social networks is a significant resource for analyzing the pandemic's diverse impact on society. Using approaches such as topic modeling, this data can be evaluated to identify common themes, feelings, and worries voiced by people from various demographics and geographic areas. Armed with these insights, big organizations and government agencies can create educated policies and focused awareness guidelines to meet unique social requirements and issues caused by the epidemic.

Furthermore, combining patient health records and social network data allows healthcare practitioners to provide tailored medical interventions, modifying treatment programs, and prescribing medications based on specific patient needs and circumstances. This comprehensive strategy not only improves patient outcomes but also optimizes resource use within the healthcare system ([36]), emphasizing the importance of social network data analysis and patient health records in addressing the pandemic's impact at both the societal and individual levels.

The enormous data on social media allows for the prediction of the relationship between physical health and socio-cultural factors, revealing how cultural norms and online interactions influence behaviors and outcomes. By examining this data, researchers can develop personalized interventions and policies to alleviate regional health disparities. Initiatives such as ”Healthy People 2030” in the United States demonstrate how data-driven tactics may generate substantial change and promote health equity across varied geographic locations.

6. Conclusion. Digitalization has offered us with various unstructured health-related data, which contains information and knowledge of great potential. Topic Modeling from this unstructured health-related data can easily identify the hidden pattern which can prove to be helpful in improving the treatment of the patients and the healthcare facility. In this paper, we surveyed various research papers and addressed the work of those researchers. We also did an analysis on Dreaddit dataset, categorizing the type of cause which can lead to stress and presented the way, in which topic modeling can be used to benefit the health sector. Many topic modeling techniques such as LDA, ATAM, FLSA, CTM, NMF, and LSA have been discussed throughout the paper and it was observed that LDA is superior to all the techniques discussed in the paper as it is easily able to manifest nouns and adjectives in a topic and its capability of handling very large documents. Also, we then discussed the challenges researchers can face in the healthcare sector while using topic modeling techniques. Moreover, future research opportunities and applicability of topic modeling techniques in sectors other than healthcare, such as bioinformatics, IT (Predication and Recommendation system), and Social Network Analysis have been discussed, which can help the researchers in understanding the requirement of topic modeling and the areas where it can be used. The main aim of conducting this research was to understand the work done in the aspect of topic analysis in the healthcare sector and use it to form a conclusion, to improvise the strategy to deliver a better treatment and healthcare facility.

REFERENCES

[1] A. Abbas, L. Zhang, and S. U. Khan, *A literature review on the state-of-the-art in patent analysis*, World Patent Information, 37 (2014), pp. 3–13.

[2] C. Arnold and W. Speier, *A topic model of clinical reports*, in Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 2012, pp. 1031–1032.

[3] C. W. Arnold, S. M. El-Saden, A. A. Bui, and R. Taira, *Clinical case-based retrieval using latent topic analysis*, in AMIA annual symposium proceedings, vol. 2010, American Medical Informatics Association, 2010, p. 26.

[4] M. Asghari, D. Sierra-Sosa, and A. Elmaghraby, *Trends on health in social media: Analysis using twitter topic modeling*, in 2018 IEEE international symposium on signal processing and information technology (ISSPIT), IEEE, 2018, pp. 558–563.

[5] T. Asou and K. Eguchi, *Predicting protein-protein relationships from literature using collapsed variational latent dirichlet allocation*, in Proceedings of the 2nd international workshop on Data and text mining in bioinformatics, 2008, pp. 77–80.

[6] M. Badawy, N. Ramadan, and H. A. Hefny, *Healthcare predictive analytics using machine learning and deep learning techniques: a survey*, Journal of Electrical Systems and Information Technology, 10 (2023), p. 40.

[7] D. Blei, L. Carin, and D. Dunson, *Probabilistic topic models*, IEEE signal processing magazine, 27 (2010), pp. 55–65.

[8] D. M. Blei and J. D. Lafferty, *A correlated topic model of science*, The annals of applied statistics, 1 (2007), pp. 17–35.

[9] B. Carron-Arthur, J. Reynolds, K. Bennett, A. Bennett, and K. M. Griffiths, *What's all the talk about? topic modelling in a mental health internet support group*, BMC psychiatry, 16 (2016), pp. 1–12.

[10] X. Chen, Y. Lun, J. Yan, T. Hao, and H. Weng, *Discovering thematic change and evolution of utilizing social media for healthcare research*, BMC Medical Informatics and Decision Making, 19 (2019), pp. 39–53.

[11] W. Chipidza, E. Akbaripourdibazar, T. Gwanzura, and N. M. Gatto, *Topic analysis of traditional and social media news coverage of the early covid-19 pandemic and implications for public health communication*, Disaster medicine and public health preparedness, (2021), pp. 1–8.

[12] R. Dantu, I. Dissanayake, and S. Nerur, *Exploratory analysis of internet of things (iot) in healthcare: A topic modeling approach*, (2019).

[13] S. S. Erzurumlu and D. Pachamanova, *Topic modeling and technology forecasting for assessing the commercial viability of healthcare innovations*, Technological Forecasting and Social Change, 156 (2020), p. 120041.

[14] A. Garcia-Rudolph, S. Laxe, J. Saurí, M. B. Guitart, et al., *Stroke survivors on twitter: sentiment and topic analysis from a gender perspective*, Journal of medical Internet research, 21 (2019), p. e14077.

[15] A. Gupta and H. Fatima, *Topic modeling in healthcare: A survey study*, NEUROQUANTOLOGY, 20 (2022), pp. 6214–6221.

[16] H. Hao, K. Zhang, et al., *The voice of chinese health consumers: a text mining approach to web-based physician reviews*, Journal of medical Internet research, 18 (2016), p. e4430.

[17] T. Hofmann, *Probabilistic latent semantic analysis*, arXiv preprint arXiv:1301.6705, (2013).

[18] Z. Huang, W. Dong, and H. Duan, *A probabilistic topic model for clinical risk stratification from electronic health records*, Journal of Biomedical Informatics, 58 (2015), pp. 28–36.

[19] T. L. James, E. D. V. Calderon, and D. F. Cook, *Exploring patient perceptions of healthcare service quality through analysis of unstructured feedback*, Expert Systems with Applications, 71 (2017), pp. 479–492.

[20] L. JD, *A correlated topic model of science. the annals of applied statistics 2007*, 17 (2007), pp. 17–35.

[21] H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao, *Latent dirichlet allocation (lda) and topic modeling: Models, applications, a survey. arxiv*, arXiv preprint arXiv:1711.04305, (2017).

[22] I. Kagashe, Z. Yan, I. Suheryani, et al., *Enhancing seasonal influenza surveillance: topic analysis of widely used medicinal drugs using twitter data*, Journal of medical Internet research, 19 (2017), p. e7393.

[23] A. Karami, A. Gangopadhyay, B. Zhou, and H. Kharrazi, *Fuzzy approach topic discovery in health and medical corpora*, International Journal of Fuzzy Systems, 20 (2018), pp. 1334–1345.

[24] P. Kathiria and H. Arolkar, *Study of different document representation models for finding phrase-based similarity*, in Information and Communication Technology for Intelligent Systems, Springer, 2019, pp. 455–464.

[25] P. Kathiria and H. Arolkar, *Document clustering based on phrase and single term similarity using neo4j*, International Journal of Innovative Technology and Exploring Engineering (IJITEE), 9 (2020), pp. 3188–3192.

[26] P. Kathiria and H. Arolkar, *Trend analysis and forecasting of publication activities by indian computer science researchers during the period of 2010–23*, Expert Systems, 39 (2022), p. e13070.

[27] P. Kathiria, U. Patel, and N. Kansara, *Document classification using deep neural network with different word embedding techniques*, International Journal of Web Engineering and Technology, 17 (2022), pp. 203–222.

[28] P. Kathiria, U. Patel, S. Madhwani, and C. S. Mansuri, *Smart crop recommendation system: A machine learning*

approach for precision agriculture, in Machine Intelligence Techniques for Data Analysis and Signal Processing, D. S. Sisodia, L. Garg, R. B. Pachori, and M. Tanveer, eds., Singapore, 2023, Springer Nature Singapore, pp. 841–850.

[29] L. Liu, L. Tang, W. Dong, S. Yao, and W. Zhou, *An overview of topic modeling and its current applications in bioinformatics*, SpringerPlus, 5 (2016), pp. 1–22.

[30] Y. Lu, P. Zhang, J. Liu, J. Li, and S. Deng, *Health-related hot topic detection in online communities using text clustering*, Plos one, 8 (2013), p. e56221.

[31] J. Mantas et al., *Application of topic modeling to tweets as the foundation for health disparity research for covid-19*, The Importance of health Informatics in Public Health during a Pandemic, 272 (2020), p. 24.

[32] Y. Meng, W. Speier, M. Ong, and C. W. Arnold, *Hcet: Hierarchical clinical embedding with topic modeling on electronic health records for predicting future depression*, IEEE Journal of Biomedical and Health Informatics, 25 (2020), pp. 1265–1272.

[33] T. M. Mitchell and T. M. Mitchell, *Machine learning*, vol. 1, McGraw-hill New York, 1997.

[34] B. Mohanta, P. Das, and S. Patnaik, *Healthcare 5.0: A paradigm shift in digital healthcare system using artificial intelligence, iot and 5g communication*, in 2019 International Conference on Applied Machine Learning (ICAML), Los Alamitos, CA, USA, may 2019, IEEE Computer Society, pp. 191–196.

[35] E. National Academies of Sciences, Medicine, et al., *Future directions for NSF advanced computing infrastructure to support US science and engineering in 2017-2020*, National Academies Press, 2016.

[36] S. Neely, C. Eldredge, and R. Sanders, *Health information seeking behaviors on social media during the covid-19 pandemic among american social networking site users: Survey study*, J Med Internet Res, 23 (2021), p. e29802.

[37] M. D. T. Nzali, S. Bringay, C. Lavergne, C. Mollevi, and T. Opitz, *What patients can tell us: topic analysis for social media on breast cancer*, JMIR medical informatics, 5 (2017), p. e7779.

[38] K. Odongo, *Uncovering consumer preferences for a novel apple variety using latent dirichlet allocation*, (2022).

[39] P. C.-I. Pang and S. Chang, *The twitter adventure of# myhealthrecord: an analysis of different user groups during the opt-out period*, Studies in Health Technology and Informatics, 266 (2019), pp. 142–148.

[40] P. C.-I. Pang and L. Liu, *Why do consumers review doctors online? topic modeling analysis of positive and negative reviews on an online health community in china*, (2020).

[41] M. J. Paul and M. Dredze, *Discovering health topics in social media using topic models*, PloS one, 9 (2014), p. e103408.

[42] K. R. Prasad, M. Mohammed, and R. Noorullah, *Hybrid topic cluster models for social healthcare data*, International Journal of Advanced Computer Science and Applications, 10 (2019).

[43] P. Purohit, P. Khanpara, U. Patel, and P. Kathiria, *Iot based ambient assisted living technologies for healthcare: Concepts and design challenges*, in 2022 Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), IEEE, 2022, pp. 111–116.

[44] A. K. Rider and N. V. Chawla, *An ensemble topic model for sharing healthcare data and predicting disease risk*, in Proceedings of the international conference on bioinformatics, computational biology and biomedical informatics, 2013, pp. 333–340.

[45] P. Sampath, G. Packiriswamy, N. Pradeep Kumar, V. Shanmuganathan, O.-Y. Song, U. Tariq, and R. Nawaz, *Iot based health—related topic recognition from emerging online health community (med help) using machine learning technique*, Electronics, 9 (2020), p. 1469.

[46] W. Shi, D. Liu, J. Yang, J. Zhang, S. Wen, and J. Su, *Social bots' sentiment engagement in health emergencies: A topic-based analysis of the covid-19 pandemic discussions on twitter*, International Journal of Environmental Research and Public Health, 17 (2020), p. 8701.

[47] R. Sperandeo, G. Messina, D. Iennaco, F. Sessa, V. Russo, R. Polito, V. Monda, M. Monda, A. Messina, L. L. Mosca, et al., *What does personality mean in the context of mental health? a topic modeling approach based on abstracts published in pubmed over the last 5 years*, Frontiers in psychiatry, 10 (2020), p. 938.

[48] T. Wang, Z. Huang, and C. Gan, *On mining latent topics from healthcare chat logs*, Journal of biomedical informatics, 61 (2016), pp. 247–259.

[49] C. Wenli, *Application research on latent semantic analysis for information retrieval*, in 2016 Eighth International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), 2016, pp. 118–121.

[50] S. Yang, G. Huang, and B. Cai, *Discovering topic representative terms for short text clustering*, IEEE Access, 7 (2019), pp. 92037–92047.

Edited by: Chiranji Lal Chowdhary
Special issue on: Scalable Machine Learning for Health Care: Innovations and Applications
Received: Jun 22, 2023
Accepted: Mar 14, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1361–1372, DOI 10.12694/scpe.v25i3.2392

CONSTRUCTION OF AN INTELLIGENT IDENTIFICATION MODEL FOR DRUGS IN NEAR INFRARED SPECTROSCOPY AND RESEARCH ON DROG CLASSIFICATION BASED ON IMPROVED DEEP ALGORITHM

JIULIN XIA[*]

Abstract. Near-infrared spectroscopy has important applications in drug and food identification. Combining machine learning with near-infrared spectroscopy to achieve intelligent identification of drugs has become a research hotspot in recent years. To solve the problem of machine learning's inefficiency in classifying small-scale data, a drug identification model based on near-infrared spectroscopy combined with a random fading depth belief network is proposed. Aiming at the problem that the training time of the machine learning algorithm is too long, the extreme learning machine is used to replace the back propagation algorithm to optimize the stack sparse auto-encoder network. Additionally, the stack sparse auto-encoder algorithm based on extreme learning machine algorithm is constructed. The study found that the precision of the Dropout Deep Belief Network model was 99.12%, which was higher than the other three models. Additionally, the area under the curve value of the Dropout Deep Belief Network model was 0.87, which was 0.04 higher than the binary whale optimization algorithm model, 0.26 higher than the factor decomposition machine and depth neural network model, and 0.05 higher than the random forest network model. The sparse auto-encoder algorithm based on the extreme learning machine algorithm achieved a precision of 99.72%. The study proposes two algorithm models that can effectively identify drugs using near-infrared spectroscopy. This has a positive impact on the medical industry and the safety of patients' lives and health.

Key words: Machine learning; Near-infrared spectroscopy; Deep belief network; Extreme learning machine; Dropout-DBN; ELM-SAE; Drug identifications

1. Introduction. Counterfeit and substandard drugs not only cause huge economic losses, but also cause harm to human body and endanger the health and safety of patients. Therefore, the identification of fake and substandard drugs has always been the focus of scholars from all walks of life [1, 2]. Near-infrared Spectroscopy (NIRS) can detect samples efficiently and accurately without damaging the samples to be tested, so it is used in the testing of pharmaceutical, food, chemical and other industries [3, 4]. However, current drug identification technology is mostly limited to identifying genuine and fake drugs in the second classification. Its effectiveness in identifying multiple varieties of drugs is poor. To solve this problem, a drug identification model based on Dropout Deep Belief Network (DBN) and NIRS is proposed. In addition, aiming at the problem that the training time of machine learning algorithm is too long, a stack Sparse Auto-encoder Algorithm based on Extreme Learning Machine (ELM-SAE) is proposed. The paper has realized the high-precision and efficient identification of drugs, which has a certain role in promoting the development of China's medical industry.

2. Review. The application achievements of NIRS in various fields have gradually matured, and it is also well applied in drug identification. This part lists the application of NIRS technology in drug recognition and explores the latest progress of drug classification models based on deep learning algorithms. The relevant study contents are presented in Table 2.1. In summary, near-infrared spectroscopy technology provides strong support for intelligent drug recognition. When combined with deep learning algorithms, drugs can be effectively classified. This paper reviews relevant literature and elaborates on the important application of NIRS technology in drug recognition. It emphasizes the value of drug classification research based on improved deep learning algorithms. Future research can optimize models to improve recognition accuracy, providing powerful tools for drug quality control and regulation.

[*] Basic Teaching Department, Chongqing College of Science and Creation, Chongqing, 402160, China (jiulin_xia@outlook.com)

Table 2.1: Literature review

Literature	Author	Time	Time method	conclusion
The importance of wavelength selection in on-scene identification of drugs of abuse with portable NIRS [6].	R. F. Kranenburg, Y. Weesepoel, M. Alewijn, S. Sap, W. F. Peter, A. V. Esch	2022	Use of NIRS for the identification of drugs of abuse	Through visual inspection of the near-infrared spectra, the results yielded new insights into the usability of individual spectrometers
Vibrational spectroscopy in analysis of pharmaceuticals : Critical review of innovative portable and handheld NIR and Raman spectrophotometers [7].	R. Deidda, P. Y. Sacre, M. Clavaud, L. Coic, H. Avohou, P. Hubert, E. Ziemons	2019	Study of portable/handheld near infrared and Raman spectrophotometers in drug identification	Infrared spectroscopy and Raman spectroscopy had consistent advantages in drug analysis, demonstrating enormous potential
Multi-manufacturer drug identification based on near infrared spectroscopy and deep transfer learning [8].	L. Li, X. Pan, W. Chen, M. Wei, Y. C. Feng, L. H. Yin, C. Q. Hu, H. H. Yang	2020	Drug recognition using NIRS combined with deep learning algorithms	CNN-based drug recognition method achieved higher classification accuracy and scalability in multi-species and multi-manufacturer near-infrared spectral classification experiments

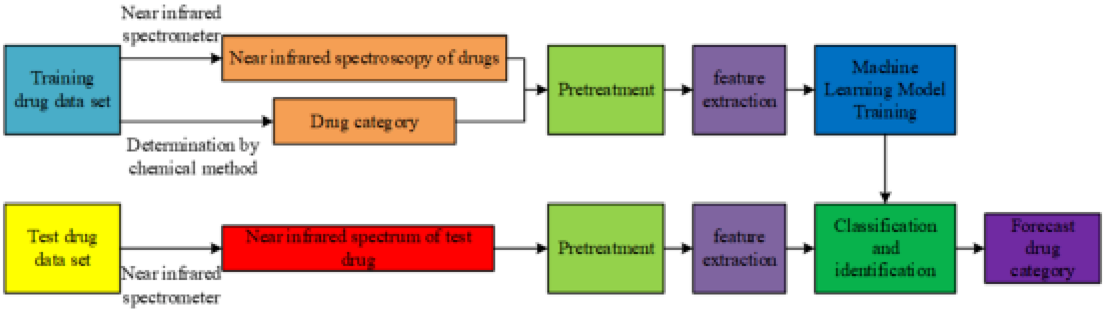
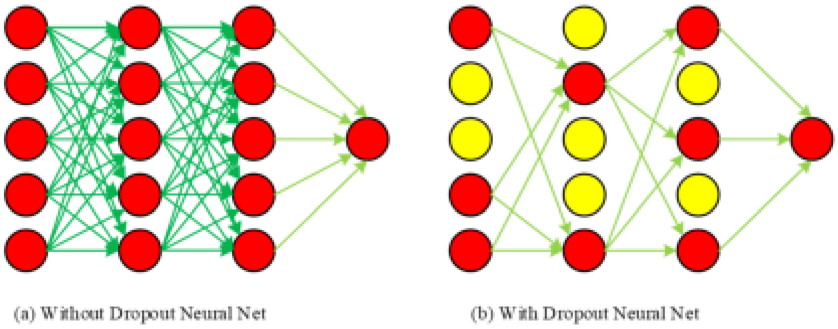

Near infrared spectrometer
Training drug data set
Near infrared spectroscopy of drugs
Drug category
Determination by chemical method
Pretreatment
feature extraction
Machine Learning Model Training
Test drug data set
Near infrared spectrum of test drug
Near infrared spectrometer
Pretreatment
feature extraction
Classification and identification
Forecast drug category

Fig. 3.1: Drug identification process combining NIRS technology and machine learning algorithm

3. Construction of Drug Identification Model Combined with NIRS.

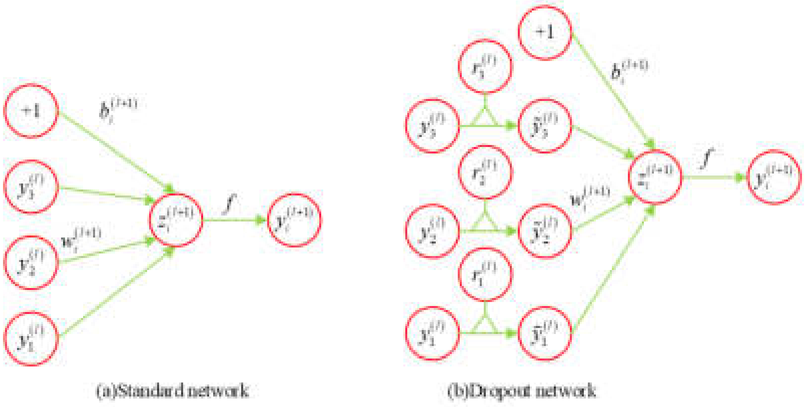
3.1. Construction of Dropout DBN Model. Counterfeit and substandard drugs will not only cause economic losses to patients, but also seriously threaten the life and health safety of patients. Therefore, drug identification has received extensive attention. To avoid the drawbacks of low efficiency, high cost, and potential errors in traditional manual identification methods, many researchers have combined NIRS technology with machine learning algorithms to achieve intelligent drug identification in large quantities with high precision [9, 10]. The drug identification process combining NIRS technology and machine learning algorithm is shown in Figure 3.1.

In Figure 3.1, there are five main steps, namely: NIRS sample data collection, sample data preprocessing, sample data feature extraction, machine learning model construction, and identification, classification and prediction of drugs to be tested. The extraction of sample data features is one of the most important links, which relates to the precision of drug identification models. At present, feature extraction of sample data mainly depends on various algorithms, such as binary whale optimization algorithm (BWOA) [11], random forest network model (RFNM) [12], factor decomposition machine and depth neural network (FM-DNN) [13]. However, the drug identification precision and speed of these algorithms are not ideal, and they need to be further improved [14, 15]. Because of deep learning deep network structure and nonlinear activation function, deep learning model has an excellent performance in high-dimensional, nonlinear big data modeling. Therefore, the application effect of the DBN in drug identification is discussed. However, the DBN model is more suitable for training large-scale data. It is less effective for sample sets with fewer sample numbers and data feature dimensions, and is prone to over-fitting [16]. In drug identification, due to the complexity and difficulty of sample collection and chemical analysis, the number of samples is often small. Therefore, the Dropout technology is introduced into the DBN model. Dropout technology means that during the training process, part of the neural nodes in the hidden layer of the deep learning network and their weights are temporarily removed from the network randomly, as shown in Figure 3.2.



(a) Without Dropout Neural Net (b) With Dropout Neural Net

Fig. 3.2: Application effect of Dropout technology



(a)Standard network (b)Dropout network

Fig. 3.3: Common network and Dropout network

Through the process shown in Figure 3.2, the Dropout technology can avoid the synergism of some similar features, thus avoiding the decline in model precision caused by over fitting. During training, Dropout technology randomly selects a subnetwork from the learning network containing N nodes to generate 2n subnetwork sets. The weight of the sub network set is shared, so its time complexity can still be considered as O(2n). For 2n sub network set, only partial training is conducted. Supposing that the hidden layer of a deep learning network has an L layer, we have the index number of the hidden layer, the input vector of layer, the output vector of layer, and the connection weight and offset value of layer, respectively. At this time, the forward network can be described as Formula 3.1.

$$y_i^{l+1} = f(z_i^{(l+1)}) = W_i^{(l+1)} + b_i^{(l+1)} \tag{3.1}$$

In Formula (1.1), $f()$ is the activation function. After the Dropout technology is introduced, the forward propagation neural network and its changes can be shown in Figure 3.3.

In Figure 3.3(b), $r^{(l)}$ is a vector in the hidden layer, and all elements in $r^{(l)}$ obey Bernoulli distribution, which can be expressed as Formula 3.2.

$$r_j^{(}l) = Bernoulli(p) \tag{3.2}$$

In Formula 3.2, p refers to the sampling probability of Dropout. Using $r^{(l)}$ to sample the output y of the upper layer network, a sub network $\widehat{y}$ is obtained as the input of the lower layer network, as shown in Formula 3.3

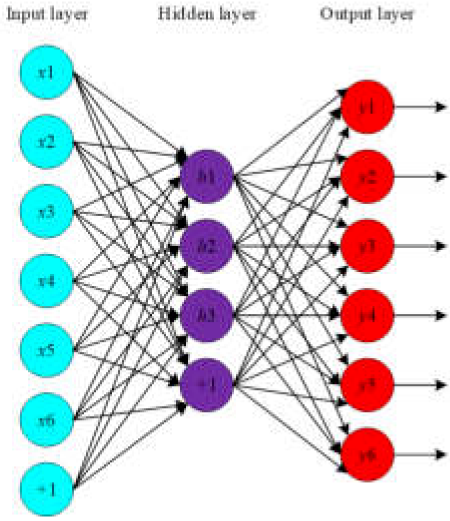


Fig. 3.4: Approximate structure of AEN

and Formula 3.4.

$$\widetilde{y}^{l} = r(l) * y^{(}l) \tag{3.3}$$

$$y_i^{l+1} = f(z_i^{(}l+1)) = W_i^{l+1}\widetilde{y}^{l} + b_i^{l+1} \tag{3.4}$$

In terms of structure, Dropout DBN is composed of multi-layer restricted Boltzmann machine (RBM) and a layer of BP neural network (BPNN), so it has excellent performance in prediction and classification of high-dimensional feature vectors [17]. The application process of Dropout DBN model in drug identification is as follows: The first is to preprocess the NIRS spectral data of drugs, and input them into the network as training data. Then to set the RBM network structure and number of hidden layers, the learning rate of RBM, and the activation function. Using the training data to train the RBM in the DBN network. After completing the training, to fine-tune the parameters through backpropagation to improve precision. In this process, Logistic classifier is used for the second classification output, and softmax classifier is used for the multi classification output.

3.2. Optimization of Auto-encoder Network. The Dropout DBN model has a good processing effect on small-scale data, but when the drug data is large, its identification accuracy is not ideal. Therefore, for large-scale data, Auto-encoder Network (AEN) network is generally used as the identification model. AEN is the basic model of deep learning algorithms [18]. AEN utilizes an artificial neural network (ANN) to construct a three-layer symmetric network with equal input and output layers. The network is then trained to minimize the error of input and reconstruction data, optimizing the connection weight and offset value of SAE to obtain the internal structural characteristics of the data [19, 20]. The general structure of AEN is shown in Figure 3.4.

The input layer of AEN is set to be x, and there are d input neural units in the input layer. Assuming that the hidden layer of AEN is y, in which there are h nerve units. The input layer of AEN is set as z, which contains the same number of nerve units as the input layer. During training, the data of the model input layer is mapped to the hidden layer, which is called coding, as shown in Formula 3.5.

$$y = f(x) = s(W_y x + b_y) \tag{3.5}$$

In Formula 3.5, $s()$ is a nonlinear mapping function, usually represented by a sigmoid function. b_y is the offset value of the hidden layer, and W_y is the weight matrix between the input layer and the hidden layer. After encoding, the data characteristics of the hidden layer are reconstructed and mapped to the output layer to complete decoding, as shown in Formula 3.6.

$$z = g(x) = s(W_z x + b_z) \tag{3.6}$$

In Formula 3.6, b_z represents the offset value of the output layer, and W_z represents the weight matrix between the output layer and the hidden layer. The eigenvalue z obtained after decoding can be approximately regarded as the eigenvalue of the input data. Therefore, in the process of decoding and reconstruction, weight binding is required to make the weight matrix $W_y = W_z = W$. Therefore, in the training of AEN model, only three groups of parameters, namely W, and b_y, need to be learned. When processing large-scale data with the AEN model, the batch random gradient descent method is typically utilized to obtain the error of small batches of data, which is then used to update the network's weight and bias. Therefore, the AEN model can be expressed as the solution of the optimization problem in Formula 3.7.

$$L(\theta) = arg\ min_{\theta} \frac{1}{n} \sum_{i=1}^{n} [x_{ik} log(z_{ik}) + (1 - x_{ik}) log(1 - z_{ik})] \tag{3.7}$$

In Formula 3.7, $L()$ is the reconstruction error function expressed by cross entropy function, $\theta = \{W, b_y, b_z\}$, that is, three groups of parameters to be learned. x_{ik} is a symbolic function. z_{ik} represents the probability that sample ik is predicted to be positive. The gradient descent method can be used to solve θ, to obtain the required parameters. However, if only the information of input data is saved, AEN model cannot learn and obtain effective feature representation. When the dimensions of the hidden layer and the input layer of the AEN model are the same, it is only necessary to learn an identity mapping to achieve zero error data reconstruction. But this kind of identity mapping does not have enough expression ability for high-level abstract representation. In drug identification, AEN model is required to learn a more complex nonlinear function. To address this issue, constraints must be added to AEN. One approach is to decrease the number of nodes in the hidden layer to reduce the dimensionality of the data. Another option is to include a penalty factor at the input of the network to filter out data noise. To solve these problems, a Sparse Auto-encoder Network (SAEN) is proposed by introducing the sparse idea into DBN. The basic idea of SAEN is that when a hidden layer node has a high probability of being activated, the node represents relatively little information. SAEN can add constraints to the activation function of hidden layer neurons, so that hidden layer neurons are in a state of inhibition for a long time, thus expressing more information. Generally speaking, the activation function of neurons adopts the sigmoid function. At this time, the average activation value of hidden layer node j in all samples can be expressed as Formula 3.8.

$$\widehat{\rho_j} = \frac{1}{m} \sum_{i=1}^{m} f_{\theta}^{j}(x^{(i)}) \tag{3.8}$$

In Formula 3.8, m is the number of training samples. In AEN, $\widehat{\rho} = \rho$ is used to limit the activation probability of hidden layer nodes. ρ is a sparsity parameter, which is generally close to 0. After the AEN is limited by the sparse parameter, the sparse penalty term should be added to the loss function of AEN. The relative entropy function, namely KL divergence algorithm, is used to add Formula 3.9 to the loss function of AEN.

$$KL(\rho || \widehat{\rho}) = \rho log \frac{\rho}{\widehat{\rho_j}} + (1 - \rho) log(\frac{1 - \rho}{1 - \widehat{\rho_j}}) \tag{3.9}$$

After adding restrictions, the self-coding cost function of AEN can be expressed by Formula 3.10.

$$L_{\rho}(\theta) = L(theta) + \beta \sum_{j=1}^{h} KL(\rho || \widehat{\rho}) \tag{3.10}$$

In Formula 3.10, β is the weight parameter of the sparse penalty. Due to the sparse penalty term, the sigmoid function causes the activation value of most hidden layer neural units to be close to 0, with only a few data

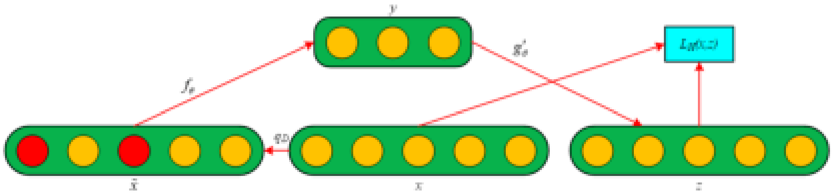


Fig. 3.5: Noise reduction AEN

points producing large activation values. In this case, AEN can learn higher and more complex abstract features. To improve the robustness of AEN, a noise reduction AEN is proposed. The basic idea is to input noisy sample data at the input layer, and then encode and decode the noisy data. In this process, the output sample data after AEN reconstruction should keep the original information of the input data as much as possible. Through the above operations, the robustness of the hidden layer can be improved to better learn high-level abstract features. If the original input vector is, after the noise is added, the input vector is represented as $\widetilde{x}$. At this time, the coding and decoding of the noise reduced AEN can be represented by Formula 3.11.

$$\begin{cases} y = f(\widetilde{x}) = s(W_y x + b_y) \\ z = g(y) = s(\widetilde{W}_z x + b_y) \end{cases} \tag{3.11}$$

In Formula 3.11, $\widetilde{W}$ is the weight after adding noise. There are generally two ways to add noise in noise reducing AEN. The first is to add Gaussian noise to the input sample data, as shown in Formula 3.12.

$$\widetilde{x} = x + \varepsilon \tag{3.12}$$

In Formula 3.12, ε is Gaussian noise. The second method is to randomly assign part of the vectors of the input sample data to 0 to add binary masking noise. The specific process of this method is to set a scale at first, then to select the components of this scale in the input sample data, and yo assign these components to 0. The details are shown in Figure 3.5.

3.3. Stack SAE Based on ELM. To enable AEN to learn high-level abstract features, to achieve drug identification, AEN has been improved and a sparse noise reduction AEN has been proposed [21, 22]. SAE is a neural network model. SAE adopts layer by layer greedy training method, and trains each layer of SAE from front to back. The hidden layer representation of the SAE layer is used as the input vector for the next layer of SAE during training. The weight and offset values of SAE are obtained layer by layer. Higher-level abstract features can be extracted from the above operations. At the top level of SAE, Logistic or Softmax classifiers are introduced to achieve binary or multi classification of data. Finally, BPNN algorithm is used to fine-tune the entire neural network. However, the way of BPNN algorithm to optimize weights and offsets is based on gradient descent method, so all parameters need to be modified in all iteration processes, and the training time is extremely expensive. In addition, the neural network trained based on the gradient descent method is easy to obtain local optimal solutions [23]. To solve this problem, ELM is introduced to replace the BPNN algorithm to adjust and optimize SAE, reduce training time, and improve training efficiency and practicality. During training, ELM can obtain the output weight of the hidden layer as long as the number of hidden layer neurons is set in advance [24]. Therefore, as long as the number of neurons in the hidden layer is set in advance, the output weight of the hidden layer can be obtained. The ELM model is shown in Figure 3.6.

In Figure 3.6, N groups of independent input samples are set, with and . When the ELM model has M hidden layer nodes, the ELM model can be expressed as Formula 3.13.

$$f(x) = \sum_{k=1}^{M} \beta_k G(W_k X_j + b_k) = o_j, 1, 2, \ldots, N \tag{3.13}$$

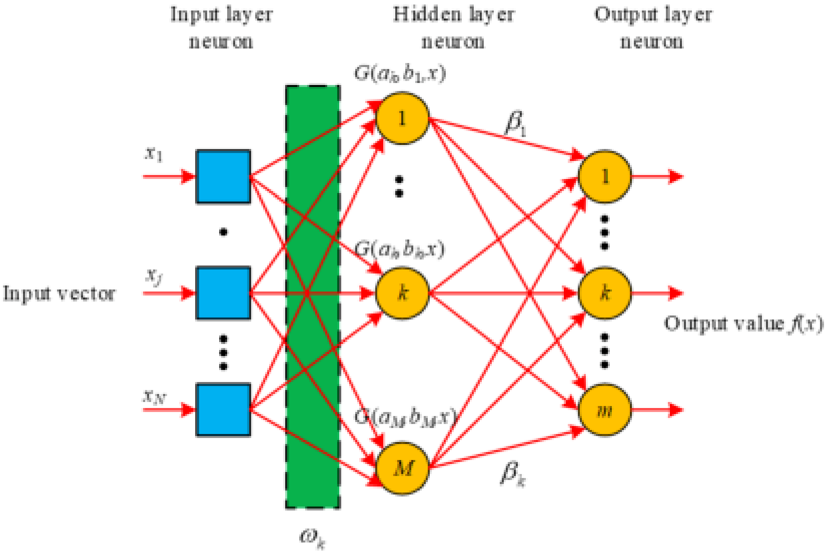


Fig. 3.6: ELM model

In Formula 3.13, $G()$ is the activation function. $W_k = [\omega_{k1}, \omega_{k2}, \ldots \omega_{kn}]^T$ is the input weight matrix. $\beta_k = [\beta_{k1}, \beta_{k2}, \ldots \beta_{kn}]^T$ is the output weight matrix. b_k is the k-th offset value in the hidden layer. o_j is the network output value. Compared with the BPNN algorithm, ELM has excellent generalization ability and training efficiency, and does not require too much manual intervention [25]. Therefore, ELM is used to adjust SAE and construct the ELM-SAE algorithm. The output of the last hidden layer of the SAE model is set to be H_{n-1}, the nth hidden layer contains $\widehat{N}$ nodes, and the $n-1$th hidden layer contains m nodes. Then the neural network function can be expressed by Formula (1.14).

$$f(x) = \sum_{n=1}^{\widehat{N}} \beta_n G(W_n H_{n-1} + b_n) = o_j, 1, 2, \ldots, m \tag{3.14}$$

In Formula 3.14, W_N represents the connection weight between the $n-1$-th hidden layer and the n-th hidden layer. β_n is the connection weight between the n-th hidden layer and the output layer. b_n represents the offset value between the $n-1$-th hidden layer and the n-th hidden layer. For SAE, its training goal is to minimize the error between the actual output and the expected output, as shown in Formula 3.15.

$$\sum_{n=1}^{\widehat{N}} = \|o_j - t_j\| \tag{3.15}$$

Formula 3.15 can be converted to Formula (1.16).

$$H_n \beta = T \tag{3.16}$$

In Formula 3.16, H_n is the output matrix of hidden layer nodes. β is the output weight. T is the desired output. To minimize the output error of the ELM-SAE algorithm, the $\widehat{W_k}, \widehat{b_k}, \widehat{\beta_k}$ value is obtained through continuous updating and iteration to make it meet the Formula 3.17.

$$\left\| H_n(\widehat{W_k}, \widehat{b_k})\widehat{\beta_k} - T \right\| = min_{w,b,\beta}(W_k, b_k)\beta - T \tag{3.17}$$

In Formula 3.17, $k = 1, 2, \ldots, \widehat{N}$. Formula 3.17 can be equivalent to the minimum loss function, as shown in Formula 3.18.

$$E = \sum_{j=1}^{m} (\sum_{k=1}^{\widehat{N}} \beta_k G(W_k H_{(n-1)j} + b_k) - t_j)^2 \tag{3.18}$$

Randomly to initialize the connection weight between the $n-1$-th hidden layer and the n-th hidden layer, and the offset value between the $n-1$-th hidden layer and the n-th hidden layer to obtain the unique output matrix H_n of the hidden layer. The training process of the ELM-SAE algorithm can be converted into a solution problem of a linear system, as shown in Formula (1.16). At this time, the output weight can be determined, as shown in Formula 3.19.

$$\widehat{\beta} = H_n^T T \tag{3.19}$$

In Formula 3.19, $H_n^T T$ is the Moore penrose generalized inverse of the hidden layer output matrix, which can be solved by the generalized inverse theorem and singular value decomposition method. Based on the above contents, the ELM-SAE algorithm is constructed to make up for the long training time of machine learning.

4. Analysis of the Application Effect of Dropout DBN and ELM-SAE in Drug Identification.

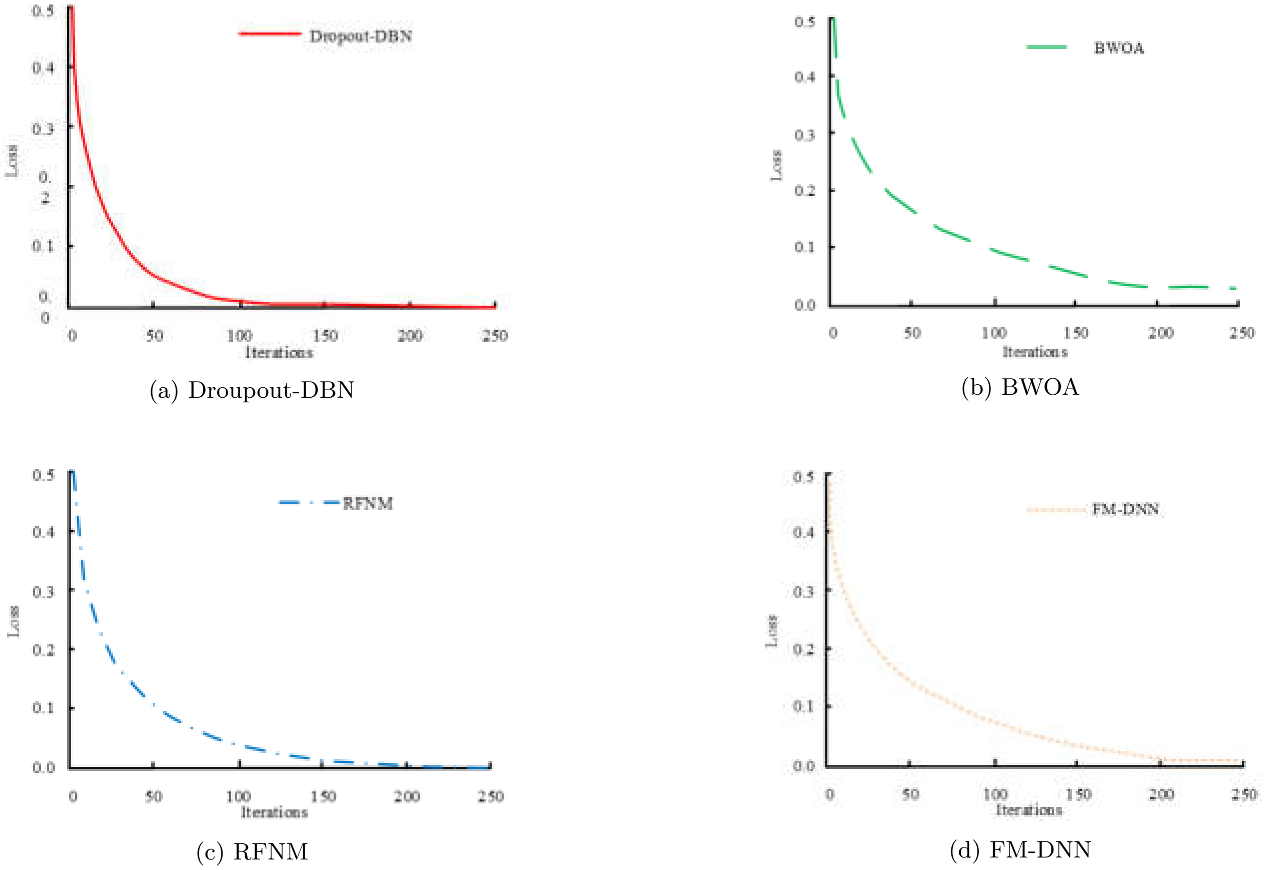
4.1. Training Effect Analysis of Dropout DBN Model. In drug identification, the amount of sample data is generally small. To solve the problem of binary classification and multi-classification modeling of NIRS in small sample datasets, a Dropout DBN algorithm based on machine learning is proposed. To verify the identification effect of the Dropout DBN algorithm on infrared spectrum drugs in a small sample data set, the research uses the data collected by the China Institute for Food and Drug Control to test the performance of the algorithm. 500 spectral sample data of erythromycin ethylsuccinate were selected as the sample data, 300 of which were used as the training sample set, and the other 200 were used as the test sample set. The performance of the Dropout DBN algorithm with BWOA, RFNM, FM-DNN and other common drug identification algorithms were compared. During the training process, the loss values of several algorithms [26, 27, 28] change as shown in Figure 4.1.

In Figure 4.1, after 100 iterations, the loss change rate of the four models slows down. Among them, the loss value of the Dropout DBN model is always lower than that of the other three models. After 150 iterations, the loss value of the four models is the lowest, and the loss value of the Dropout DBN model is 0.02. The loss value of the BWOA model is 0.05, which is 0.03 higher than that of the Dropout DBN model. The loss value of RFNM model is 0.02, but the number of iterations when it reaches the lowest value is significantly more than that of Dropout DBN. The loss value of FM-DNN is 0.03, which is 0.01 higher than the Dropout DBN model. This shows that the Dropout DBN model has better convergence and stability, higher prediction precision and a more ideal model.

4.2. Classification Accuracy Analysis of the Dropout DBN Model. After training, the classification precision of the test sample set for the four models was used, and the test results are shown in Figure 4.2.

In Figure 4.2, the precision of the Dropout DBN model is significantly higher than the other three algorithms. When the number of iterations is less than 50, the precision rate of the four models is constantly improving. After the number of iterations reaches 100, the precision rate of the four models slows down. After the precision becomes stable, the precision of the Dropout DBN model reaches 99.12%. The precision of the BWOA model is 98.63%, 0.49% lower than that of the Dropout DBN model. The precision of the RFNM model is 96.82%, and that of the FM-DNN model is 94.12%, 2.30% and 5.00% lower than the Dropout DBN model respectively. The above results show that when the number of samples is small, the Dropout DBN model has high precision in drug identification. The performance of the above four models was tested using ROC, as shown in Figure 4.3.

In Figure 4.3, the Area under the Curve (AUC) value of the Dropout DBN model is obviously superior to the other three models. The AUC value of the Dropout DBN model is 0.87, and that of the BWOA model is 0.83, 0.04 lower than that of the Dropout DBN model. The AUC value of the FM-DNN model is 0.61, 0.26 lower than that of Dropout DBN model. AUC value of RFNM model is 0.82, which is 0.05 lower than that of Dropout DBN model. This shows that the Dropout DBN model proposed in the study can effectively identify the infrared spectrum of drugs.



(a) Droupout-DBN

(b) BWOA

(c) RFNM

(d) FM-DNN

Fig. 4.1: Loss value change of several algorithms

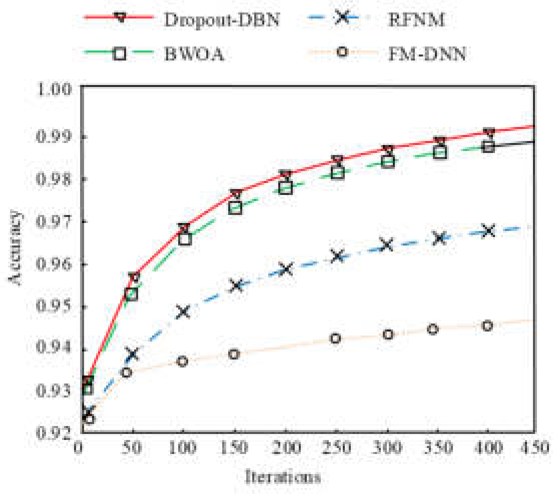


Fig. 4.2: Drug identification precision of four models

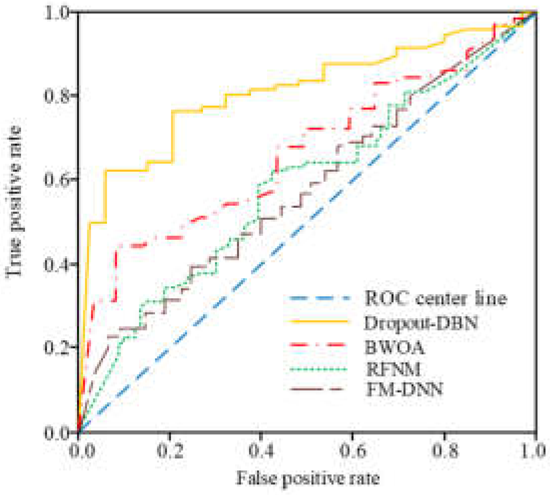


Fig. 4.3: AUC values of four models

Table 4.1: Classification precision of three algorithms for erythromycin ethylsuccinate data sets

Training Sets/Test Sets	Precision/%		
	ELM	SAE-BPNN	ELM-SAE
60/189	91.31	94.48	95.86
80/169	92.78	97.30	97.34
100/149	95.56	98.08	98.26
120/129	95.81	99.05	98.68
140/109	97.32	99.42	98.85
160/89	98.05	99.03	99.34
180/69	98.41	98.96	99.72

4.3. Classification Accuracy Analysis of the ELM-SEA Algorithm. The training time of the machine learning algorithm is slow. To solve this problem, ELM is used to optimize SAE and construct the ELM-SAE algorithm. To verify the performance of the ELM-SEA algorithm, the data of erythromycin ethylsuccinate collected by China Institute for Food and Drug Control was used to test it. ELM algorithm, SAE-BPNN algorithm and ELM-SAE algorithm are respectively used to process and identify erythromycin ethylsuccinate data sets of different orders of magnitude, and the recognition precision and learning efficiency of several algorithms are compared. The classification precision of the three algorithms for erythromycin ethylsuccinate data sets is in Table 4.1.

Table 4.1 shows the classification precision of the SAE-BPNN model is the highest, reaching 99.05%, except when the test set/training set is 120/129. In other data sets, the classification precision of the ELM-SAE model is the highest among the three models. The precision of both ELM-SAE and ELM models increases with an increase in training samples. The highest precision achieved by ELM-SAE is 99.72%, while the highest precision achieved by the ELM model is 98.41%. The classification precision of the SAE-BPNN model starts to decline when the training samples exceed a certain number. This is because the SAE-BPNN model has produced an over-fitting phenomenon and is trapped in local optimization, resulting in a decline in model precision. The training time of the three algorithms on data sets of different orders of magnitude is shown in Table 4.2.

In Table 4.2, among the three models, the training time of ELM model on each dataset is much lower than that of the other two models. However, Table 1 shows that the drug classification precision of ELM model is not ideal. In Table 2, the training time of ELM-SAE model on each data set is significantly lower than that of

Table 4.2: Training time of three algorithms on data sets of different orders of magnitude

Training Sets/Test Sets	Training time/s		
	ELM	SAE-BPNN	ELM-SAE
60/189	0.011	28.78	19.34
80/169	0.013	26.84	17.43
100/149	0.014	27.87	18.64
120/129	0.016	28.68	19.25
140/109	0.016	29.66	20.08
160/89	0.016	42.89	30.13
180/69	0.016	46.56	33.10

SAE-BPNN model. This shows that using ELM algorithm to optimize SAE can greatly improve the training efficiency and reduce the training time of near-infrared spectral data on the premise of ensuring the precision of drug identification. It is an effective method for near-infrared spectral drug identification. To sum up, the two algorithm models proposed in the study can effectively achieve NIRS drug identification, indicating that machine learning has broad application prospects in NIRS drug identification.

5. Conclusion. Machine learning has important applications in various fields. The application ways and effects of machine learning in NIRS drug identification were studied. Aiming at the problem of binary and multi-classification modeling of NIRS in small sample datasets, a Dropout DBN algorithm based on machine learning was proposed. Aiming at the shortcomings of low learning efficiency and long learning time of machine learning algorithms, an ELM-SAE algorithm was proposed. After testing the model with experimental data, the Loss value of the Dropout DBN model was 0.02, 0.03 lower than the BWOA model and 0.01 lower than the FM-DNN model. The precision of the Dropout DBN model reached 99.12%, 0.49%, 2.30% and 5.00% higher than the BWOA model, RFNM model and FM-DNN model, respectively. On the data sets of each order of magnitude, the classification precision of the ELM-SAE model was the highest among the three models. The precision of both ELM-SAE and ELM models increased with the increase of training samples. The highest precision of ELM-SAE and ELM model was 99.72% and 98.41%. However, the classification precision of the SAE-BPNN model started to decline after the training samples exceeding a certain number, which indicates that it has produced an over-fitting phenomenon. The training time of the ELM-SAE model on each data set was significantly lower than that of the SAE-BPNN model. To sum up, the two algorithm models proposed in the study can effectively achieve NIRS drug identification, indicating that machine learning has broad application prospects in NIRS drug identification. The study was conducted only with an erythromycin ethylsuccinate data set without testing the identification effect of the model on other drugs, which may lead to deviation in the experimental results. Therefore, the scope of the study needs to be expanded in the future.

Fundings. The research is supported by Science and Technology Research Program of Chongqing Municipal Education Commission: Preprocessing and Comparison Analysis of Near Infrared Spectrum Data Based on Different Objects (No. KJQN202205402); Natural Science Foundation of Yongchuan District, Chongqing: Research on Food and Drug Identification Model Based on Near Infrared Spectrum and HHT Transform (No. 2022yc-jckx20041).

REFERENCES

[1] Kademi, H., Ulusoy, B. & Hecer, C. Applications of miniaturized and portable near infrared spectroscopy (NIRS) for inspection and control of meat and meat products. *Food Reviews International.* **35**, 201-220 (2019)

[2] Zhang, H., Feng, J., Chen, S., Zhao, Z., Li, B., Wang, Y., Jia, J., Li, S., Wang, Y., Yan, M., Lu, K. & Hao, H. Geographical patterns of nirS gene abundance and nirS-type denitrifying bacterial community associated with activated sludge from different wastewater treatment plants. *Microbial Ecology.* **77**, 304-316 (2019)

[3] Vesoulis, Z., Mintzer, J. & Chock, V. Neonatal NIRS monitoring: recommendations for data capture and review of analytics. *Journal Of Perinatology.* **41**, 675-688 (2021)

[4] Paul, A., Wander, L., Becker, R., Goedecke, C. & Braun, U. High-throughput NIR spectroscopic (NIRS) detection of microplastics in soil. *Environmental Science And Pollution Research.* **26**, 7364-7374 (2019)

[5] Han, C., M"uller, K. & Hwang, H. Enhanced performance of a brain switch by simultaneous use of EEG and NIRS data for asynchronous brain-computer interface. *IEEE Transactions On Neural Systems And Rehabilitation Engineering.* **28**, 2102-2112 (2020)

[6] Kranenburg, R., Weesepoel, Y., Alewijn, M., Sap, S., Peter, W. & Esch, A. The importance of wavelength selection in on-scene identification of drugs of abuse with portable near-infrared spectroscopy. *Forensic Chemistry.* **30** (2022)

[7] Deidda, R., Sacre, P., Clavaud, M., Coic, L., Avohou, H., Hubert, P. & Ziemons, E. Vibrational spectroscopy in analysis of pharmaceuticals: Critical review of innovative portable and handheld NIR and Raman spectrophotometers. *TrAC Trends In Analytical Chemistry.* **114** pp. 251-259 (2019)

[8] Li, L., Pan, X., Chen, W., Wei, M., Feng, Y., Yin, L., Hu, C. & Yang, H. Multi-manufacturer drug identification based on near infrared spectroscopy and deep transfer learning. *Journal Of Innovative Optical Health Sciences.* **13** pp. 50016-20500 (2020)

[9] Mehmood, A., Kaushik, A., Wang, Q., Li, C. & Wei, D. Bringing structural implications and deep learning-based drug identification for KRAS mutants. *Journal Of Chemical Information And Modeling.* **61**, 571-586 (2021)

[10] Ding, Y., Tang, J. & Guo, F. Identification of drug–target interactions via fuzzy bipartite local model. *Neural Computing And Applications.* **32**, 10303-10319 (2020)

[11] Yusof, N., Muda, A., Pratama, S. & Abraham, A. A novel nonlinear time-varying sigmoid transfer function in binary whale optimization algorithm for descriptors selection in drug classification. *Molecular Diversity.* **2022** pp. 1-10 (2022)

[12] Zhao, X., Chen, L., Guo, Z. & Liu, T. Predicting drug side effects with compact integration of heterogeneous networks. *Current Bioinformatics.* **14**, 709-720 (2019)

[13] Wang, J., Wang, H., Wang, X. & Chang, H. Predicting drug-target interactions via FM-DNN learning. *Current Bioinformatics.* **15**, 68-76 (2020)

[14] Dargan, S., Kumar, M., Ayyagari, M. & Kumar, G. A survey of deep learning and its applications: a new paradigm to machine learning. *Archives Of Computational Methods In Engineering.* **27**, 1071-1092 (2020)

[15] Ma, W., Liu, Z., Kudyshev, Z., Boltasseva, A., Cai, W. & Liu, Y. Deep learning for the design of photonic structures. *Nature Photonics.* **15**, 77-90 (2021)

[16] Wang, Y., Pan, Z., Yuan, X., Yang, C. & Gui, W. A novel deep learning based fault diagnosis approach for chemical process with extended deep belief network. *ISA Transactions.* **96** pp. 457-467 (2020)

[17] Li, L., Pan, X., Yang, H., Zhang, T. & Liu, Z. Supervised dictionary learning with regularization for near-infrared spectroscopy classification. *IEEE Access.* **7** pp. 923-10093 (2019)

[18] Otto, S. & Rowley, C. Linearly recurrent autoencoder networks for learning dynamics. *SIAM Journal On Applied Dynamical Systems.* **18**, 558-593 (2019)

[19] Karimpouli, S. & Tahmasebi, P. Segmentation of digital rock images using deep convolutional autoencoder networks. *Computers & Geosciences.* **126** pp. 142-150 (2019)

[20] Wang, M., Zhao, M., Chen, J. & Rahardja, S. Nonlinear unmixing of hyperspectral data via deep autoencoder networks. *IEEE Geoscience And Remote Sensing Letters.* **16**, 1467-1471 (2019)

[21] Khamparia, A., Saini, G., Pandey, B., Tiwari, S., Gupta, D. & Khanna, A. KDSAE: Chronic kidney disease classification with multimedia data learning using deep stacked autoencoder network. *Multimedia Tools And Applications.* **79**, 35425-35440 (2020)

[22] Cui, M., Wang, Y., Lin, X. & Zhong, M. Fault diagnosis of rolling bearings based on an improved stack autoencoder and support vector machine. *IEEE Sensors Journal.* **21**, 4927-4937 (2020)

[23] Li, D., Fu, Z. & Xu, J. Stacked-autoencoder-based model for COVID-19 diagnosis on CT images. *Applied Intelligence.* **51**, 2805-2817 (2021)

[24] Yaseen, Z., Sulaiman, S., Deo, R. & Chau, K. An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction. *Journal Of Hydrology.* **569** pp. 387-408 (2019)

[25] Manoharan, J. Study of variants of Extreme Learning Machine (ELM) brands and its performance measure on classification algorithm. *Journal Of Soft Computing Paradigm (JSCP), V Qzol..* **3** pp. 02 (2021)

[26] Gouda, W., Almurafeh, M., Humayun, M. & Jhanjhi, N. Detection of covid-19 based on chest x-rays using deep learning. *Healthcare.* **10**, 343 (2022)

[27] Khalil, M., Humayun, M., Jhanjhi, N., Talib, M. & Tabbakh, T. Multi-class segmentation of organ at risk from abdominal ct images: A deep learning approach. *Intelligent Computing And Innovation On Data Science: Proceedings Of ICTIDS 2021.* pp. 425-434 (2021)

[28] Aldughayfiq, B., Ashfaq, F., Jhanjhi, N. & Humayun, M. Yolo-based deep learning model for pressure ulcer detection and classification. *Healthcare.* **11**, 1222 (2023)

Edited by: Chiranji Lal Chowdhary
Special issue on: Scalable Machine Learning for Health Care: Innovations and Applications
Received: Jul 4, 2023
Accepted: Feb 19, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1373–1386, DOI 10.12694/scpe.v25i3.2687

PERFORMANCE COMPARISON OF APACHE SPARK AND HADOOP FOR MACHINE LEARNING BASED ITERATIVE GBTR ON HIGGS AND COVID-19 DATASETS

PIYUSH SEWAL[*] AND HARI SINGH[†]

Abstract. In the realm of distributed computing frameworks, such as Apache Spark and MapReduce Hadoop, the efficacy of these frameworks varies across diverse applications and algorithms contingent upon distinctive evaluation metrics and critical parameters. This research paper diligently scrutinizes the extant body of research that compares these two frameworks concerning said evaluation metrics and parameters. Subsequently, it conducts empirical investigations to authenticate the performance of these frameworks in the context of an iterative Gradient Boosting Tree Regression (GBTR) algorithm. Remarkably, the comparative analyses in previous studies encompass a spectrum of iterative machine learning regression and classification techniques, batch processing, SQL, and Graph processing algorithms. Furthermore, numerous investigations have explored the application of machine learning algorithms encompassing logistic regression, Page Rank, K-Means, KNN, and the HiBench suite. This paper presents the comparison between the two distributed computing platforms on iterative GBTR for classification task on the HIGGS dataset from the physics domain and for the regression task on the Covid-19 dataset from the healthcare domain. The empirical findings corroborate that Apache Spark exhibits superior execution speed in iterative tasks when the available physical memory significantly exceeds the dataset size. Conversely, Hadoop outperforms Spark when dealing with substantial datasets or constrained physical memory resources.

Key words: MapReduce Hadoop, Spark, Machine Learning, Iterative, In-memory computation, Gradient Boost Tree Regression, Covid-19

1. Introduction. Today, a huge amount of data is being generated from different sources like social media platforms, IoT devices, sensors and digital devices. This data is popularly known as "Big Data" which is being generated in different forms like structured, semi-structured and unstructured [1]. As a result, the key challenge is not only to store this huge amount of data but also to process the data to gain useful insights and knowledge discovery. At present, there are several distributed data processing frameworks available such as Hadoop MapReduce, Spark, Flink, Storm, Samza etc.[2, 3].

Among these frameworks, Apache Spark and MapReduce Hadoop are two popular open-source frameworks that are widely used by different enterprises for processing large-scale data. Google's MapReduce aimed for scalability, security and fault tolerance for big data processing. The Apache Hadoop, an open-source implementation of the MapReduce model is a disk-based data processing framework suitable for batch processing jobs. However, it faced the issue of high disk access in each epoch which resulted in high I/O costs due to its inability to reuse intermediate results during the execution phase. Hence it resulted in low performance for iterative jobs. The performance of Apache Hadoop for spatial data is improved by indexing [4, 5]. The Apache Spark framework having the in-memory computational capability, overcame this limitation with a special type of data structure known as Resilient Distributed Datasets (RDDs) that supports reusability and is capable to store intermediate results in the physical memory of the system. A critical analysis of Hadoop and Spark [6], along with the high accuracy, scalability, and execution efficiency of distributed Spark MLib regression algorithms [7], as well as the performance prediction of Spark workloads using I/O parameters [8], covered in prior studies, sheds light on distributed processing frameworks.

In this study, firstly, a detailed literature survey is carried out that compares the two distributed computing frameworks – Apache Spark and MapReduce Hadoop on performance evaluation metrics and key parameters. Secondly, the two frameworks are compared through experimental work on the iterative Gradient Boost Tree Regression algorithm on metric execution time, the effect of memory size and varying dataset size on execu-

[*] CSE & IT Department, Jaypee University of Information Technology, Solan, HP, India. (piyush.sewal@gmail.com)
[†] CSE & IT Department, Jaypee University of Information Technology, Solan, HP, India. (hsrawat2016@gmail.com)

tion time. The experimental work uses the HIGGS dataset [9] and the Covid-19 datasets [10] on clusters of varying sizes.

The rest of the paper is as follows. The related work is presented in Section 2. Section 3 compares experimental results for the iterative Gradient Boost Tree Regression algorithm under different cluster configurations and datasets. Finally, Section 4 concludes the paper.

2. Related Work. This section presents a detailed literature review on the distributed computing frameworks – Apache Spark and MapReduce Hadoop. The review is carried out in two different sets of parameters in sub-section 2.1 and 2.2.

2.1. Review of evaluation metrics. This sub-section presents the review of evaluation metrics execution speed, memory usage, cluster size, data characteristics, CPU utilization and network usage. It is presented in Table 2.1 along with its description in the text. The issues of latency, I/O and deserialization cost are analyzed in a distributed memory structure Resilient Distributed Datasets (RDDs) in the Spark that uses in-memory computations. Then Spark and Hadoop are compared on Logistic Regression, K-means and Page Rank algorithms. The experimental results validate the high execution speed of Spark than Hadoop for iterative and graph-based applications. It is also observed that Spark recovers the lost RDD partitions very quickly in case of node failures. However, the performance of the Spark degrades more than the Hadoop when memory is not sufficient [11]. Similar results were obtained when researchers compared the Hadoop and Spark frameworks using the Page Rank algorithm [12].

The running time of the K-Means algorithm of the HiBench benchmark on Hadoop and Spark clusters with different sizes of memories allocated to data nodes shows that Spark performs better as long as the memory size is sufficient enough for the data size [13]. Another similar work evaluated the two frameworks on execution speed for the K-Means algorithm using sensor datasets of varying size on different cluster sizes and obtained similar results [14]. In another paper, the K-Means algorithm is applied to the satellite images dataset with modified values of K in three phases which include the Initialization Phase, Clustering Phase and Validation Phase [15]. The experimental results show that the speed-up performance and scalability of the algorithm is improved on both the Spark and MapReduce clusters. In another research work, the K-Means algorithm and Page-Rank algorithm show that the Spark performance improves for the iterative algorithm of data reuse [16]. The Spark is about 40 times faster than Hadoop for a data quantity of about 40 thousand points. However, the Spark performance declines and then saturates but it remains 8 times fast as compared to Hadoop with an increase in data quantity. The results also show that Spark has a significant performance for iterative jobs with a low latency schedule when the size of the dataset exceeds the memory size. In another work, the authors compare the two frameworks on execution time, CPU utilization, memory and network usage for the K-Nearest Neighbor (KNN) algorithm on different size datasets and cluster configurations. The Spark is observed to have better execution speed and CPU utilization but memory size is the bottleneck. The Hadoop is observed to consume more network resources but the memory size does not create any performance problems [17].

The Spark is reported to perform well on the HiBench benchmark suit when different datasets are used [18]. The authors classified thirteen workloads benchmark suites into four categories Micro Benchmarks, Web Search, SQL and Machine Learning. Among these, eight benchmarks Aggregation, Bayesian, Join, Pagerank, Scan, Sleep, Sort and Terasort are taken into consideration for measuring the performance. Then the performance is evaluated using three metrics execution time, throughput and speed-up. It is observed that the execution time is very less in Spark as compared to Hadoop with a factor of 18. The reason behind the more execution time in Hadoop is due to multiple object creation for single input, slow data sharing due to replication and serialization and disk-based I/O. Similarly, the throughput and speedup are also better for the Spark cluster than the Hadoop cluster.

The wordcount program is used to compare the two frameworks for different size datasets [19]. The experimental results show that Hadoop takes more time as compared to Spark irrespective of the size of the dataset. The Spark counts the occurrence of each word in less time due to its in-memory computational capabilities.

In a similar work, the wordcount program on four different datasets validates the better performance of Spark over Hadoop on execution time [23]. In another work, the wordcount program is implemented on a publically available word file and logistic regression is applied to a dataset of bankruptcy conditions of companies

[20]. The experimental results show that Spark performs better than MapReduce for both normal and iterative queries with a performance ratio of 2.8 and 2.2 for wordcount and logistic regression respectively. The study is performed on a single machine and does not consider varying cluster sizes. Another work on the same algorithm with a cluster of fixed size validated the high performance of Spark for iterative and streaming data processing whereas Hadoop is found suitable for batch data processing [24].

In another research work, the authors compared Hadoop and Spark on streaming data for comparing the execution time. The Spark engine is used to process twitter's tweets in a short interval of less than a second. The results indicate that Spark is better than Hadoop for streaming data processing due to its in-memory processing, lower disk-access rate and event-driven task scheduling [22]. In another work, Spark outperformed MapReduce for processing streams of text and video on GPU for execution time and throughput [25]. The video data is captured from Youtube with different road and traffic scenarios. The text data is collected from sensors and social networks by using Apache Spark Streaming. For video-related big data, the processing time of large videos increased considerably in MapReduce Hadoop in comparison to Spark.

In another work, the authors used the Apriori algorithm in three different execution approaches IMRAprioriAcc (Improved MapReduce Apriori Accelerated), DPC (Dynamic Passes Combined-Counting) and CPA (Complete Parallel Apriori) along with their adaption on Spark with different size datasets and varying cluster configuration [21]. Four performance metrics runtime, speed-up, size-up and scale-up are used for the performance evaluation of the Hadoop MapReduce and Spark. The experiment results of the work validate the better performance of Spark over Hadoop MapReduce. The implementation of CPA with MapReduce gives better results than Spark when the size of the dataset is large and physical memory is not sufficient. In recent works, the authors compared various machine learning regression algorithms [26] and Hadoop and Spark for execution time and throughput using different size Covid-19 datasets on a fixed-size cluster [27]. In the latter, the experiment results validate the high execution speed of Spark for small datasets.

2.2. Comparison of key parameters. This sub-section presents the comparison of key parameters data processing, performance, latency, fault-tolerance, scalability, security, cost, scheduling, resource management, inbuilt capabilities, usability and language support. It is presented in Table 2.2 along with its description in the text. The Hadoop is best suited for batch data processing as it uses MapReduce which splits large datasets among various clusters and processes these in parallel [28]. In the MapReduce architecture, data passes in four phases which are splitting, mapping, shuffling and reducing. On the other hand, Spark is suitable for iterative and live streaming data which is mostly in the unstructured form. It uses the concept of in-memory processing of data and works with the help of RDDs to perform various operations [29, 30]. Spark creates DAG (Directed Acyclic Graph) that contains vertices and edges where vertices represent RDDs and edges represent the operations to be performed on the RDDs [11].

The MapReduce Hadoop is not efficient for iterative operations because it cannot keep reused data and state information during execution [17]. It is a high-latency framework in which integrative mode is not available. Thus it persists intermediate data onto the disk that further results in slow data processing. However, Spark is a low-latency computing framework and it can process data interactively. It results in faster processing of data as it reads the disk only once and then all the intermediate operations are performed within the RAM [2, 11].

The MapReduce Hadoop accesses disk for data storage and processing which comparatively results in slower processing whereas the Spark performs in-memory data processing that results in faster processing of data [12, 17, 31, 32, 33, 34]. The processing capabilities of the Spark are affected significantly by the size of available memory in comparison to the Hadoop [13]. The memory utilization is proven better in Hadoop whereas CPU utilization is proven better in Spark [17].

The Hadoop uses the concept of 3X replication and erasure coding for backing up data in case of any node failure. 3X replication generates 200% overhead in storage space as compared to just 50% in the case of erasure coding. On the other side, Spark uses DAG to rebuild the data using RDD across the nodes which also avoids storage space overhead. Hence both platforms have good fault tolerance mechanisms [11, 35].

Scalability is another important parameter in big data processing frameworks. Nodes and disks can be added easily on the fly and the latest versions of Hadoop are capable to add more than ten thousand nodes at a time. On the other hand, scalability is a bit challenging in Spark because it depends upon the computational capabilities of machines which may be different. However, Spark also supports thousands of nodes in a cluster[33, 36, 37].

Table 2.1: Performance comparison of Hadoop and Spark on various applications/algorithms

Application	System Configuration				Dataset Size	Dataset Type	Evaluation Parameters						Better Performer
	RAM (GB)	Disk (GB)	Disk Type	Cluster Nodes			A	B	C	D	E	F	
K-Means, Page Rank [16]	4	NA	NA	4	10K to 20M points	Log data	✓	x	✓	x	x	x	Spark
Page Rank, Logistic Regression, K -Means [11]	15	NA	NA	4	54GB	Wikipedia dump	✓	✓	✓	x	x	x	Spark
Page Rank [12]	4	NA	NA	8	NA	Graph data	✓	x	✓	✓	x	x	Spark
K-Means [14]	4	500	HDD	2	64 MB, 1240 MB	Sensor data	✓	✓	✓	x	x	x	Spark
Multiple K-Means++ [15]	8	NA	NA	5	1GB to 4GB	Satelite images data	✓	x	✓	x	x	x	Spark & Hadoop
Wordcount, Logistic Regression [20]	8	1000	HDD	1	NA	Text and numeric data	✓	x	✓	x	x	x	Spark
Hibench suit (8 benchmarks) [18]	4	40	SSD	1	Different for each benchmark	NA	✓	x	✓	✓	✓	x	Spark
K-Means [13]	16	500	HDD	3	1GB to 8GB	NA	✓	✓	✓	✓	x	x	Spark & Hadoop
Apriori [21]	8	500	HDD	20	NA	Synthetic data	✓	✓	✓	x	x	x	Spark & Hadoop
Flume, Spark streaming [22]	5	40	NA	1	NA	Twitter data	x	x	✓	x	x	x	Spark
Wordcount [19]	4	1000	HDD	4	1 MB to 300 MB	Text data	✓	x	✓	x	x	x	Spark
Wordcount, Logistic Regression, K-Means [19]	NA	NA	NA	1	500 MB to 40 GB	Wikipedia and Enron	✓	x	✓	x	x	x	Spark
K-NN [17]	4	NA	NA	6	8 GB	CSV data	✓	✓	✓	✓	✓	✓	Spark& Hadoop
Wordcount [23]	2	NA	NA	3	34 MB to 202 MB	Text data	✓	x	✓	x	x	x	Spark

(Abbreviations used: A: Varing dataset size, B: Varing cluster size, C: Execution time, D: Memory Usage, E: System Throughput, F: Networking, GB: GigaByte, MB: MegaByte RAM: Random Access Memory, HDD: Hard Disk Drive, SSD: Solid State Drive, NA: Not Available.)

Security is always a major concern while processing large datasets and Hadoop address this issue very effectively. The Hadoop supports ACLs, SLAs, LDAP and Kerberos which makes it extremely secure. The Spark does not provide such level of security and its security is turned off by default. However, Spark provides authentication with the help of event logging or shared secrets which is not sufficient. Thus, Spark integrates with Hadoop to achieve a significant security level [38].

Along with some key performance parameters, cost also plays an important role in the big data processing. Although both Hadoop and Spark are open-source platforms, when hardware resources are considered, Hadoop is less expensive as it relies on disks for storage and processing whereas Spark performs in-memory processing which slightly increases the data processing cost [20, 33, 36].

Table 2.2: Comparison between Hadoop and Spark on key parameters

Parameters	**Hadoop**	**Spark**
Data Processing [6, 11, 28, 29, 30]	Suitable for batch processing	Best for iterative and streaming data processing
Latency [2, 11]	High	Low
Performance [11, 12, 13, 17, 21, 24, 34]	Comparatively slower	Comparatively faster
Fault Tolerance [6, 11, 35]	Supported	Supported
Scalability [33, 36, 37]	Easily scalable	Scalability is a bit challenging
Security [33, 36, 38]	Extremely secure	Comparatively less secure
Cost [20, 33, 36]	Less expansive	A bit more expensive
Scheduling and Resource Management [33, 36, 39]	Use external solutions	Built-in solutions available
In-built capabilities [6, 22, 24, 40, 41, 42]	HDFS, YARN, MapReduce	Spark Core, Spark SQL, Spark Streaming, SparkML, GraphX
Usability and Language Support [33, 36]	A bit difficult	User friendly

For scheduling and resource management, Hadoop uses external solutions which include ResourceManager, NodeManager and YARN for resource management, CapacityScheduler and FairScheduler for resource allocation, and workflow scheduling is done by Oozie [39]. On the other hand, Spark has in-built support for job scheduling, resource management and monitoring. Spark uses DAG Scheduler for dividing the operations into stages and each stage consists of various tasks that need to be done by the Spark computation engine. Both Hadoop and Spark are equipped with some in-built components. In Hadoop, in-built components include HDFS which is used as a file system, YARN for resource management and MapReduce as the processing engine. On the other side, Spark has Spark core as the processing engine, Spark streaming for near real-time data processing, GraphX for graph processing and Spark SQL for structured data processing. Both Hadoop and Spark support machine learning libraries [24]. Hadoop uses the external library Mahout for machine learning whereas Spark has a in-built library MLlib. The experimental results show that due to the in-memory processing of Spark, MLlib is much faster than Apache Mahout but when data is extremely large MLlib sometimes crashes because of insufficient memory whereas Mahout process the data continuously even with a slow speed [40, 41, 42].

Lastly, if we compare the usability and language support of both platforms, Hadoop was developed in Java language whereas Apache Spark was developed in Scala language. If we consider the usability of both data processing platforms, Hadoop has limited language support and it uses Java and Python languages for MapReduce applications. Spark on the other hand is more user-friendly and allows interactive shell mode. Spark has large language support and its APIs can be written in Scala, Python, R, Java and Spark SQL [33, 36].

3. Performance Comparison of the Apache Spark and MapReduce Hadoop on the iterative GBTR. This section covers the performance comparison of Hadoop and Spark frameworks on the iterative GBTR. The performance of both frameworks has been analyzed on the benchmark dataset HIGGS [9] and two real-life Covid-19 [10] datasets. This section first briefs the GBTR algorithm and then presents experimental results and discussions.

3.1. Gradient Boost Tree Regression Background. The GBTR improves the mistakes of the previous learner with the help of the next learner. This algorithm is similar to the AdaBoost algorithm (adaptive boosting) and uses an ensemble tree for predicting the target variables. However, the depth of the tree is more than one here. During the implementation of the algorithm, firstly a base model is initialised with constant values and an average value of the target variable is calculated using the following equation:

$$F_0(x) = arg_\gamma min \sum_{i=1}^{n} L(y_i, \gamma) \tag{3.1}$$

Here L denotes the loss function, y_i is the observed value, γ is the predicted value and arg_γmin is the predicted value for which the value of the loss function is minimum. The Loss function L for the target variable can be

Algorithm 1: Gradient Boost Tree Regression for predicting the target variable

```
Data: Dataset file (in CSV format)
Result: Final Prediction concerning target variable
Description: This algorithm is used to predict the value of target variable using the Gradient Boost
  Tree Regression algorithm.
Step 1. Choose the input dataset (HIGGS/Covid-19) and select the IndependentFeatures and
  DepenedentFeatures as input values;
Step 2. Intialize a base model with IndependentFeatures and DependentFeatures;
$TargetVariable \leftarrow DependentVariable$;
and compute y using equation 3.1 where y is the average value of TargetVariable;
Step 3. while $Residuals = Null$ do Compute Residuals using equation 3.1 and 3.2 ;
$Residuals \leftarrow ActualOutputValue - PredictedOutputValue$ ;
end while
Step 4. Initialize an empty ensemble list: Ensemble = [] ;
Initialize i = 1 ;
while $i \leq N$ do /* Where, N is the number specified by hyperparameter tuning */ Construct
  $DecisionTree_i$;
$PredictTargetValue_i$ using $DecisionTree_i$ within the ensemble;
Compute New $Residual_i$ using equation 3.3 ;
$AddDecisionTree_i$ to the ensemble ;
$i \leftarrow i + 1$
end while
Step 5. Use all $DecisionTree_i$ within the ensemble for final prediction as to the value of
  TargetVariable;
```

calculated as:

$$L = \frac{1}{n}\sum_{i=1}^{n} L(y_i - \gamma_i)^2 \tag{3.2}$$

Here residuals are calculated by taking the difference between the actual and predicted values. The predicted value is the average value which is calculated in the first step of the base model. After this, a decision tree is constructed and new values of the target variable are predicted. Mathematically, pseudo residuals can be calculated using the following equation:

$$\gamma_i m = \left[\frac{L(y, F(x_i))}{F(x_i)}\right] \forall i \in \{1toN\}, WhereF(x) = F_{m-1}(x) \tag{3.3}$$

Here F(x) is used to calculate the value of the updated model by using the previous model $F_{m-1}(x)$. To prevent low bias and high variance, a learning rate variable (between 0 to 1) is multiplied with newly calculated residuals which is important to use for improving the accuracy of the model in long run. With each successive step, new residuals are calculated again and steps are repeated till the value matches the value of the hyperparameter. In the end, all the decision trees are used within the ensemble and the final prediction is calculated concerning the target variable. An algorithm to demonstrate the working of GBTR has been presented in Algorithm 1.

3.2. Results and Analysis on the Benchmark Dataset - HIGGS. In the first level of the execution stage, the Spark and Hadoop frameworks are tested on the HIGGS dataset [9]. The dataset has been produced using Monte Carlo simulations and contains 11 million samples with 28 features for each. The first 21 features are the kinematic properties and the last 7 features are the functions of the first 21 features. In our execution environment, we have used five nodes cluster where each node has 4 GB RAM, 512 GB HDD, Intel Core i5

Table 3.1: Details of HIGGS sample datasets and cluster configuration

HIGGS Dataset Samples	**Sample Size (MB)**	**Number of Records**	**Number of Features**	**Cluster Size**
H1	20	30000	28	5
H2	40	60000	28	5
H3	80	120000	28	5
H4	160	240000	28	5
H5	320	480000	28	5
H6	6400	960000	28	5
H7	1000	1375000	28	5
H8	2000	2750000	28	5
H9	4000	5500000	28	5
H10	8000	1100000	28	5

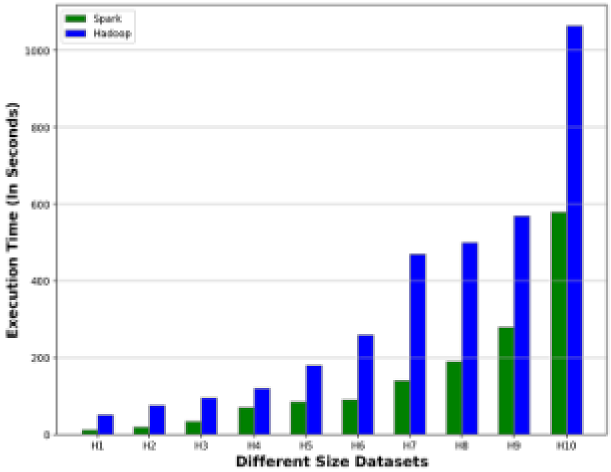


Fig. 3.1: Comparison of the Execution time of GBTR algorithm on varying size datasets on a fixed cluster size=5

processor, Hadoop 3.2 and Spark 3.2. All the systems are connected with 100Mbps local area network. The performance is evaluated using two different scenarios which are covered in the next sections.

3.2.1. Varying size datasets on a fixed cluster size. Ten different samples of HIGGS datasets (H1 to H10) with sizes from 20 MB to 8000 MB are used for the execution. The GBTR is used for execution on Hadoop and Spark clusters on a cluster size of five. The details are presented in Table 3.1.

It is clear from FIG. 3.1 that in the case of small size datasets, the dominance of Spark over Hadoop is nearly 4 to 5 times but as the size of the dataset keeps increasing, this dominance starts reducing. In the case of samples H9 and H10, it is observed that the Spark is 1.5x to 2x faster than the Hadoop. So, the Hadoop is not suitable for small size datasets but when the dataset size is large enough then the Hadoop performs well. On the other hand, Spark is good for small size datasets but if the size of the dataset is large then the Spark either need sufficient physical memory or its performance will start degrading. So the size of the dataset plays an important role in the performance evaluation of Hadoop and Spark.

3.2.2. Fixed size dataset on varying cluster size. In the second scenario, a sample HIGSS dataset of 320 MB is used for experimentation on five different cluster configurations C1, C2, C3, C4 and C5 having one,

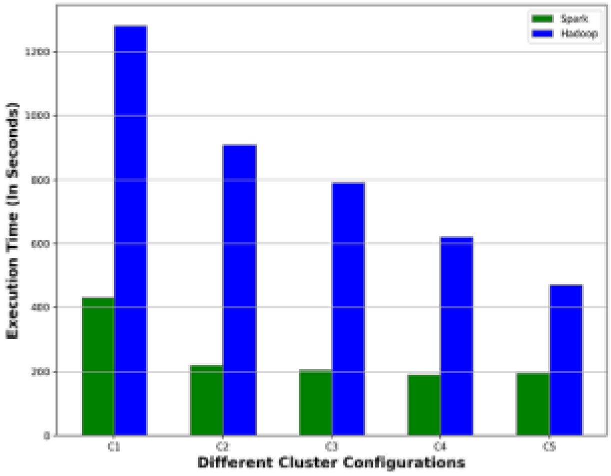


Fig. 3.2: Comparison of the Execution time of the GBTR algorithm on different cluster configurations with the same size dataset

two, three, four and five machines respectively. It observed the impact of varying system configurations on the execution time of the GBTR algorithm with the same dataset. It is clear from FIG. 3.2 that the execution time of both Hadoop and Spark is very high in C1.

Although, with the increase in system configuration, additional resources are available for execution that results in a decrease in execution time for parallel tasks which can be validated from the execution statistics of configuration C2. The execution time of the Spark does not vary in the case of C3, C4 and C5. This is because Spark got sufficient resources for execution and after a certain limit of computational resources there is no effect on execution time. The execution time of the Spark in C5 is slightly more than in C4. On the other hand, the execution time of Hadoop gradually decreases from C1 to C5. It means that the allocation of additional resources in Hadoop is helping in reducing the execution time. So it can be concluded from this experiment that granting the additional resources in distributed processing framework can be useful in reducing execution time but after the saturation point, the execution time will remain constant or it may start increasing and that saturation point of cluster configuration is directed related with dataset size and execution algorithm.

However, this fact cannot be neglected that the execution time of Spark is still less than Hadoop. The main reason behind this is the capability of Spark to perform in-memory computations with the help of RDDs. The intermediate operations performed on RDDs can be visualized through a graph which is known as Directed Acyclic Graph (DAG). The DAG is basically a set of vertices and edges where vertices represent the RDDs and edges represent the operations performed on RDDs. The Spark RDDs splits into stages by job scheduler on the basis of various transformations. During the execution phase of Spark framework, the GBTR application is divided into 203 stages. Each stage performs the transformations operation on intermediate RDDs and finally performs the action operation in the last stage. Stages 1 to 6 perform distinct operations on RDDs and then Stage 7 to 200 performs the same operations as Stage 5 and 6 but on different intermediate RDDs. Then all the intermediate orations are consolidated in Stage 201, Stage 202 shuffles the final results and Stage 203 shows the final results of GBTR on the console. An overview of Spark stages for GBTR application in the form of DAG has been presented in FIG. 3.3.

3.3. Results and Analysis on the Real-life Covid-19 Datasets. The performance evaluation of Hadoop and Spark clusters was conducted using the iterative GBTR algorithm on two datasets, Dataset-1 and Dataset-2, of varying sizes from the Covid-19 datasets of India and the world available on Kaggle [10]. After data pre-processing phase, data cleaning and data transformation, the experiment is performed on one, two and five machines equipped with 64-bit Windows OS, 4 GB RAM, 512 GB HDD, Apache Spark, Hadoop, Mahout, Python, Java, Eclipse and Anaconda Navigator respectively. The same system configuration is used for Hadoop

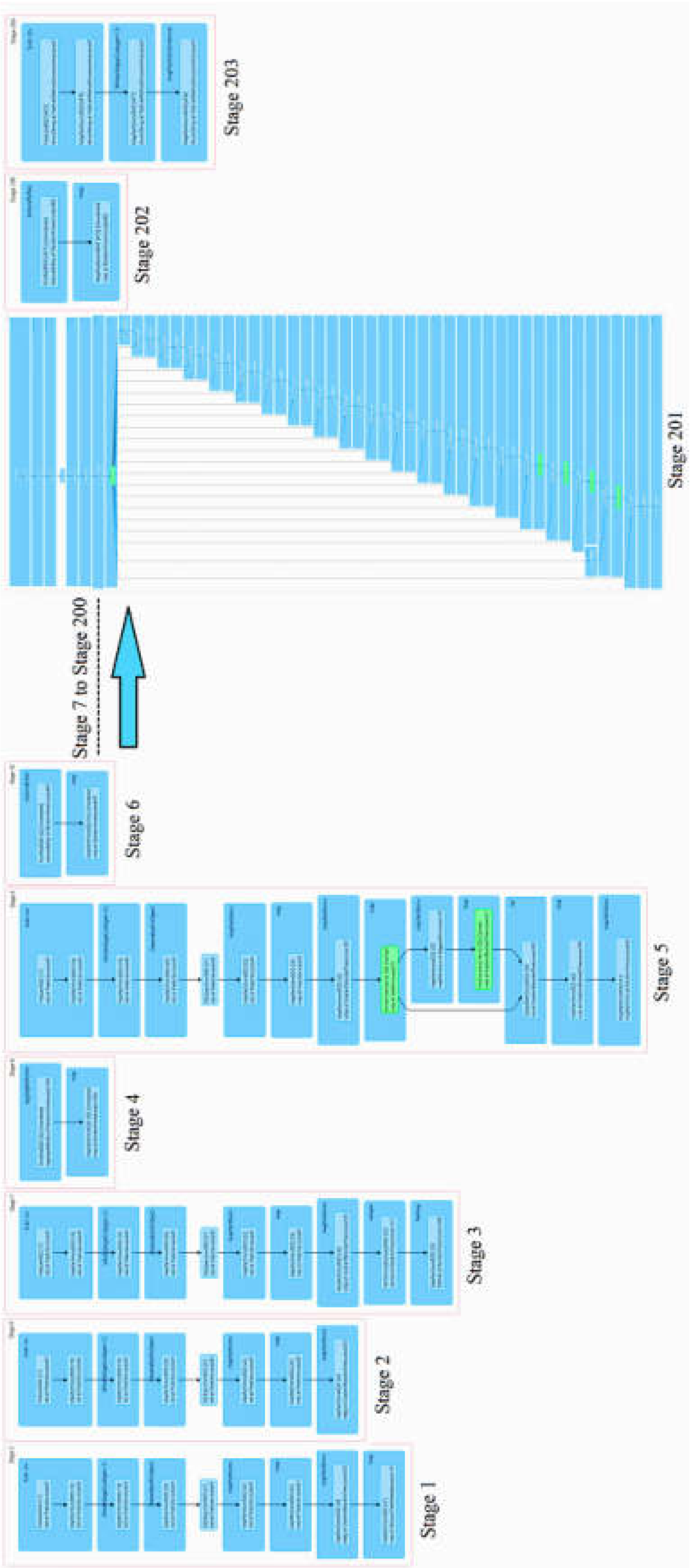


Fig. 3.3: DAG visualization of GBTR Algorithm at stage level during execution phase

Table 3.2: Cluster Specification of Hadoop and Spark

Title	**Hadoop Cluster**			**Spark Cluster**		
ClusterSize (Nodes)	1	2	5	1	2	5
Master Node	1	1	1	1	1	1
Worker Nodes	1	1	4	1	1	4
Number of CPU Cores	4	8	20	4	8	20
Total Internal Memory (GB)	4	8	20	4	8	20
Total Secondary Memory (GB)	512	1024	2560	512	1024	2560
Secondary Memory Type	HDD			HDD		
Processing Framework & Version	Hadoop 3.2			Spark 3.2		
Machine Learning Library	Mahout			Spark MLib		
Scala Version	2.12.15			2.12.15		
Java Version	11.0.13			11.0.13		
IDE	Eclipse			Jupyter Notebook		
API	Java			Python		
OS	Windows 10			Windows 10		
Processor	Intel Core i5-6500 CPU @ 3.20 GHz					

Table 3.3: Comparison of Hadoop and Spark for GBTR Algorithm on different size datasets and cluster configurations

Dataset	**Execution Case**	**No. of Records**	**Cluster Size**	**Execution Time (sec)**	
				Hadoop (Mahout)	**Spark (MLib)**
Dataset 1	Case I	560	1 Node	541	220
			2 Node	317	18.6
			5 Node	206	17.2
	Case II	18110	1 Node	754	311
			2 Node	429	22.4
			5 Node	293	17.8
Dataset 2	Case III	494	1 Node	821	371
			2 Node	472	21.9
			5 Node	325	18.8
	Case IV	306429	1 Node	1019	489
			2 Node	589	24.2
			5 Node	382	21.7

and Spark as given in Table 3.2.

In Case-I, 18,110 records from dataset-1 were grouped based on the number of days, ranging from day 1 to day 560. Consequently, day-wise data was consolidated, reducing the record count to 560, and the algorithm was executed under three distinct cluster configurations. Case-II involved the use of dataset-1 in its entirety, encompassing all 18,110 records. In Case-III, a similar day-wise grouping approach as in Case-I was followed, but this time with 306,429 records from dataset-2, resulting in a dataset of 494 records for algorithm execution. Finally, in Case-IV, the complete dataset-2, consisting of all 306,429 records, was employed.

Findings, illustrated in Table 3.3, highlight that across all four cases, the execution time for both Hadoop and Spark is notably high when executed on a single node. However, statistical analysis clearly reveals that

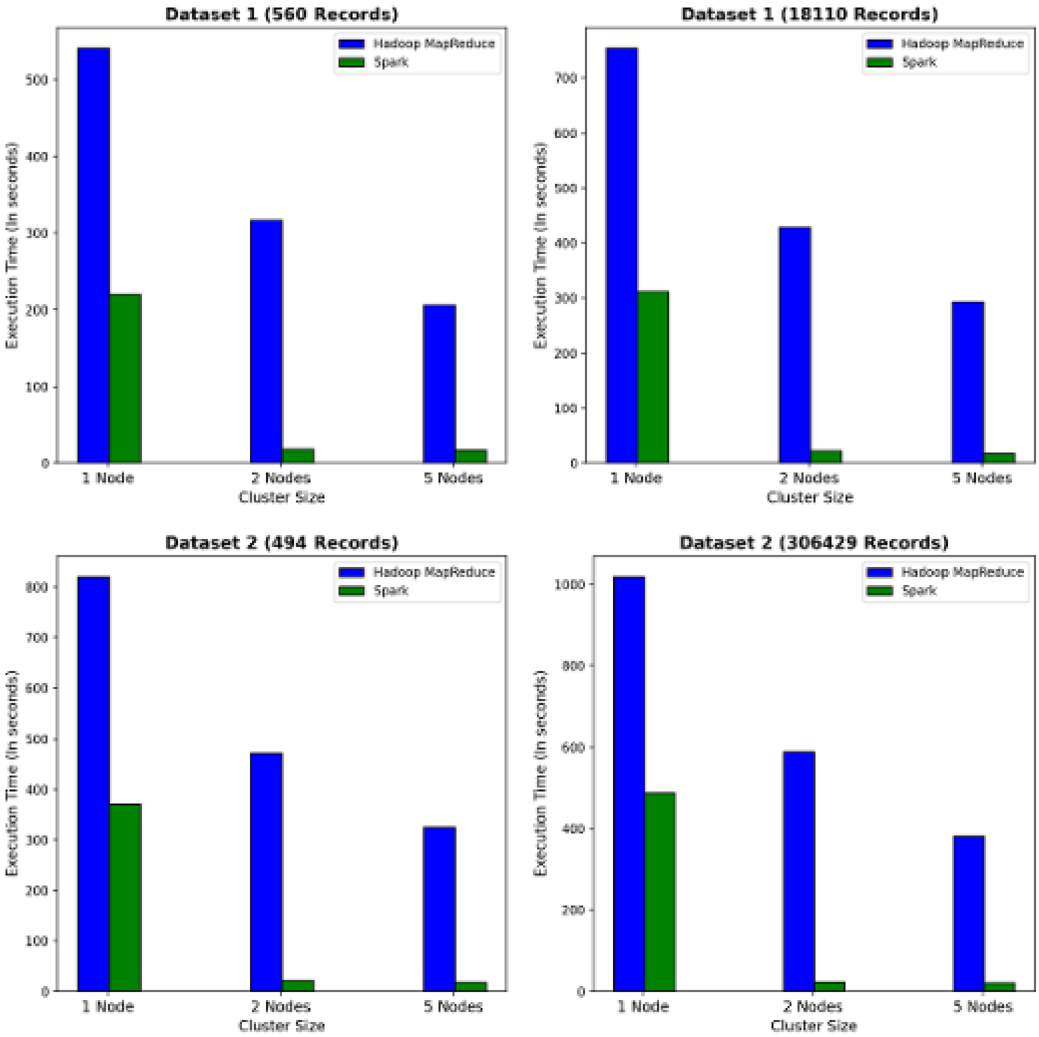


Fig. 3.4: Comparison of the execution time of Hadoop and Spark for Gradient Boosting Algorithm under different cluster configurations

Hadoop exhibits higher execution times than Spark, as indicated in FIG. 3.4. This disparity arises because Hadoop accesses the disk multiple times during iterative tasks, leading to increased latency. In contrast, Spark leverages in-memory data processing, creating RDDs and executing transformations and actions without repeated disk reads. An important observation is that the addition of nodes to the cluster, from one to two and then five, results in a sharp reduction in Spark's execution time. Conversely, Hadoop MapReduce exhibits a gradual decrease in execution time under similar conditions, as depicted in FIG. 3.4. This phenomenon is attributed to the increased number of nodes, which also augments available physical memory, contributing to Spark's enhanced processing speed. It's noteworthy that the execution time remains consistent for Spark clusters with 2 nodes and 5 nodes, indicating that the resources of the two-node cluster suffice for processing the dataset, and additional resources do not impact execution time. Conversely, while disk space increases for Hadoop, it has limited utility in expediting data processing tasks, particularly with small datasets. However, in the context of batch processing of larger datasets, Hadoop surpasses Spark when physical memory resources are constrained.

4. Conclusions and Future Work. In summary, this paper presents an extensive comparative analysis of evaluation metrics and critical parameters between Apache Spark and MapReduce Hadoop across a diverse array of algorithms, including K-Means, Page Rank, Word Count, Logistic Regression, Apriori, and the HiBench Suite. Our analysis reveals that both Hadoop and Spark exhibit commendable data processing capabilities, efficient scalability, and robust fault tolerance mechanisms. Nonetheless, it is worth noting that Hadoop excels in batch data processing, albeit at the cost of frequent disk memory access, resulting in increased disk latency and relative sluggishness. Conversely, Spark emerges as the preferred choice for iterative and streaming data processing, owing to its in-memory computational prowess, translating into superior performance vis-à-vis Hadoop. While Spark consistently outperforms Hadoop across most scenarios, it is prudent to acknowledge Hadoop's superior performance in scenarios characterized by substantial data sizes and constrained physical memory resources. Finally, our literature review findings find empirical validation through an experimental examination that compares these two frameworks, employing the iterative Gradient Boost Tree Regression algorithm.

This validation encompasses a two-tiered approach, the initial phase employing benchmark HIGGS datasets across distinct scenarios, encompassing varying dataset sizes under identical system configurations, as well as uniform dataset sizes across divergent cluster configurations. In the second phase of execution, both Hadoop and Spark are applied to actual Covid-19 datasets, illustrating practical scenarios for both frameworks and shedding light on their respective advantages for future endeavours in constructing real-world application models. The ensuing performance evaluations affirm Spark's consistent superiority over Hadoop across all scenarios, with a marked reduction in execution time as cluster size increases for Spark, whereas Hadoop MapReduce exhibits linear execution time degradation under analogous conditions.

The findings of this research paper can guide future research efforts by highlighting the suitability of Apache Spark and MapReduce Hadoop for specific use cases. Researchers can consider these frameworks' strengths and weaknesses when choosing platforms for various data processing needs, considering factors such as data size and type, memory resources, nature of algorithm, and processing requirements. Additionally, the paper's experimental validation using real-world datasets provides practical insights into the performance of these frameworks, aiding future endeavours in building real-world application models.

REFERENCES

[1] P. Muthulakshmi and S. Udhayapriya,*A Survey on big data issues and challenges*, Int. J. Comput. Sci. Eng., Vol. 6, No. 6, pp. 1238–1244, 2018, doi: 10.26438/ijcse/v6i6.12381244.

[2] T. R. Rao, P. Mitra, R. Bhatt, and A. Goswami,*The big data system, components, tools, and technologies: a survey*, Knowl. Inf. Syst., Vol. 60, No.3, pp. 1165–1245, 2019, doi: 10.1007/s10115-018-1248-0.

[3] C. Dobre and F. Xhafa,*Parallel programming paradigms and frameworks in Big Data Era*, Int. J. Parallel Program., Vol. 42, No.5, pp. 710-738, 2014, doi: 10.1007/s10766-013-0272-7.

[4] H. Singh and S. Bawa,*A mapreduce-based efficient H-bucket PMR quadtree spatial index*, Computer System Science and Engineering, Vol. 32, No. 5, pp. 405–415, 2017.

[5] H. Singh and S. Bawa,*IGSIM: An improved integrated Grid and MapReduce-Hadoop architecture for spatial data: Hilbert TGS R-Tree-based IGSIM*, Concurrency Computation : Practice and Experience, John Wiley & Sons, Vol. 31, Iss. 17, 2019, doi:https://doi.org/10.1002/cpe.5202.

[6] P. Sewal and H. Singh, *A Critical Analysis of Apache Hadoop and Spark for Big Data Processing*, in 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC), pp. 308–313, 2021, doi: 10.1109/ISPCC53510.2021.9609518.

[7] P.Sewal and H. Singh,*Analyzing distributed Spark MLlib regression algorithms for accuracy, execution efficiency and scalability using best subset selection approach*, Multimedia Tools and Applications, 2023, doi: 10.1007/s11042-023-17330-5.

[8] P. Sewal and H. Singh, *A Machine Learning Approach for Predicting Execution Statistics of Spark Application*, in the proceedings of the 2022 7th Int. Conf. Parallel, Distrib. Grid Comput., pp. 331–336, 2022, doi: 10.1109/PDGC56933.2022.10053356.

[9] Daniel Whiteson, *UCI Machine Learning Repository: HIGGS Data Set*, https://archive.ics.uci.edu/ml/datasets/HIGGS (accessed Dec. 05, 2022).

[10] Kaggle, *Your Machine Learning and Data Science Community*,https://www.kaggle.com/ (accessed Mar. 23, 2022).

[11] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker andI. Stoica, *Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing*, in Proceedings of NSDI 2012: 9th USENIX Symposium on Networked Systems Design and Implementation, pp. 15–28., 2012

[12] L. Gu and H. Li, *Memory or Time: Performance Evaluation for Iterative Operation on Hadoop and Spark*, in 2013 IEEE 10th International Conference on High Performance Computing and Communications & 2013 IEEE International Conference

on Embedded and Ubiquitous Computing, IEEE, pp. 721–727, 2013, doi: 10.1109/HPCC.and.EUC.2013.106.

[13] S. Han, W. Choi, R. Muwafiq, and Y. Nah, *Impact of Memory Size on Bigdata Processing based on Hadoop and Spark*, in Proceedings of the International Conference on Research in Adaptive and Convergent Systems, New York, NY, USA: ACM, pp. 275–280, 2017 doi: 10.1145/3129676.3129688.

[14] S. Gopalani and R. Arora, *Comparing Apache Spark and Map Reduce with Performance Analysis using K-Means*, Int. J. Comput. Appl., Vol. 113, No. 1, pp. 8–11, Mar. 2015, doi: 10.5120/19788-0531.

[15] T. Sharma, D. V. Shokeen and D. S. Mathur, *Multiple K Means++ Clustering of Satellite Image Using Hadoop MapReduce and Spark*, Int. J. Adv. Stud. Comput. Sci. Eng., Vol. 5, No. 4, pp. 23–31, May 2016, [Online]. Available: http://arxiv.org/abs/1605.01802

[16] X. Lin, P. Wang, and B. Wu, *Log analysis in cloud computing environment with Hadoop and Spark*, Proc. 2013 5th IEEE Int. Conf. Broadband Netw. Multimed. Technol. IEEE IC-BNMT(2013), pp. 273–276, 2013, doi: 10.1109/ICB-NMT.2013.6823956.

[17] A. Mostafaeipour, A. J. Rafsanjani, M. Ahmadi, and J. A. Dhanraj, *Investigating the performance of Hadoop and Spark platforms on machine learning algorithms*, J. Supercomput., vol. 77, no. 2, pp. 1273–1300, 2021, doi: 10.1007/s11227-020-03328-5.

[18] Y. Samadi, M. Zbakh, and C. Tadonki, *Comparative study between Hadoop and Spark based on Hibench benchmarks*, Proc. 2016 Int. Conf. Cloud Comput. Technol. Appl. CloudTech 2016, pp. 267–275, 2017, doi: 10.1109/CloudTech.2016.7847709.

[19] A. Singh, A. Khamparia, and A. K. Luhach, *Performance comparison of Apache Hadoop and Apache Spark*, in Proceedings of the Third International Conference on Advanced Informatics for Computing Research - ICAICR '19, New York, New York, USA: ACM Press, pp. 1–5, 2019. doi: 10.1145/3339311.3339329.

[20] A. V. Hazarika, G. J. S. R. Ram, and E. Jain, *Performance comparision of Hadoop and spark engine*, in 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), IEEE, pp. 671–674, 2017. doi: 10.1109/I-SMAC.2017.8058263.

[21] E. P. S. Castro, T. D. Maia, M. R. Pereira, A. A. A. Esmin, and D. A. Pereira, *Review and comparison of Apriori algorithm implementations on Hadoop-MapReduce and Spark*, Knowl. Eng. Rev., Vol. 33, No. e9, pp. 1–25, Jul. 2018, doi: 10.1017/S0269888918000127.

[22] K. Aziz, D. Zaidouni, and M. Bellafkih, *Real-time data analysis using Spark and Hadoop*, Proc. 2018 Int. Conf. Optim. Appl. ICOA 2018, pp. 1–6, 2018, doi: 10.1109/ICOA.2018.8370593.

[23] Y. Benlachmi, A. El Yazidi, and M. L. Hasnaou, *A Comparative Analysis of Hadoop and Spark Frameworks using Word Count Algorithm*, Int. J. Adv. Comput. Sci. Appl., Vol. 12, No. 4, pp. 778–788, 2021, doi: 10.14569/IJACSA.2021.0120495.

[24] S. Ketu, P. K. Mishra, and S. Agarwal, *Performance Analysis of Distributed Computing Frameworks for Big Data Analytics: Hadoop Vs Spark*, Comput. y Sist., vol. 24, no. 2, pp. 669–686, 2020, doi: 10.13053/CyS-24-2-3401.

[25] M. M. Rathore, H. Son, A. Ahmad, A. Paul, and G. Jeon, *Real-Time Big Data Stream Processing Using GPU with Spark Over Hadoop Ecosystem*, Int. J. Parallel Program., vol. 46, no. 3, pp. 630–646, 2018, doi: 10.1007/s10766-017-0513-2.

[26] H.Singh and S. Bawa, *Predicting Covid-19 statistics using machine learning regression models Li-MuLi-Poly*, Multimedia Systems, Vol. 28, pp. 113-120, 2022, doi: 10.1007/s00530-021-00798-2.

[27] M. M. George and P. S. Rasmi, *Performance Comparison of Apache Hadoop and Apache Spark for COVID-19 data sets*, 2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 1659–1665, Feb. 2022, doi: 10.1109/ICSSIT53264.2022.9716232.

[28] J. Dean and S. Ghemawat, *MapReduce: Simplified Data Processing on Large Clusters*, Commun. Acm, Vol. 51, No. 1, pp. 107–113, 2008, doi: 10.1145/1330000/1327492.

[29] S. Shahrivari, *Beyond batch processing: Towards real-time and streaming big data*, Computers, Vol. 3, No. 4. MDPI AG, pp. 117–129, Dec. 01, 2014. doi: 10.3390/computers3040117.

[30] S. Salloum, R. Dautov, X. Chen, P. X. Peng, and J. Z. Huang, *Big data analytics on Apache Spark*, International Journal of Data Science and Analytics, Vol. 1, No. 3–4. Springer International Publishing, pp. 145–164, 2016. doi: 10.1007/s41060-016-0027-9.

[31] J. G. Shanahan and L. Dai, *Large Scale Distributed Data Science using Apache Spark*, in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM, pp. 2323–2324, 2015. doi: 10.1145/2783258.2789993.

[32] S. Pan, *The Performance Comparison of Hadoop and Spark*,Culminating Proj. Comput. Sci. Inf. Technol. 7, 2016, [Online]. Available: https://repository.stcloudstate.edu/csit_etds/7/

[33] Apache SparkTM, *Unified Analytics Engine for Big Data*, https://spark.apache.org/ (accessed Jan. 05, 2023).

[34] J. Shi et al., *Clash of the titans: Mapreduce vs. spark for large scale data analytics*, Proc. VLDB Endow., Vol. 8, No. 13, pp. 2110–2121, 2015, doi: 10.14778/2831360.2831365.

[35] Y. Liu and W. Wei, *A Replication-Based Mechanism for Fault Tolerance in MapReduce Framework*, Math. Probl. Eng., Vol. 2015, pp. 1–7, 2015, doi: 10.1155/2015/408921.

[36] Apache Hadoop, *https://hadoop.apache.org/ (accessed Dec. 23, 2023).*

[37] H. Singh and S. Bawa, *A MapReduce-based scalable discovery and indexing of structured big data*, Futur. Gener. Comput. Syst., Vol. 73, pp. 32–43, Aug. 2017, doi: 10.1016/j.future.2017.03.028.

[38] B. Saraladevi, N. Pazhaniraja, P. V. Paul, M. S. S. Basha, and P. Dhavachelvan, *Big data and Hadoop-A study in security perspective*, Procedia Comput. Sci., Vol. 50, pp. 596–601, 2015, doi: 10.1016/j.procs.2015.04.091.

[39] V. K. Vavilapalli et al., *Apache Hadoop YARN*, in Proceedings of the 4th annual Symposium on Cloud Computing, New York, NY, USA: ACM, Oct. 2013, pp. 1–16. doi: 10.1145/2523616.2523633.

[40] K. Aziz, D. Zaidouni, and M. Bellafkih, *Big Data Processing using Machine Learning algorithms: MLlib and mahout use case*, in ACM International Conference Proceeding Series, 2018, pp. 2–7. doi: 10.1145/3289402.3289525.

[41] M. Assefi, E. Behravesh, G. Liu, and A. P. Taft, *Big data machine learning using apache spark MLlib*, Proc. - 2017 IEEE Int. Conf. Big Data, Big Data 2017, Vol. 2018-Janua, pp. 3492–3498, 2017, doi: 10.1109/BigData.2017.8258338.
[42] X. Meng et al., *MLlib: Machine learning in Apache Spark*, J. Mach. Learn. Res., Vol. 17, pp. 1–7, 2016.
[43] A. Sarkar, J. Guo, N. Siegmund, and S. Apel, *Cost-Efficient Sampling for Performance Prediction of Configurable Systems*, 2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE), pp. 342–352, 2015, doi: 10.1109/ASE.2015.45.
[44] M. Last, *Improving data mining utility with projective sampling*, Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., pp. 487–495, 2009, doi: 10.1145/1557019.1557076.

Edited by: Chiranji Lal Chowdhary
Special issue on: Scalable Machine Learning for Health Care: Innovations and Applications
Received: Oct 12, 2023
Accepted: Mar 12, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1387–1401, DOI 10.12694/scpe.v25i3.2691

CLASSIFICATION OF COVID-19 USING DIFFERENTIAL EVOLUTION CHAOTIC WHALE OPTIMIZATION BASED CONVOLUTIONAL NEURAL NETWORK

D.P. MANOJ KUMAR[*], SUJATA N PATIL[†], PARAMESHACHARI BIDARE DIVAKARACHARI [‡], PRZEMYSŁAW FALKOWSKI-GILSKI[§] AND R. SUGANTHI[¶]

Abstract. COVID-19, also known as the Coronavirus disease-2019, is an transferrable disease that spreads rapidly, affecting countless individuals and leading to fatalities in this worldwide pandemic. The precise and swift detection of COVID-19 plays a crucial role in managing the pandemic's dissemination. Additionally, it is necessary to recognize COVID-19 quickly and accurately by investigating chest x-ray images. This paper proposed a Differential Evolution Chaotic Whale Optimization Algorithm (DECWOA) based Convolutional Neural Network (CNN) method for identifying and classifying COVID-19 chest X-ray images. The DECWOA based CNN model improves the accuracy and convergence speed of the algorithm. This method is evaluated by Chest X-Ray (CXR) dataset and attains better results in terms of accuracy, precision, sensitivity, specificity, and F1-score values of about 99.89%, 99.83%, 99.81%, 98.92%, and 99.26% correspondingly. The result shows that the proposed DECWOA based CNN model provides accurate and quick identification and classification of COVID-19 compared to existing techniques like ResNet50, VGG-19, and Multi-Model Fusion of Deep Transfer Learning (MMF-DTL) models.

Key words: Chest X-ray Images, Convolutional Neural Network, COVID-19, Inertia Weight, Residual Blocks.

1. Introduction. The Transfer Learning (TL) is a popular method for developing deep learning models. In the TL, the neural networks are trained in double phases, such as pretraining and fine-tuning [1]. In the first phase, the network is trained commonly on the huge-scale standard dataset which presents an extensive variety of classes [2]. In the next stage, the pretrained network is again trained on the precise target which contains some branded samples than the pretrained dataset [3]. This pretrained stage is useful for the network to learn common features which are to be reprocessed on the target task [4]. These two categories are enormously widespread in numerous situations specifically in therapeutic images. In TL, the benchmark structures are considered for ImageNet with matching pre-trained weights that are fine-tuned on clinical tasks extending from the COVID-19 diagnosis [5]. COVID-19 is considered by huge transmittable disease and death rate. Every country has employed various productive measures to the safety of their citizens [6].

The major promising study parts in the healthcare domain and the technical group are concentrated on medical applications like creating of Computer-Aided Diagnosis (CAD) system for chest X-ray images [7]. By utilizing transfer learning, healthcare experts can influence the knowledge and proficiency within these pre-trained models and apply it to several healthcare tasks such as disease diagnosis, prediction, and medical image analysis [8]. This technique saves time and improves the accuracy and efficiency of healthcare systems [9]. COVID-19 is a transportable disease produced by the SARS-CoV-2 virus. It rapidly circulates and a wide variety of people endure and die from this unive rsal pandemic [10]. Coronavirus is a large family virus and SARS-CoV-2 is a ribonucleic acid (RNA) virus which comes under coronaviruses [11]. COVID-19 is identified over various approaches such as chest X-ray, positive pathogenic test, CT images, epidemiological history and medical syndromes such as pneumonia, cough, dyspnea and fever [12]. The utilization of chest X-ray and CT scans is beneficial in the premature treatment and diagnosis of the disease. The advantage of using these CT

[*] Department of Computer Science and Engineering, Kalpataru Institute of technology, Tiptur -572201, India

[†] Radiation Oncology Department (Pathology lab), Thomas Jefferson University Philadelphia, PA, 19107, USA

[‡] Department of Electronics and Communication Engineering, Nitte Meenakshi Institute of Technology, Bengaluru 560064, Visvesvaraya Technological University, Belagavi, India. (paramesh@nmit.ac.in).

[§] Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Narutowicza 11/12, 80-233 Gdansk, Poland

[¶] Department of Electronics and Communication Engineering, Panimalar Engineering College, Chennai, India

and Chest X-rays is high speed, cost-effectiveness, and a wide range of applications [13]. The detection of COVID-19 in X-ray images is a difficult task because of the requirements of trademark and the accuracy is amplified by the process of segmentation. In recent times, methods based on deep learning has majorly utilized in medical image classification. There is a chance for acquiring much noises while obtaining the X-ray images. So, the technique of noise removal is important for decreasing the noise. The selection of features plays a major role in the part of classification, because it decreases the evaluation time and maximizes the performance of classification. The application of deep learning produces an ideal solution because it extracts many prominent features from whole image than the manually defined features. In the pandemic situation, the critical problem of COVID-19 is the distribution of rapid treatment to the patients. Due to the rapid spread of pandemic, the patients are severely admitted to the hospitals which leads the requirement of faster diagnosis models need to be solved.The major contribution of this manuscript is as follows:

- The preprocessing is done by using the which removes the noise from chest X-ray images and feeds into the Multilevel threshold segmentation process which enhances the training process and minimizes the overfitting issues.
- After the segmentation, the segmented chest X-ray image features are extracted by shape, texture, color and ResNet50 then selected by using DECWOA, and these features are given as input to the CNN classifier.
- By integrating the chaotic nature of WOA enhances the exploration of the solution space which is utilized to find optimal solutions in the high dimensional and complex space. The DECWOA helps in tuning the parameters of CNN to adopt the complex data patterns and enhancing the model performance.

The rest part of research is as follows: Section 2 defines literature review. Section 3 defines details of proposed methodology. Section 4 defines results and discussion and Section 5 defines conclusion and lastly this paper finish with the references.

2. Literature Review. Amin Ul Haq et al. [14] introduced a deep learning (DL) and transfer learning (TL) technique for accurate diagnosis of COVID-19 by employing X-ray images of medical information in healthcare. The developed model utilized a 2D Convolutional Neural Network (2DCNN) to enhance the training process. The TL was pretrained by using ResNet-50 which was transferred to 2DCNN model and fine-tuned through X-ray image. Additionally, data augmentation was employed for the training of the (ResNet-50+2DCNN) multiclassification (R2DCNNMC) model. The developed model utilizes computer vision tasks for effectively classifying the images. However, this model is not suitable for all scenarios due to its computational complexity and data requirements.

Rohit Kundu et al. [15] developed an ensemble of three various transfer learning methods for predicting COVID-19 infection through chest CT images. The bagging or bootstrap aggregation of three example models such as Inception v3, ResNet34 and DenseNet201 are utilized to boost the individual model performance. The developed ET-NET method was estimated on publicly available datasets by using 5-fold cross-validation. The developed model required minimum time for attaining the test output and it has minimum FNR. The developed model is incapable of detecting the current COVID-19 infections from the initial infection stage.

Md. Belal Hossain et al. [16] implemented a TL with fine-tuned ResNet50 to classify COVID-19 from chest images. The developed model was altered by attaching a dual fully connected layer to the actual ResNet50 method by employing fine-tuning. The experiments are conducted on COVID-19 Radiography dataset by applying ten various pre-trained weights trained on large-scale datasets. The developed model obtains better accuracy in classifying normal and COVID patients on chest X-ray images. The developed model accomplished transfer learning on limited clinical image datasets and computing resources.

Soarov Chakraborty et al. [17] introduced a transfer learning approach for classifying pneumonia and COVID-19-affected chest X-ray images by utilizing deep CNN on pre-trained VGG-19 architecture. This developed model utilized a MongoDB database for storing original images and respective classes. The developed model performance was measured for classification of COVID-19, pneumonia-affected, and healthy people from chest images. By utilizing a pre-trained model, it achieves better accuracy. However, MongoDB cannot retain a document file which crosses 16MB in size.

A. Siva Krishna Reddy et al. [18] developed a Multi-Model Fusion of Deep Transfer Learning (MMF-DTL)

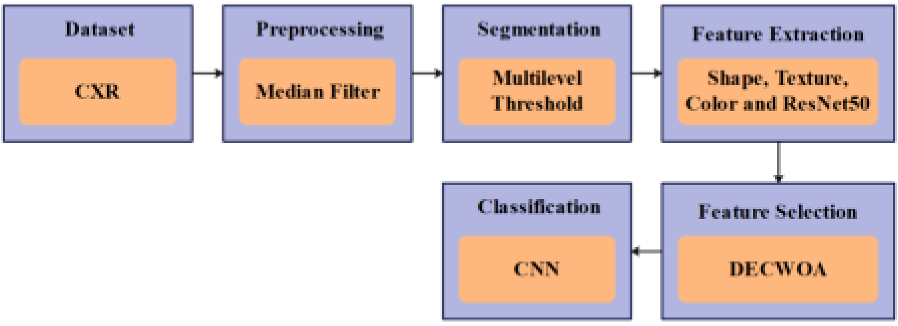


Fig. 3.1: Block Diagram of the Proposed Methodology

approach for the diagnosis and categorization of COVID-19 chest images. The developed model utilized three various DL methods such as Inception v3, VGG16, and ResNet50 for feature selection. The solitary modality was not suitable to obtain an efficient detection rate, the combination of three techniques utilized an MMF to enhance the detection value. At last, the softmax classifier utilized sample images to group six variants. The developed model is utilized to reduce the diagnosis procedure and manage the current epidemic. Hence the developed model cannot be experienced on actual-time images.

Placido L. Vidal et al. [19] suggested a multiple-phase transfer learning for lung segmentation employing transportable X-ray images to COVID-19 patients. The developed model adopts the knowledge from well-known field with numerous samples to a new field with minimized numbers and better complexity. Transfer learning of multiple stages among created consecutive image fields works with a restricted quantity of transportable X-ray models. The advantage of using this model was to train with a huge number of images from the same image field. The limitation of the developed method was established in the images as a minor loss in accuracy and softness in marginal segmentation areas that depend on the image rescaling.

Md. Milon Islam et al. [20] presented a transfer learning based combined Convolutional Neural Network and Recurrent Neural Network (CNN-RNN) technique for COVID-19 diagnosis. The VGG19, InceptionV3, ResNetV2, and DenseNet121 are utilized in this experiment where the CNN was applied to extract the difficult features from samples and RNN was applied to classify them. At last, the images were visualized in the decision-making region by using gradient-weighted class activation mapping. The developed model works only with anterior-posterior view in chest X-ray so, it is unable to classify other views like lordotic, apical, etc.

N. Kumar et al. [21] implemented a deep transfer learning method for detecting COVID-19 patients by utilizing chest X-ray images. The developed method combined several transfer learning methods like Xception-Net, GoogLeNet, and EfficientNet. This model can classify the patient as infected with pneumonia, COVID-19, tuberculosis, or healthy. The developed model employed pre-trained models to extract the features and classify them by utilizing pre-trained models. This model enhances the ability of a classifier for both COVID-19 binary and multiclass datasets. The developed model achieves better diagnosis results with a reduction in errors. However, this model maximizes the training and testing time.

3. Proposed Method. The chest X-ray (CXR) dataset is utilized in this paper which includes 305 images with six different classes. The preprocessing utilizes the median filter which removes the noise from CXR images. The preprocessed data are segmented by using the multilevel threshold image segmentation method. Multilevel thresholding is utilized to choose the attribute to split the image grayscale into more than twofold sets. After segmentation, the features are extracted using shape, texture, color and ResNet50-based feature extraction. The DECWOA is used for selecting the features. The selected features are classified by using Convolutional Neural Network (CNN). The block diagram of the proposed methodology is presented in Figure 3.1.

3.1. Dataset. The dataset used in this analysis is Chest X-Ray (CXR) dataset [22] which is publicly available on the Kaggle. This dataset includes 305 images with six different classes ARDS, COVID-19, No findings, Pneumocystis, SARS, and Streptococcus. The size of every input image is 256×256. These classes

Table 2.1: Key characteristics of previous studies and the proposed solutions

Author	Advantage	Limitation
Amin Ul Haq et al. [14]	The ResNet-50+2DCNN model utilizes computer vision tasks for effectively classifying the images.	However, this model is not suitable for all scenarios due to its computational complexity and data requirements.
Rohit Kundu et al. [15]	The ET-NET model required minimum time for attaining the test output and it has minimum FNR.	The developed model is incapable of detecting the current COVID-19 infections from the initial infection stage.
Md. Belal Hossain et al. [16]	The ResNet50 model obtains better accuracy in classifying normal and COVID patients on chest X-ray images.	The developed model accomplished transfer learning on limited clinical image datasets and computing resources.
Soarov Chakraborty et al. [17]	The VGG-19 model performance was measured for the classification of COVID-19, pneumonia-affected, and healthy people from chest images. By utilizing a pre-trained model, it achieves better accuracy.	However, MongoDB cannot retain a document file that crosses 16MB in size.
A. Siva Krishna Reddy et al. [18]	The MMF-DTL model is utilized to reduce the diagnosis procedure and manage the current epidemic.	Hence the developed model cannot be experienced on actual-time images.
Placido L. Vidal et al. [19]	The U-Net CNN model was utilized to train with a huge number of images from the same image field.	The limitation of the developed method was established in the images as a minor loss in accuracy and softness in marginal segmentation areas that depend on the image rescaling.
Md. Milon Islam et al. [20]	The VGG19-RNN achieves better diagnosis results with a reduction in errors.	The developed model works only with anterior-posterior view in chest X-ray so, it is unable to classify other views like lordotic, apical, etc.
N. Kumar et al. [21]	The ensemble model enhances the ability of a classifier for both COVID-19 binary and multiclass datasets.	However, this model maximizes the training and testing time.
Proposed methodology	By integrating the chaotic nature of WOA enhances the exploration of the solution space which is utilized to find optimal solutions in the high dimensional and complex space. The DECWOA helps in tuning the parameters of CNN to adopt the complex data patterns and enhancing the model performance.	

Table 3.1: Description of CXR Dataset

Classes	ARDS	COVID-19	No findings	Pneumocystis	SARS	Streptococcus
Labels	0	1	2	3	4	5
No of Images	15	220	27	15	11	17

along with corresponding labels and number of images are represented in Table 3.1.The Figure 3.2 presents the sample dataset images.

3.2. Preprocessing. The median filter is a process of nonlinear method which removes the noise [23] from chest X-ray images. Median filter process by shifting pixel in image, modifying every value with median value of adjacent pixel. The pixel is measured by dividing whole values of pixel from neighborhood pattern into the mathematical order and modifying pixel which is considered as average value of pixel. Median filter removes

Fig. 3.2: Sample dataset image

the noise effectively without minimizing the image sharpness which is represented in Eq. 3.1,

$$f(x,y) = median\{g(s,t)\}, where(s,t) \in S_{xy} \tag{3.1}$$

where, S_{xy} represents a group of coordinates in rectangular image window that contains center at (x,y). The $f(x,y)$ is a restored image, $g(s,t)$ is a calculated and corrupted area under S_{xy}.

3.3. Segmentation. The preprocessed data are segmented by using multilevel threshold image segmentation method. Applying multiple thresholds enables the enhancement of contrast a visibility of different structures within the chest X-ray. The thresholding is utilized to choose the attribute to split the image grayscale into more than double sets. It is generally established based on histograms produced by gray-level images. The image is unable to be distributed by noise and histogram of a segmented image has more than two peaks during an ideal condition. Then the threshold is set at trough and'an image is divided into numerous objects and backgrounds. The image is distributed in various noises in an actual picture and image grayscale data is not accurate. There is no peak on the histogram then it is distributed by using noise. The result of image segmentation with a threshold on troughs is incorrect or deprived. In this paper, Kapur's entropy MIS is utilized by non-local means 2D-histogram. When an image is polluted by noise, this technology efficiently minimizes the noise interface and it has a better segmentation effect. The particular procedure is to achieve grayscale image according to an actual image and accomplish a non-local mean noise decrease process on a gray scale image and attained image is known as NLM image. Next, the 2D -histogram is attained corresponding NLM and grayscale images. The highest Kapur's entropy evaluation is performed based on a 2D-histogram then the threshold set according to the highest entropy at last image segmentation is performed with corresponding thresholds. The three-majority image segmentation calculation approaches utilized in this paper which is illustrated in the following section.

3.3.1. Kapur's Entropy. Kapur's entropy is according to the image gray scale which is stored in 8 bits and the range of gray value 0 to 255 . Consider L $= 256, n_i$ is the pixel number grayscale is i. Kapur's entropy H is represented by using Eq. 3.2, Eq. 3.3 and Eq. 3.4.

$$N = \sum_{i=0}^{L-1} n_i \tag{3.2}$$

$$p_i = \frac{n_i}{N} \tag{3.3}$$

$$H = -\sum_{i=0}^{L-1} p_i \ln p_i \tag{3.4}$$

where p_i is incidence of gray scale probability i. For Multilevel threshold Image Segmentation (MIS), images are separated into m subclasses. Where, the $C_0 = \{1, 2, \ldots, t_1 - 1\}, C_1 = \{t_1, \ldots t_2 - 1\}, C_2 = \{t_2, \ldots t_3 - 1\}, \ldots,$ $C_{m-1} = \{t_{m-1}, \ldots L - 1\}$ then the Kapur's entropy H_c is represented by using Eq. 3.5, Eq. 3.6, Eq. 3.7 and Eq. 3.8,

$$H_C = \sum_{i=0}^{m-1} H_{C_i} \tag{3.5}$$

$$H_{C_i} = -\sum_{j=t_i}^{t_{i+1}-1} \frac{p_j}{\omega_i} \ln \frac{p_j}{\omega_i} \tag{3.6}$$

$$\omega_i = \sum_{n=t_i}^{t_{i+1}-1} p_j \tag{3.7}$$

$$t^* = \text{argMax}\left(H_c\right) \tag{3.8}$$

where the t^* separates the point set when H_c take the highest value which is determined by the threshold.

3.3.2. Non-local means 2D histogram. The non-local means 2D histogram procedure is utilized for maintaining and denoising the highest features of an image. Consider the original image is O, denoising image is N, grayscale of pixel p in image O is represented as $O(p)$. The non-local mean filtering a gray scale $N(p)$ of pixel p is attained by the Eq. 3.9, Eq. 3.10, Eq. 3.11 and Eq. 3.12,

$$N(p) = \frac{\sum_{q \varepsilon O} O(q)\omega(p, q)}{\sum_{q \varepsilon O} \omega(p, q)} \tag{3.9}$$

$$\omega(q, p) = \exp^{-\frac{|\mu(p)-\mu(q)|^2}{\partial^2}} \tag{3.10}$$

$$\mu(p) = \frac{\sum_{i \varepsilon O(p)} O(i)}{m \times m} \tag{3.11}$$

$$\mu(q) = \frac{\sum_{i \varepsilon O(q)} O(i)}{m \times m} \tag{3.12}$$

where, $\omega(q, p)$ is the pixel p and q weights, $L(p)$ and $L(q)$ is the local and centered image on pixel p and q respectively, $\mu(p)$ and $\mu(q)$ are local mean of pixels p and q and ∂ is the standard deviation. By combining the grayscale image O and denoising image N, the 2D view of the histogram is generated and non-local means 2D-histogram is formulated as Eq. 3.13,

$$P_{ij} = \frac{h_{ij}}{m \times n} \tag{3.13}$$

where i and j denotes the pixel rate of image $O(x, y)$ and $N(x, y), h_{ij}$ represents the number of times occurs at gray scale vector (s, t), the pixel size in the image is $m \times n$.

3.3.3. Kapur's entropy-based 2D histogram. The transverse of 2D histogram includes suitable image data. The optimal solution found by $\{t, t_2 \dots t_n - 1\}$ is the optimal threshold. This manuscript estimates Kapur's entropy as an objective function and subareas of major diagonal by utilizing Eq. 3.14,

$$H(s,t) = -\sum_{i=0}^{s_1}\sum_{j=0}^{t_1} \frac{P_{ij}}{P_1} \ln \frac{P_{ij}}{P_1} - \sum_{i=s_1+1}^{s_2}\sum_{j=t_1+1}^{t_2} \frac{P_{ij}}{P_2} \ln \frac{P_{ij}}{P_2} \dots - \sum_{i=s_{L-2}+1}^{s_{L-1}}\sum_{j=t_{L-2}+1}^{t_{L-1}} \frac{P_{ij}}{P_{L-1}} \ln \frac{P_{ij}}{P_{L-1}} \tag{3.14}$$

3.4. Feature Extraction. After the segmentation process, features are extracted using shape, texture, color and ResNet50-based feature extraction. By using this technique, the CXR dataset is much more informative. The shape captures geometric properties, texture accounts for surface patterns, color represents visual information and RsNet50 captures deep learning features.

3.4.1. Shape-based Feature Extraction. The shape-based feature extraction is primarily performed to the shape properties like region, moment, and boundary in the image. Such extraction simplifies the transmission, identification, recognition, and comparison of the shape. Shape-based features must be robust to conversion, scaling, and rotation. In this concern, there is no arithmetical transformation is implicated in shape features. The image has three values in every pixel and in the shape feature extraction the color image is converted into the greyscale images. For this purpose, the Eq. 3.15 is introduced by Craig as shown below:

$$I_g = [I_r I_g I_b] \times 0.29890.5870.114 \tag{3.15}$$

where I_g is the grey-level image, $I_r I_g I_b$ is the component of the color. Neurotrophic clustering is used to divide pixels with the nearest value and avoid pixels from greyscale images.

3.4.2. Texture-based Feature Extraction. The texture-based feature extraction is based on extracting several features from a Grey Level Co-occurrence Matrix (GLCM) model. This GLCM is an effective and robust methodology that analyzes images and it is represented as a combined 2D matrix between pixels and pairs with distance d and the direction θ. To classify the texture features, 14 features such as contrast, correlation, energy, homogeneity, angular second moment, sum of squares variance, inverse different moment, sum average, sum variance, sum entropy, entropy, difference variance, difference entropy and information measure of correlation are extracted from the GLCM method. The texture-based feature extraction using GLCM is presented in the following steps:

- The colored image is converted into a greyscale image.
- Input image is filtered by using 5×5 matrix of Gaussian filter.
- The filtered image is separated into 4×4 matrix blocks.
- The GLCM evaluates every block of energy, contrast, mean value, standard deviation, and homogeneity. The evaluation is done with four directions of these features namely vertical, horizontal, and diagonal direction.
- These extracted features are stored in a database.

3.4.3. Color-based Feature Extraction. The color space demonstrates the color with the appearance of intensity value, it can visualize, specify and create the color through the color space technique. There are three various techniques in color-based feature extraction. The first is the histogram intersection (HI) method which examines global color features. In the HI method, the number of bins creates effects on the performance. The huge number of bins shows the image is difficult and it enhances the computational complexity. The second is Zernike Chromaticity Distribution moments which is derivative from chromaticity space. This model provides accurate length and efficient computation presentation of an image that includes the color of the image but the size differs under flipping and rotation. The third is color histogram that illustrates the image from various perceptions. In this, the regularly distributed color bin is presented and it counts the pixels that are the same and keeps it.

3.4.4. ResNet50-based Feature Extraction. The ResNet utilized the residual block for solving gradient vanishing and degradation problems that happen in CNN model. The residual block increases network potency and performance. It is proficient of generating best output in classifications. The residual block of this model implements residual below an inclusion of current residual block and replication overcome of the residual block. The residual function is presented in Eq. 3.16,

$$y = F(x, W) + x \tag{3.16}$$

where x, y, and W denote the input, output, and weight of the residual block. It contains various residual blocks in which a kernel size of convolution layer has differed. After feature extraction, the extracted features are selected by the DECWOA model.

3.5. Feature Selection. After extracting features, the DECWOA is used for feature selection. The DECWOA is utilized to overcome the limitations of the WOA [24] like low convergence speed, insufficient ability of global optimization, and easy fall into local optimization. In DECWOA, initial population is produced through presenting Sine Chaos theory at start of an algorithm for improving popularity diversity. Then, new adaptive inertia weight is familiarized into whale individual position update equation to untrained global search and enhancing the performance of optimization. At last, a Differential Evolution algorithm DEA is presented to improve WOA accuracy and global search speed.

3.5.1. Sine Mapping Population Initialization. The population initialization technique affects the accuracy and convergence speed of a particular algorithm. The WOA is utilized as initial random populations in the nonappearance of appropriate experimental data which results in incapability to confirm that whales are regularly circulated through the solution space. Chaotic mapping produces sequences randomly from inevitable schemes that are stochastic and ergodic. 1D chaotic mapping like sine mapping and logistic mapping has a simple structure and high computational speed. Thus, sine chaos is utilized for population initialization of WOA. The sine chaos self-mapping is formulated in Eq. 3.17,

$$x_{n+1} = \sin\left(\frac{2}{x_n}\right), \quad x = 0, 1, \ldots, N \tag{3.17}$$

where x_n represents primary value which cannot be 0 , by evading zero point and immobility range of $[-1, 1]$. At the iteration, the scheme results traverse every solution space.

3.5.2. Adaptive Inertia Weights. The inertia weight parameter is significant in WOA and a persistent inertia weight minimizes an algorithm's effectiveness which cannot behelpful for global optimization algorithms. The maximum inertia weights are helpful for global optimum and minimum inertia weights are helpful for local mining. The ideal inertia weight contains some features: At initial iteration, it has maximum weights which ensures the algorithm has a robust global search ability. At the final iteration, it has minimum weights which ensures the algorithm has a robust local search ability. Thus, the inertia weight helps balance the local and global exploitation capability. The adaptive inertia weight ω is presented in Eq. 3.18,

$$\omega = 0.5 + \exp\left(\frac{-f_{\text{fit}}\ (x)}{u}\right)^t \tag{3.18}$$

where, $f_{fit}(t)$ is the whale x fitness value, t is the present number of iterations and u is the greatest fitness score at the initial iteration of evaluation. The property ω of dynamic nonlinear is applied to the degree of effect of the new position. The weighted update is formulated in Eq. 3.19 and Eq. 3.20,

$$\left\{\begin{array}{l} X(t+1) = \omega \cdot X_{\text{best}}\ (t) - A \cdot D_{\text{dist}} \\ D_{\text{dist}}\ = |C \cdot X_{\text{best}}\ (t) - X(t)| \end{array}\right. \tag{3.19}$$

$$\left\{\begin{array}{l} X(t+1) = O_{\text{dist}}\ \cdot \cos(2\pi l) \\ W = \omega \cdot X_{\text{best}}\ (t) \\ D_{\text{dist}}\ = |X_{\text{best}}\ (t) - X(t)| \end{array}\right. \tag{3.20}$$

The minimum adaptation scores ensure that it has highest inertia weight and maximum adaptation scores ensure it has a minimum inertia weight that is helpful for the performance of global optimum.

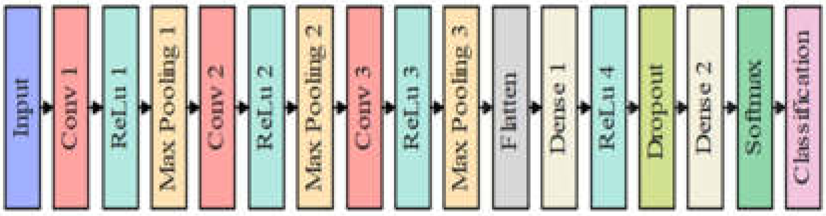


Fig. 3.3: Architecture of CNN

3.5.3. Differential Evolutionary Algorithm. The DEA includes three different processes such as variation, crossover and selection. Three controller parameters are there, such as differential variation parameter F, crossover probability CR and population size. Initially, the DE provides a new variance vector production measured by F. Then, crossover operation among target vector and variance is established, and the trial vector is produced. Atlast, the greedy selection is accomplished on target and trial vector then select an individuals with best fitness and come to next iteration procedure. After population initialization, three equally various vectors X_{r1}, X_{r2}, X_{r3} are randomly selected and a new variation vector is produced which is represented in Eq. 3.21,

$$V_i = X_{r1} + F \cdot (X_{r2} - X_{r3}) \tag{3.21}$$

where the F is a variance vector which is a random number within the range of $[0, 1]$. After mutation operation which produces a mutation vector, crossover is accomplished among target vector and variance to produce a test vector. The two crossover techniques are exponential and binomial crossover. Among them, binomial is utilized which is formulated in Eq. 3.22,

$$U'_{i,j} = \begin{cases} V_{i,j}, & \text{rand}_{i,j}[0,1] \leq CR \\ x_{i,j}, & \text{otherwise} \end{cases} \tag{3.22}$$

where $V_{i,j}$ is the i th individual of j th dimension produced in the above steps. $\text{rand}_{i,j}[0, 1]$ are the random number ranges of $[0, 1]$. The CR is a factor of crossover random number range of $[0, 1]$. After generating the test vectors, the fitness scores are related to the target vector. The individual with best fitness score is chosen for the next iteration. The f_{fit} is fitness function and the scientific equation of the selection operation is presented in Eq. 3.23,

$$X_i(t+1) = \begin{cases} U_i(t), & f_{\text{fit}}\ (X_i(t)) \\ X_i(t), & \text{Otherwise} \end{cases} \tag{3.23}$$

The selection process is separated into binary categories such as synchronous and asynchronous selection in which the asynchronous provides better performance than synchronous selection. In asynchronous, after recently produced test vector is associated with a target vector and best test vector instantly exchanges an equivalent target vector in a population. Hence, convergence speed of this algorithm is quicker.

3.6. Classification. The selected features are classified by using the Convolutional Neural Network (CNN) model which provides significant results in various areas such as image processing, Natural Language Processing (NLP) and diagnosis systems. The Multi-Layer Perceptron (MLP) and CNN decrease the number of parameters and neurons that results in quick adaptation with minimum complexity. The CNN has important applications in clinical image classification. The CNN is a kind of Feed-Forward Neural Network (FFNN) and DL model [25]. The convolution operation captures convention invariance that means the filter is independent in position that decreases a number of parameters. The CNN has three types of layers Convolution, pooling, and Fully Connected FC layers. Figure 3.3 represents CNN architecture.

These layers are essential for accomplishing dimensionality reduction, feature extractions and classifications. Through forward pass of a convolution operation, the filter is slid on computers and the input capacity of an activation map evaluates point-wise result of every score added and obtains the activation. The sliding filter is employed by linear and convolution operators, it is stated as a quick distribution of dot product. Consider w is the kernel function, x is the input, $(x \times w)(a)$ on time t is formulated as Eq. 3.24,

$$(x \times w)(a) = \int x(t)w(a-t)da \tag{3.24}$$

where a is R^n for each $n \geq 1$. The parameter t is the discrete which is presented in Eq. 3.25,

$$(x \times w)(a) = \sum_a x(t)w(t-a) \tag{3.25}$$

In this paper, the CNN is utilized for multi-classification problems. The 2D image I as input, K is a 2D kernel and convolution is formulated as Eq. 3.26,

$$(I \times K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n) \tag{3.26}$$

To improve the non-linearity, two different activation functions are utilized ReLU and softmax. The ReLU is represented as Eq. 3.27,

$$\text{ReLU}(x) = \max(0,x)x \in R \tag{3.27}$$

The gradient $\text{ReLU}(x) = 1$ for $x > 0$ and $\text{ReLU}-(x) = 0$ for $x < 0$. The ReLU convergence ability is better than the sigmoid non-linearities. The next layer is softmax, it is preferable when the result requires to include two or more classes which is mathematically formulated as Eq. 3.28,

$$\text{softmax}\,(x_i) = \frac{\exp\,(x_i)}{\sum_j \exp\,(x_i)} \tag{3.28}$$

The pooling layers are applied to result in a statistic of input and rescale the structure of output without missing essential data. There are various types of pooling layers, this paper utilized the highest pooling which individually produces large values in a rectangular neighbor of individual points (i,j) in 2D information for every input feature correspondingly. The FC is a last layer with m and n output and input are illustrated. The parameter of the output layer is stated as weight matrix $W \in M_{m,n}$. Where m and n an rows and columns and the bias vector $b \in R^m$. Consider as an input vector $x \in R^n$, the FC layer output through an activation function f is formulated as Eq. 3.29,

$$FC(x) := f(Wx = b) \in R^m \tag{3.29}$$

where Wx is the matrix product while function f is employed as a component. This fully connected layer is applied for classification difficulties. The FC layer of CNN is commonly involved at the topmost level. The CNN production is compressed and displayed as a single vector.

4. Experimental Result. In this paper, the proposed Whale Optimization Algorithm (WOA) based Convolutional Neural Network (CNN) model is stimulated by utilizing a python environment with the system configuration: RAM:16GB, processor: intel core i7 and operating system: windows 10. The parameters like accuracy, precision, sensitivity, specificity and F1-score are utilized to evaluate the model performance. The mathematical representation of these parameters is shown in Eq. 4.1, Eq. 4.2, Eq. 4.3, Eq. 4.4 and Eq. 4.5,

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4.1}$$

$$\text{Precision } = \frac{TP}{TP+FP} \tag{4.2}$$

Table 4.1: Memory usage

Image Size	RAM (GB)	
	With Feature Selection	Without Feature Selection
50	3.5	3.0
100	4.0	4.5
150	4.0	4.0
200	4.5	5.0
250	4.5	5.0
305	5.0	5.5

Table 4.2: Performance of Optimization Algorithm

Methods	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
PSO	85.83	85.52	85.34	85.04	85.75
GJO	87.62	87.41	87.29	87.59	87.31
ABC	88.71	88.63	88.52	88.49	88.09
WOA	90.98	90.54	90.61	90.47	90.76
DECWOA	92.86	92.73	92.29	92.01	92.57

$$\text{Sensitivity } = \frac{TP}{TP+FN} \tag{4.3}$$

$$\text{Specificity } = \frac{TN}{TN+FP} \tag{4.4}$$

$$F1-\text{score } = 2 \times \frac{\text{Precision } \times \text{sensitivity}}{\text{Precision } + \text{sensitivity}} \tag{4.5}$$

where TP, TN, FP and FN illustrate the True Positive, True Negative, False Positive and False Negatives respectively.

4.1. Quantitative Analysis. This section shows the quantitative analysis of the DECWOA model in terms of accuracy, precision, sensitivity, specificity and f1-score are shown in Table 4.1 4.2, 4.3 and 4.4. Table 4.1 illustrates the memory usage of various image size in terms of with and without feature selection. Table 4.2 illustrates the quantitative analysis of various optimizations by employing chest X-ray images dataset. Table 4.3 illustrates the quantitative analysis of various classifiers with default features. Table 4.4 illustrates the quantitative analysis of various classifiers after feature selection.

Table 4.2 and Figure 4.1 represent the performance of the optimization algorithm by using performance metrics like accuracy, precision, sensitivity, specificity and f1-score. The performance of Particle Swarm Optimization (PSO), Golden Jackal Optimization (GJO), Artificial Bee Colony (ABC) and Whale Optimization Algorithm (WOA) are compared with DECWOA. The attained result displays that the DECWOA attains an accuracy of 92.86%, precision of 92.73%, sensitivity of 92.29%, specificity of 92.01%, and f1-score of 92.57% which is comparatively higher than the existing optimization algorithms.

Table 4.3 and Figure 4.2 represents the performance of classification with default features by using performance metrics like accuracy, precision, sensitivity, specificity and f1-score. The performance of Artificial Neural Network (ANN), K-Nearest Neighbour (KNN), Support Vector Machine (SVM) and Deep Neural Network (DNN) are compared with CNN model. The attained result displays that the CNN model attains accuracy of 93.97%, precision of 93.57%, sensitivity of 93.71%, specificity of 93.49% and f1-score of 93.85% which is higher than the existing classifiers.

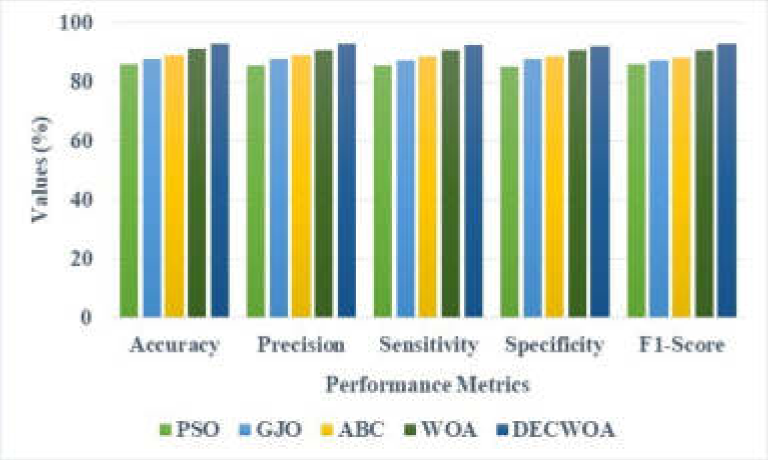


Fig. 4.1: Performance of Optimization Algorithm

Table 4.3: Performance of classification with default features

Methods	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
ANN	84.53	84.21	84.01	84.33	84.49
KNN	87.91	87.02	87.63	87.47	87.89
SVM	89.76	89.24	89.54	89.21	89.72
DNN	91.63	91.27	90.49	91.03	91.43
CNN	93.97	93.57	93.71	93.49	93.85

Table 4.4 and Figure 4.3 represent the performance of classification after feature selection by using evaluation metrics like accuracy, precision, sensitivity, specificity, and f1-score. The performance of ANN, KNN, SVM and DNN are compared with DECWOA-CNN model. The attained result displays that the DECWOA-CNN model attains accuracy of 99.89%, precision of 99.83%, sensitivity of 99.81%, specificity of 98.92% and f1-score of 99.26% which is comparatively higher than the existing methods.

4.2. Comparative Analysis. This section illustrates the comparative analysis of the proposed DECWOA-CNN model with performance metrics like accuracy, precision, sensitivity, specificity and f1-score as shown in Table 4.5. The existing result such as [15] [16], [17], [18] and [20] are utilized for estimating an ability of the classifier. The DECWOA-CNN is trained, tested and validated by using CXR dataset. The result obtained from Table 4.5 shows that the DECWOA-CNN attains better performance when compared with the existing methods. The accuracy was improved to 99.89%, precision of 99.83%, sensitivity of 99.81%, specificity of 98.92% and f1-score of 99.26%.

4.2.1. Discussion. In this section, the advantages of the proposed method and the limitations of existing methods are discussed. The existing method has some limitations such as the ET-NET [15] was incapable of detecting the current COVID-19 infections from the initial infection stage. The ResNet50 [16] model carrying out transfer learning on the limited clinical image dataset and computing resources. The VGG-19 [17] model employed MongoDB for storing images, once it crossed 16MB size then it cannot be retained. The MMF-DTL [18] model cannot be tested on real-time images. The VGG19-RNN [20] has works only with anterior-posterior view in chest X-ray so, it is unable to classify other views like lordotic, apical. The proposed DECWOA-CNN model overcomes these existing model limitations. The proposed model improves the accuracy and convergence speed of the optimization algorithm.

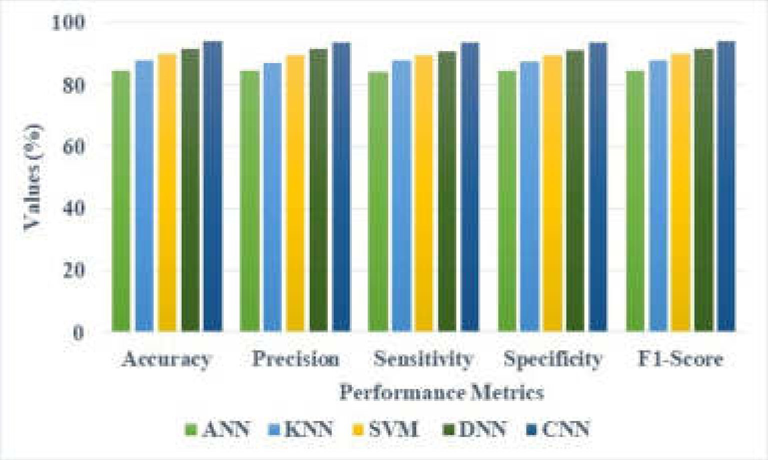


Fig. 4.2: Performance of classification with default features

Table 4.4: Performance of classification after feature selection

Methods	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1- Score (%)
ANN	84.53	84.21	84.01	84.33	84.49
KNN	87.91	87.02	87.63	87.47	87.89
SVM	89.76	89.24	89.54	89.21	89.72
DNN	91.63	91.27	90.49	91.03	91.43
DECWOA-CNN	99.89	99.83	99.81	98.92	99.26

5. Conclusion. This paper proposed a DECWOA-CNN model for accurately and quickly identifying and classifying COVID-19. The median filter is utilized for data pre-processing which removes the noise from chest X-ray images and fed into threshold segmentation for segmenting chest-ray images. The shape, texture, color and ResNet-50 feature extraction is utilized for extracting features from segmented images. The DECWOA is utilized for selecting features and CNN is utilized for classifying COVID-19 from chest images. In DECWOA, initial population is produced by presenting the Sine Chaos theory at the start of an algorithm for improving popularity diversity. Then, the new adaptive inertia weight is familiarized into whale individual position update formulation to untrained global search and enhancing optimization performance. At last, DEA is presented to enhance whale optimization accuracy and global search speed. The proposed DECWOA-CNN model is estimated on CXR dataset and attains better results by using performance metrics like accuracy, precision, sensitivity, specificity, and F1-score values of about 99.89%, 99.83%, 99.81%, 98.92%, and 99.26% respectively. In the future, the parameter tuning will be applied in the optimization algorithm to improve the model performance.

REFERENCES

[1] M. M. Taresh, N. Zhu, T. A. A. Ali, A. S. Hameed, and M. L. Mutar, *Transfer learning to detect covid-19 automatically from x-ray images using convolutional neural networks*, Int. J. Biomed. Imaging, (2021), pp. 1–9.

[2] E. Jangam, A. A. D. Barreto, and C. S. R. Annavarapu, *Automatic detection of COVID-19 from chest CT scan and chest X-Rays images using deep learning, transfer learning and stacking*, Appl. Intell., (2022), pp.1–17.

[3] Y. Brima, M. Atemkeng, S. Tankio Djiokap, J. Ebiele, and F. Tchakounté, *Transfer learning for the detection and diagnosis of types of pneumonia including pneumonia induced by COVID-19 from chest X-ray images*, Diagnostics, 11(8) (2021), p. 1480.

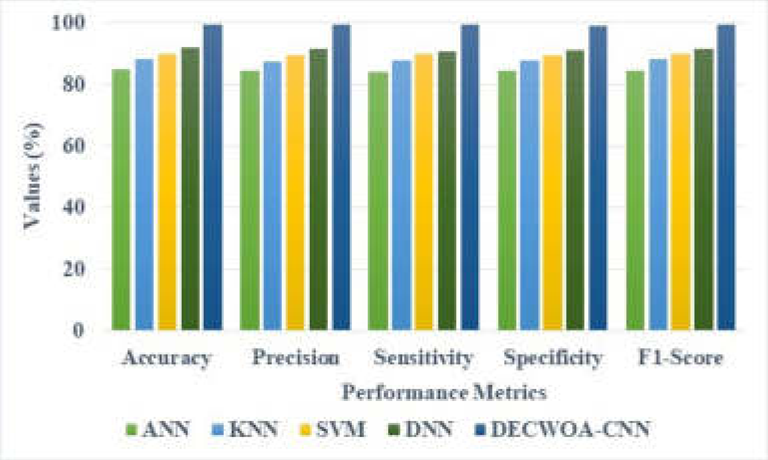


Fig. 4.3: Performance of classification after feature selection

Table 4.5: Comparative Analysis

Author	Dataset	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)	F1-Score (%)
Rohit Kundu et al. [15]	CXR	97.81	97.77	97.81	N/A	97.77
Md. Belal Hossain et al. [16]		99.17	99.31	99.03	N/A	99.17
Soarov Chakraborty et al. [17]		97.11	97	97	N/A	97
A. Siva Krishna Reddy et al. [18]		98.80	93.60	92.96	98.54	93.26
Md. Milon Islam et al. [20]		99.86	99.78	99.78	N/A	99.78
Proposed DECWOA-CNN Model		99.89	99.83	99.81	98.92	99.26

[4] S. Showkat and S. Qureshi, *Efficacy of Transfer Learning-based ResNet models in Chest X-ray image classification for detecting COVID-19 Pneumonia*, Chemometrics and Intelligent Laboratory Systems, 224 (2022), p. 104534.

[5] S. Amin, B. Alouffi, M. I. Uddin, and W. Alosaimi, *Optimizing convolutional neural networks with transfer learning for making classification report in covid-19 chest x-rays scans*, Sci. Program., 2022 (2022).

[6] M. Umair, M. S. Khan, F. Ahmed, F. Baothman, F. Alqahtani, M. Alian, and J. Ahmad, *Detection of COVID-19 using transfer learning and grad-cam visualization on indigenously collected X-ray dataset*, Sensors, 21(17) (2021), p. 5813.

[7] M. T. Naseem, T. Hussain, C. S. Lee, and M. A. Khan, *Classification and Detection of COVID-19 and Other Chest-Related Diseases Using Transfer Learning*, Sensors, 22(20) (2022), p. 7977.

[8] G. Li, R. Togo, T. Ogawa, and M. Haseyama, *COVID-19 detection based on self-supervised transfer learning using chest X-ray images*, Int. J. Comput. Assisted Radiol. Surg., 18(4) (2023), pp. 715–722.

[9] S. Hamida, O. El Gannour, B. Cherradi, A. Raihani, H. Moujahid, and H. Ouajji, *A novel COVID-19 diagnosis support system using the stacking approach and transfer learning technique on chest X-ray images*, J. Healthcare Eng., 2021 (2021).

[10] W. A. Hamwi and M. M. Almustafa, *Development and integration of VGG and dense transfer-learning systems supported with diverse lung images for discovery of the Coronavirus identity*, Inf. Med. Unlocked, 32 (2022), p. 101004.

[11] S. Asif, Y. Wenhui, K. Amjad, H. Jin, Y. Tao, and S. Jinhai, *Detection of COVID-19 from chest X-ray images: Boosting the performance with convolutional neural network and transfer learning*, Expert Syst., 40(1) (2023), p. e13099.

[12] J. Manokaran, F. Zabihollahy, A. Hamilton-Wright, and E. Ukwatta, *Detection of COVID-19 from chest x-ray images using transfer learning*, J. Med. Imaging, 8(S1) (2021), pp. 017503–017503.

[13] S. Agrawal, V. Honnakasturi, M. Nara, and N. Patil, *Utilizing Deep Learning Models and Transfer Learning for COVID-19 Detection from X-Ray Images*, SN Comput. Sci., 4(4) (2023), p. 326.

[14] A. U. Haq, J. P. Li, S. Ahmad, S. Khan, M. A. Alshara, and R. M. Alotaibi, *Diagnostic approach for accurate diagnosis of COVID-19 employing deep learning and transfer learning techniques through chest X-ray images clinical data in E-healthcare*, Sensors, 21(24) (2021), p. 8219.

[15] R. Kundu, P. K. Singh, M. Ferrara, A. Ahmadian, and R. Sarkar, *ET-NET: an ensemble of transfer learning models for prediction of COVID-19 infection through chest CT-scan images*, Multimedia Tools Appl., 81(1) (2022), pp. 31–50.

[16] M. B. Hossain, S. H. S. Iqbal, M. M. Islam, M. N. Akhtar, and I. H. Sarker, *Transfer learning with fine-tuned deep CNN ResNet50 model for classifying COVID-19 from chest X-ray images*, Inf. Med. Unlocked, 30 (2022), p. 100916.

[17] S. Chakraborty, S. Paul, and K. A. Hasan, *A transfer learning-based approach with deep cnn for covid-19-and pneumonia-affected chest x-ray image classification*, SN Comput. Sci., 3 (2022), pp. 1–10.

[18] A. S. K. Reddy, K. B. Rao, N. R. Soora, K. Shailaja, N. S. Kumar, A. Sridharan, and J. Uthayakumar, *Multi-modal fusion of deep transfer learning based COVID-19 diagnosis and classification using chest x-ray images*, Multimedia Tools Appl., 82(8) (2023), pp. 12653–12677.

[19] P. L. Vidal, J. de Moura, J. Novo, and M. Ortega, *Multi-stage transfer learning for lung segmentation using portable X-ray devices for patients with COVID-19*, Expert Syst. Appl., 173 (2021), p. 114677.

[20] M. M. Islam, M. Z. Islam, A. Asraf, M. S. Al-Rakhami, W. Ding, and A. H. Sodhro, *Diagnosis of COVID-19 from X-rays using combined CNN-RNN architecture with transfer learning*, BenchCouncil Trans. Benchmarks Stand. Eval., 2(4) (2022), p. 100088.

[21] N. Kumar, M. Gupta, D. Gupta, and S. Tiwari, *Novel deep transfer learning model for COVID-19 patient detection using X-ray chest images*, J. Ambient Intell. Hum. Comput., 14(1) (2023), pp. 469–478.

[22] Dataset Link : https://www.kaggle.com/datasets/alifrahman/chestxraydataset.

[23] A. Sharma and P. K. Mishra , Image enhancement techniques on deep learning approaches for automated diagnosis of COVID-19 features using CXR images, Multimedia Tools Appl., 81(29) (2022), pp. 42649–42690.

[24] J. Xing, H. Zhao, H. Chen, R. Deng, and L. Xiao, *Boosting whale optimizer with quasi-oppositional learning and gaussian barebone for feature selection and COVID-19 image segmentation*, J. Bionic Eng., 20(2) (2023), pp. 797–818.

[25] D. Zhang, F. Ren, Y. Li, L. Na, and Y. Ma, *Pneumonia detection from chest X-ray images based on convolutional neural network*, Electronics, 10(13) (2021), p. 1512.

Edited by: Chiranji Lal Chowdhary
Special issue on: Scalable Machine Learning for Health Care: Innovations and Applications
Received: Oct 13, 2023
Accepted: Dec 20, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1402–1423, DOI 10.12694/scpe.v25i3.2776

A SURVEY ON AI-BASED PARKINSON DISEASE DETECTION: TAXONOMY, CASE STUDY, AND RESEARCH CHALLENGES

SHIVANI DESAI[*] DEVAM PATEL[†] KAJU PATEL, ALAY PATEL, NILESH KUMAR JADAV, SUDEEP TANWAR, AND HITESH CHHIKANIWALA[‡]

Abstract. Parkinson Disease (PD) is most common diseases from majority of disease encountered all over the world, with more than 7 million individuals being affected. PD is a type of progressive nervous system disease, causing deterioration in health or function. The timely identification of PD is a significant challenge because it rarely shows symptoms in the early stages. Moreover, it is typically encountered in older people where the symptoms sometimes coincide with age-related issues. Deep Learning (DL) can be integrated into many methodologies in diagnosing PD, such as Magnetic Resonance Imaging (MRI) and Single-Photon Emission Computed Tomography (SPECT). DL algorithms can detect PD based on observing some common symptoms. Moreover, it can also be detected using brain MRI images. So, in this study, we reviewed existing DL algorithms for timely identification of PD. We also have developed a CNN model for the timely identification of PD. We used 3D brain MRI images of PPMI datasets and achieved the 88% accuracy.

Key words: PD, DL, MRI, Pre-processing, Datasets

1. Introduction.

Parkinson's Disease (PD) is one of the most frequent worldwide diseases. According to the World Health Organisation (WHO), 7-10 million individuals are diagnosed with PD. In PD patients, men and women are in the ratio of 3:2, and people above the age of fifty develop PD in general [33]. In human brain, Dopamine sends messages between brain cells so that human body motions remain flexible and synchronised [28]. Certain parts of the human brain include dopamine-producing cells. The SN area of the brain contains a high density of these cells. When these cells are destroyed or under-operated, PD symptoms develop. The cause of these neuron disorders is currently unknown. PD is a progressive type of disease of the nervous system. A progressive disease worsens with time, resulting in deterioration in health or function. Insufficient dopamine production leads to muscle tremors, stiffness, and slowdown. A loss in scent sensitivity, muscle tremors while at rest, bending of the body posture, numbness, tingling, discomfort in the limbs, sleep difficulties, constipation, and slow movements are common symptoms in PD [58].

The early detection of PD is challenging for various reasons. Motor symptoms have traditionally been utilised to make the diagnosis of PD. The majority of patients with this disease are above the age of sixty, making it time-consuming and uncomfortable for them to have repeated scans and for neurologists and other specialists in movement disorders to diagnose the condition [11]. To diagnose and measure the severity of PD, medical professionals may use various Radiological Imaging (RI) Studies of the human brain. Some of them are Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), Single-Photon Emission Computed Tomography (SPECT), Structural Magnetic Resonance Imaging (sMRI), Diffusion Tensor Imaging (DTI), etc.

Various RI Studies are crucial in diagnosing and managing PD. These imaging techniques provide essential insights into the anatomical and functional changes taking place in PD patients' brains, allowing for a more accurate diagnosis and aiding in disease progression monitoring. MRI is frequently used as the first-line imaging technique for diagnosing PD. It provides comprehensive structural images of the brain, which can help rule out

[*] Research Scholar, Gujarat Technological University, Ahmedabad, Gujarat, India (shivani.desai@nirmauni.ac.in)

[†] Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad, Gujarat, India (21bce191@nirmauni.ac.in, 21bce201@nirmauni.ac.in, 21bce185@nirmauni.ac.in, 21FTPHDE53@nirmauni.ac.in, sudeep.tanwar@nirmauni.ac.in)

[‡] Department of Information and Communication Technology, Adani Institute of Infrastructure Engineering (AIIE), Ahmedabad, Gujarat, India (drhiteshrc1@gmail.com)

Table 1.1: Commonly used abbreviations

Acronym	**Abbreviations**
PD	Parkinson's Disease
DL	Deep Learning
MRI	Magnetic Resonance Imaging
SPECT	Single-Photon Emission Computed Tomography
CNN	Convolutional Neural Network
PPMI	Parkinson's Progression Markers Initiative
SN	Substantia Nigra
PET	Positron Emission Tomography
DTI	Diffusion Tensor Imaging
DaT	Dopamine Active Transporter
fMRI	Functional MRI
AI	Artificial Intelligence
sMRI	Structural MRI
HC	Healthy Controls
CNNs	Computational Neural Networks
ANN	Artificial Neural networks
CR-ML	Contrast Ratio classifier
RA-ML	Radiomics based classifier

other potential explanations of comparable symptoms, and it can reveal shrinkage of the brain or abnormalities in specific regions, such as the Substantia nigra (SN), which is frequently afflicted in PD. Dopamine Active Transporter (DaT) scan, an example of SPECT, is very useful in diagnosing PD. It assesses the density of dopamine transporters in the basal ganglia, where dopamine insufficiency is a characteristic of PD. Reduced dopamine transporter binding in this area is an important diagnostic marker. PET scans provide dynamic information regarding brain function, such as glucose metabolism and dopamine levels. PET scans can distinguish PD from other movement disorders and follow disease progression, which aids in therapy planning. Functional Magnetic Resonance Imaging (fMRI) measures brain activity and connectivity to detect changes during the disease [10]. Khan *et al.* [30] discussed the use of brain anatomical MRI for diagnostic classification of Alzheimer's disease, which signifies the potential of neuroimaging techniques, such as functional MRI, while examining neurodegenerative diseases like Parkinson's disease.

It enables researchers and physicians to better understand how PD affects different brain regions, offering information on how the disease affects cognition and non-motor symptoms. These scans allow clinicians to make informed treatment decisions and provide vital insights into the evolving nature of the disease, ultimately increasing the quality of care for those living with PD.

The physicians who analyse the patient's data and symptoms must have a thorough understanding of the disease. Unfortunately, many countries do not have enough qualified doctors. As a result, identifying or detecting PD is a difficult process because specialists are stressed as a result of their job. Doctors can describe medications, but they lose effectiveness as the disease worsens from its early stages. As a result, early detection of Parkinson's disease is critical for taking immediate steps to assist people preserve their independence for as long as possible. This explanation has prompted healthcare providers to develop a decision support system based on computer aided diagnosis to assist clinicians in diagnosing PD [35]. This method can serve as a second opinion for PD diagnosis while lowering the likelihood of errors due to the use of Artificial Intelligence (AI) and Deep Learning (DL). [17]

DL also enables the integration of many radiological methodologies, such as MRI and SPECT, in diagnosing PD. By using DL algorithms to identify crucial traits that are typically not used in the clinical diagnosis of PD, we may be able to detect PD in preclinical stages or atypical forms. DL models are increasingly used for early PD detection because they detect minute patterns in Radiological Imaging (RI) data. PD frequently manifests subtle symptoms that are difficult to diagnose ordinarily. DL's ability to analyse varied data sources, such as medical imaging and patient records, aids in identifying early markers and biomarkers. This allows for earlier detection and more effective therapies. These models enable unbiased, dependable data analysis while minimising human errors and biases. They are critical for improving the detection of PD and treatment because of their ability to provide personalised assessments, follow disease progression, and assist ongoing research.

Weng *et al.* [65] used SPECT imaging for diagnosing PD with 99m Tc-TRODAT-1, discusses the age-related depletion in the striatal binding in both PD patients as well as Healthy patients. and concluded that 99mTc-TRODAT-1 SPECT imaging is a helpful and remarkable diagnostic tool for clinical applications in determining the decrease of striatal DAT concentration in PD patients. Bae *et al.* [5] presented some imaging techniques used in PD and other Parkinsonian syndromes. The nigral structure includes markers to help complex neuroimaging procedures using MRI to find pathophysiologic, functional, and neuroanatomical changes. These markers can help diagnose the subtype, monitor disease severity of PD and separate PD from other movement disorders. Noor *et al.* [40] used the model that is based on 3D separable and grouped convolutions to highlight fine-grained descriptive features from sMRI and attained the accuracy of 86% with sensitivity of 87.5% and specificity of 85.7%. Chen *et al.* [14] seeking to develop a model that relies on intra/intervoxel metrics derived from DTI to facilitate automatic differentiation of PD patients without dementia into Mild Cognitive Impairment (MCI) and Normal cognition groups (NC).

This paper mainly focuses on various structural changes in PD and its symptoms. Further, different types of datasets, multiple types of data preprocessing technique which is commonly used and some implemented DL models for early detection of PD are also covered. Moreover, to support the above discussion, one case study is included with the Convolutional Neural Network (CNN) model executed on 3D brain MRI images with 88% accuracy.

1.1. Research Contributions.

The significant research contributions are listed as follows:

- We have discussed pathology, symptoms, and treatment of the Parkinson Disease along with the on brief description on symptoms and causes of the same.
- We have gone through different datasets followed by various data pre-processing methods, extended by DL algorithms for the early diagnosis of PD.
- We have implemented CNN model in our case study to diagnose Parkinson's disease using 3D-MRI from the Parkinson's progression markers Initiative (PPMI) dataset. The model 88% accuracy in predicting the disease.
- At the end we have elaborated the research challenges and future scopes of researches in this domain .

1.2. Article Layout.

The rest of the paper is as follows. Section 2 discusses the fundamental aspects of PD with some commonly observed symptoms. Section 3 gives the overview of the different datasets, various data preprocessing techniques and different DL models, some people applied in the past. Section 4 includes implementing the CNN model on 3D brain MRI images. Section 5 discusses the analysis of performance achieved by the CNN model. Challenges after and before using the DL model for early diagnosing PD are discussed in section 6. Finally, the paper is concluded in section 7.

1.3. Scope of the survey.

Gilat *et al.* [25] assessed the brain activity related to body movement. Gain impairment is the primary disability shown in PD. Even when PD is in the early stage, patients often experience this disability and gait problems can cause other physical issues, leading to severe injuries, immobilisation and mortality. They have reviewed the articles that used PET, SPECT or fMRI, which examined the neurological mechanisms behind gait impairment in PD patients. However, they have only included studies that model gait in PD patients.

Bharti *et al.* [9] discussed existing research studies on structural and functional neuroimaging studies in PD and provided an overview of knowledge regarding Freezing of Gait (FOG) condition. FOG is typically seen in later stages of PD and is generally resistant to medicinal treatment. It contributes to a severe decline in quality of life. They have considered MRI, fMRI, SPECT and PET for brain changes in FOG condition of PD.

Wang *et al.* [63] assessed the diagnostic utility of Neuromelanin sensitive MRI (NMS-MRI) in PD using a meta-analysis technique. They focused on research that examined the Substantia nigra pars compacta (SNpc) structure's signal strength, volume, or area as well as the precise sensitivity and specificity of the cutoff value in order to identify PD.

Cho *et al.* [15] also looked at research on NeuroMelanin-sensitive (NMS-MRI) of clinically diagnosed PD patients or Healthy Controls (HC). They evaluated the diagnostic effectiveness of NeuroMelanin-sensitive (NMS-MRI) for differentiating HC from patients with PD and discovered variables causing heterogeneity, which had an impact on the diagnostic effectiveness of this method across studies. They did not, however, concentrate on the patients' further radiological scans or various MRI variations.

Bergamino *et al.* [6] concentrated on improvements in DTI methodology and only looked at diffusion weighted magnetic resonance imaging (dMRI) tests in early-stage PD. It has been demonstrated that these cutting-edge techniques can identify structural White matter (WM) abnormalities in PD in its early stages. The method most commonly used to examine WM pathological alterations in symptomatic regions like the SN is DTI.

Alzubaidi *et al.* [3] summarised the research studies that contained DL techniques introduced or developed to diagnose PD. They also set some limitations on the types of publications and the research language. Peer-reviewed papers, conference proceedings, reports, theses, and dissertations in English were the only ones approved. They looked into the use of neural network algorithms, notably DL algorithms, for the early detection of PD but did not provide a full evaluation of its quality.

Feraco *et al.* [24] assessed articles containing modern MRI techniques for studying the SN to aid in diagnosing and treating PD. In their research, they used Nigrosome imaging, Neuromelanin sensitive sequences, iron-sensitive sequences, and improved dMRI to characterise SN damage in PD better. These approaches are emerging as promising early diagnostic biomarkers for PD.

Inspired by this, we prepared an exhaustive and comprehensive survey on PD detection that can help medical practitioners and researcher working in the same domain. The survey comprises of fundamental aspects of PD, symptoms, conventional detection mechanism, and its challenges. Further, we presented a thorough taxonomy that shows the AI-based PD detection mechanisms. In addition, a case study has been proposed where AI techniques are used to detect PD to showcase the competency of AI in healthcare domain. Lastly, we highlighted the research challenges that still hinder the performance of PD detection.

2. Background. The section includes the functional aspects of PD, some of its common reasons, its symptoms, some DL model and concludes with current curing methods. The detailed description is as follows.

PD is a complex neurodegenerative disorganisation that seeds unexpected or uncontrolled motions such as shaking, stiffness and unsteadiness and coordination, and the symptoms may worsen over time, which involves a small, dark-colored portion of the brain called the SN. This is a significant production ground for dopamine in the brain. Dopamine is the chemical released in the brain that acts as messenger and transmits messages between nerves, which controls muscle movements, including those of the brain's pleasure and reward centres.

A distinctive trait of PD is the aggregation of exceptional protein known as Lewy bodies in specific brain cells, including dopamine-producing neurons. These Lewy bodies mainly consist of misfolded alpha-synuclein protein. The presence of Lewy bodies intrudes cellular function and facilitates neuronal dysfunction and eventual cell death. Inflammation within the brain, commonly known as neuroinflammation, plays a role in improving PD. The activated immune cells and inflammatory molecules can cause damage to the neurons and even escalate the disease process. Mitochondria, commonly known as the powerhouse of the cells, plays a vital role in maintaining cell health. Dysfunction in mitochondria can lead to oxidative stress and impaired energy production. Oxidative stress, which, in turn, can harm neurons and even cause degeneration. PD can affect the normal secretion of abnormal proteins, including alpha-synuclein, which accumulates protein aggregation, contributing to rupture or damage to the ultimate neural system.

Neurons depend on the effective transport system to move essential molecules along their long axons. The damage caused by PD on the nervous system weakens the delivery process of nutrients and signalling molecules, leading to cell dysfunction and death. The basal ganglia, the part of or the region in the brain which controls the movements, includes motor symptoms that are motor learning. PD mainly affects this part, and the signs are evident when the nerve cells in the basal ganglia become harmed or die. These nerve cells commonly produce a vital brain chemical fluid known as dopamine, so when they are dead, the source of dopamine production becomes less, resulting in movement problems associated with the disease. Reported cases of PD can be connected with specific genetic variants, while some cases are inherited from blood. However, PD is mainly known to have genetic grounds, and the condition is scarcely found in families. Copious experts have

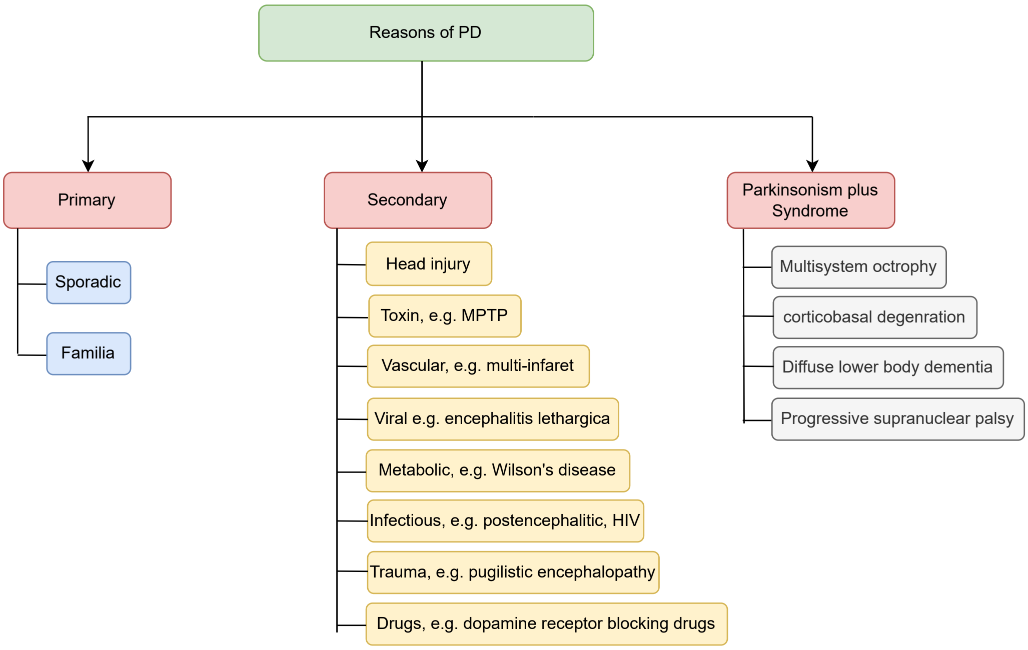

Reasons of PD
Primary
Sporadic
Familia
Secondary
Head injury
Toxin, e.g. MPTP
Vascular, e.g. multi-infaret
Viral e.g. encephalitis lethargica
Metabolic, e.g. Wilson's disease
Infectious, e.g. postencephalitic, HIV
Trauma, e.g. pugilistic encephalopathy
Drugs, e.g. dopamine receptor blocking drugs
Parkinsonism plus Syndrome
Multisystem octrophy
corticobasal degenration
Diffuse lower body dementia
Progressive supranuclear palsy

Fig. 2.1: Causes of PD

now concluded that genetic and environmental factors, including exposure to harmful chemicals, commonly contribute to the cause of PD. Some of the reasons are classified below:

- Primary: The primary reason for PD can be sporadic, which means it is a chronic progressive disorder in which idiopathic PD occur without substantial evidence. The other primary reason is Genetics, which may occur from single genes.
- Secondary: The cause of PD is reduction in dopamine level. The reason for the reduction of dopamine level can be toxins like MPTP(1-methyl 4-phenyl tetrahydropyridine), viral-like encephalitis lethargica, metabolic like Wilson disease, head injury, infectious like postencephalitic, drugs like dopamine receptors blocking drugs, vascular like multi-infact, trauma, etc.
- Parkinsonism plus syndrome: This syndrome can have various types which are multisystem octrophy, corticobasal degeneration, diffuse lower body dementia and progressive supranuclear palsy
- Environmental Factors: Sometimes, there are chances of PD in people of rural areas due to drinking well water, high pesticide exposure, oxidative stress, etc.

The Fig. 2.1 gives the overview of the following causes of PD.

The general estimation says that when almost 80 percent of the dopaminergic cells are lost before the motor symptoms are observed. The degradation of the nigrostriatal dopaminergic system is observed when there is a erosion of more than 70% of the striatum's dopamine and more than 50% erosion of dopaminergic neurons in SN [59]. So early PD detection is sometimes miscellaneous task. Therefore for PD detection some rules or criteria are proposed from the United Kingdom Brain Bank [54]. The below mentioned are criteria are as follows [20]:

- Diagnose of Parkinson's syndrome: When Parkinson is at the initial stages, it can be generally identified with a reduction in movement speed and other day-to-day activities, also known as Bradykinesia [39]. Other symptoms that can be observed are rigidness in muscles, rest tremors of 4 - 6 Hz frequency and sometimes posture instability, which cannot be generally observed at the initial stages.
- Exclusion criteria for PD: Some exclusion symptoms cannot be detected as the PD, though they are present. They are Parkinson's symptoms, which gradually increase after a history of multiple strokes, numerous head traumas in the past, Oculogyric crises, definite encephalitis in the past, beginning of

Table 2.1: Motor and non-motor symptoms of PD

Motor Symptoms	**Non Motor Symptoms**
Bradykinesia	Anosmia
Rest tremor	overtiredness
Rigidity	Low Blood pressure or hypotension
Falls & Dizziness	Bladder & Bowl problems
Freezing	Restless legs
Muscle cramps	Skin & perspiring problems
Micrographia	Speech & communication problem
Masked facies	Eye ailments
Reduced eye blinking	Pain
Drooling	Anxiety
Soft voice	Depression
Dysphagia	Hallucinations & delusions

neuroleptic therapy for symptoms, more than one affected relative, sustained remission, Strictly unilateral features after three years, supranuclear gaze palsy, Cerebellar signs, early severe autonomic involvement, early severe dementia with memory, language, and practical difficulties, Babinski sign, CT(Computerized Tomography) scan showing the presence of a brain tumour or communicating hydrocephalus, negative response to large doses of Levodopa (if malabsorption excluded), MPTP(1-methyl 4-phenyl tetrahydropyridine) exposure.

- Supportive criteria for PD: The criteria in support of the PD are unilateral onset, presence of rest tremor, progressive disorder, continuous asymmetry, which the onset side has the most effect, positive feedback to levodopa drug, severe levodopa-induced chorea, Levodopa response for five years or more, clinical course of 10 years or more.

Based on the above criteria, it can be generally concluded whether the patient has PD. Moving further, PD also shows many symptoms. These symptoms are divided into two broad categories, which further narrow down. Table 2.1 summarizes this symptoms. These broad classifications of the symptoms are as follows:

- Motor Symptoms
- Non-Motor Symptoms

2.1. Motor Symptoms. The symptoms which cause an effect on the movement and balance are known as motor symptoms. The UPDRS(Unified parkinson disease rating scale) score helps in identifying the extent to which the disease can affect [66]. Other people with naked eyes can easily observe these symptoms. Some of them are shown here.

- Tremor: This is an early-stage symptom that is observed in 70% people of the PD. The tremor means the shaking of body parts with frequency of around 4 - 6 Hz. This is most prominent at rest and worsens with emotional stress. Generally, this starts with rhythmic flexion-extension of the finger, hand, and foot or with rhythmic inward-outward forearm rotation. This is not observed during voluntary movement and sleep. At an early stage, this symptom is limited to a limb or two limbs on same side before being generalised. It can be observed in the jaw and chin but not in the head [68].
- Rigidity: It means an increase in resistance against the passive movement of the body parts, which leads to stiffness and flexed posture. This is more a sign rather than a symptom. The stiffness on the languid limb can be defined as the lead pipe rigidity, as the muscle tone is present in the whole limb's movement. When tremors coincide with rigidity, there is a feeling of ratchet-like jerkiness, known as cogwheel rigidity [4].
- Postural changes: There is a stooped body posture due to PD. This is because of the rigidity which forces the body to lean on one side. Moreover, the PD harms the automatic activities of brain, leading to stooped posture as the brain is not commanding to stand straight. This reduces body movement and functional capacity [22].

- Gait changes: The gait means walking pattern. Due to PD, there is a significant change in gait pattern, like slow turns, reduced arm movement, slow stumble, reduced blinking of the eyes, chances of falling forward, and small stride length in walking.
- Speech and Swallowing: People with PD may suffer from chewing or eating difficulties, increased salivation in the mouth, and garble.
- Changing in writing patterns: A test is performed on the patient in which the patient has to write some sentences or draw the spiral on paper. Then, based on the frequency of vibration of hands, which can be observed by seeing the work, the doctor gets more clarity on PD detection.[56]

2.2. Non-Motor Symptoms. The symptoms that are not associated with movement and balance are known as non-motor symptoms. Other people with naked eyes cannot easily observe these symptoms. Non-motor features can early than the motor clinical features, which are increasingly recognised as important features during the phase of pre-PD. Some of them are discussed here:

- Eye problems: One common issue PD patients face is eyesight problems. But it is not entirely related to PD; many other factors can affect eyesight or cause eyesight issues. PD patients may face relatively high issues when moving their eyes or trying to force them quickly. This might be more evident when looking at fast-moving objects like vehicles. This problem can also be caused by Progressive Supranuclear Palsy (PSP), which has symptoms similar to PD. Taking the medicine for PD, particularly anticholinergics, can result in blurred vision. Patients might see double images of a single object simultaneously. Poor coordination and wearouts of the muscles that move the eyeballs mean that the eyes have trouble moving together, causing double vision. Other problems, such as diabetes, can also cause double vision. Other eye problems that PD patients may feel are dry eyes, eyelid apraxia, contrast sensitivity, Colour vision problems, difficulty measuring space, etc [37].
- Depression: Depression might increase or lower mood and emotional health. Some of the problems that a depressed person experiences are difficulty concentrating, tiredness, lack of sleep or excessive sleep, loss of appetite or increased appetite, the expertise of worthlessness or guilt. In most critical cases, thoughts of death or self-harm and suicidal ideas can be the symptoms due to lack of dopamine [49].
- Fatigue: Fatigue can be expressed as an overwhelming feeling of tiredness, exhaustion and a lack of energy. Often, tiredness and fatigue are misunderstood. Tiredness usually goes away with a good nap or rest. But with fatigue, it does not improve with rest. Up to half of the PD patients experience fatigue. Fatigue might not be related to PD but could be caused by another condition like a thyroid problem.
- Anxiety: Anxiety is a state of nervousness characterised by worry or fear. If they find it difficult to face a situation, everyone in their lives exhibits anxiousness. A sense of imminent danger, constant worry, trouble focusing, an inability to relax, trouble sleeping, sweating, a racing heart, tightness in the chest, dizziness, trembling, nausea, stomach aches, appetite loss, a dry mouth, muscle pain, tension, restless legs, difficulty getting a good night's sleep, etc. are some of the symptoms that an anxious person may experience. Your daily life may be impacted if these sensations persist for an extended amount of time. Anxiety and sadness symptoms are sometimes present in patients. Although there may be overlap, there are three primary categories of anxiety. Multiple types may be experienced by many patients.
    1. Generalised Anxiety Disorder
    2. Panic attacks
    3. Phobia

  For people with and without PD, anxiety is probably caused by a combination of several things, including genetics and stressful daily life.
- Skin and sweating: Daily-life PD patients are mainly affected. The skin has glands that produce a greasy/oily substance called sebaceous matter, known as sebum. The production of sebum in PD patients is more than that of non-PD patients. This means the skin, especially the face and scalp, becomes greasy and shiny. They may suffer from seborrhoeic dermatitis. In this, the areas of the skin that consist of sebaceous glands become red and itchy. This is also a common occurrence and can happen without PD. Thus, these conditions in PD patients can be in the scalp, face, ears, chest, bends

and folds of skin like under the arms. And this is not caused by poor personal hygiene. PD patients can experience extreme sweating (hyperhidrosis). This happens if the prescribed medications lose their effects at the end of their dose. This can also occur when the drugs are working at their best. Some patients with PD may not sweat enough. This is called hypohidrosis. This may be a side effect of anticholinergics, a medication used to treat PD.

Another rarely used classification is through the sensory symptoms, including the gradual loss of smell, which can be detected by the University of Pennsylvania Smell Identification Test (UPSIT) [36]. Generally, the peripheral sensory symptoms like smell start to affect more early. This loss in smell symptoms is seen even earlier than motor or non-motor symptoms. This loss in smell is observed in around 75 to 95 percent of PD people. But for the taste sensor, it is followed. Generally, the taste loss is observed in the advanced stage, but in rare cases, the loss of taste can be observed earlier [41]. Moreover, impairment in voice is also observed as one of the common symptoms in many PD patients [2].

The above discussed are the symptoms that are in support of the PD. As discussed earlier, many of these symptoms are observed near the last stage. But to get the early detection of PD based on brain MRI, various DL models can be used, one of which is CNN, discussed below.

The CNN is a type of DL neural network architecture that is commonly used for applications in image and speech interpretation. High dimensionality of images can be filtered by convolutional layers that basically extracts features and store in matrix-grid structure without compromising with its informational data. That is why CNNs are especially suited for this image classification. Layers in CNN: a input layer, a convolutional layer, a pooling layer, flattening, a fully connected layer and a output layer. vasquez *et al.* [61] came up with an approach to model such difficulties in starting or in stopping movements taking into account the information from speech, handwriting, and gait. They used those adaptions to train CNN to classify patients and healthy subjects. Once the model finally detects the PD, the final step is to cure the same. But currently, no treatments are available that can completely cure it as it is impossible to reverse the degeneration of neurons that causes PD. However, many treatments can help your symptoms. Treatments for Parkinson's include:

- Surgical: It includes Deep Brain Stimulus, a neurosurgical technique for curing PD. It uses embedded electrodes and electrical stimulation to cure the disorder related to the movement of neurons inside the brain. In this process, external electric currents stimulate the brain cells. This current is supplied by the device kept near the clavicle [19].
- Drug-oriented: The drugs-oriented method can be chosen to get temporary relief against PD. Two drugs are commonly used, which are levodopa and carbidopa intestinal gel. The levodopa drug provides dopamine to the brain, which is deficient in PD patients. The carbidopa, a peripheral L-dopa decarboxylase inhabitor drug, ensures that the levodopa drug only increases the dopamine level of the brain as there are many sources of dopamine in our body[45].

The social and economic environment might influence treatment decisions, such as when to begin therapy, what sorts of treatment to utilise, and whether to change treatment. As the medical condition changes, treatment may need to be adjusted on a frequent basis to balance quality-of-life concerns, treatment side effects, and treatment expenses. A regular checkup with members of your healthcare team is required to adapt your therapy as your condition evolves.

Imaging can be helpful in patients with uncertainty in diagnosis (e.g., early stage, essential tremor) or research studies to ensure accuracy, but it is not necessary in routine practice. This may change when there is a disease-modifying therapy, and making a correct diagnosis as early as possible is essential. Many efforts are underway to accurately define a premotor stage of PD with high sensitivity and specificity. There is also some evidence that the diagnosis of PD, and even pre-PD, may be made based on increased iron in the SN using transcranial sonography or particular MRI protocols.

3. Taxonomy.

This section included the description of various datasets of brain MRI images, various data preprocessing techniques that can be applied to this dataset, and various DL models that can be used. The Fig. 3.1 shows the various subsection of taxonomy with their citations number. The detailed explanation is as follows.

3.1. Datasets.

Pahuja *et al.* [42] used the Parkinson's Progression Marker Initiative (PPMI) dataset for detecting PD. This dataset provides brain MRI images of patients who have PD. Identifying the biomarkers of

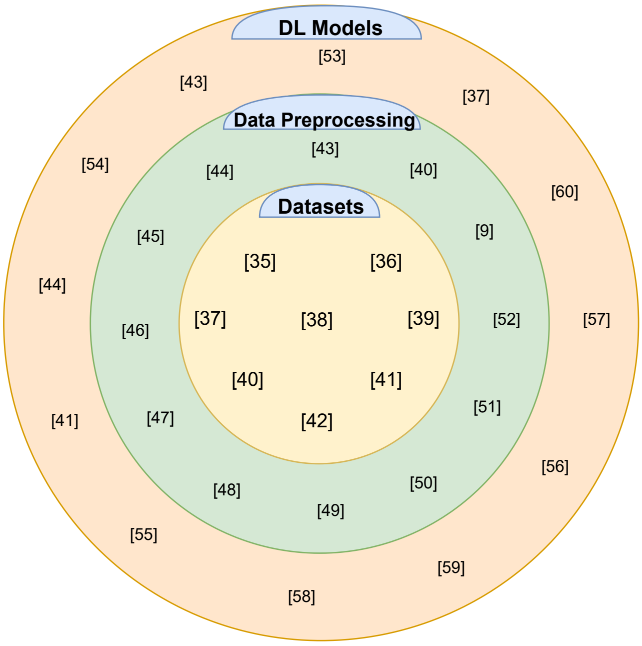


Fig. 3.1: Taxonomy chart

the Parkinson Progression is the primary goal of this dataset. They considered a 1.5T baseline 3D volumetric T1-weighted 150 brain MRI images. Half are healthy patients, and the remaining are PD patients. From the dataset, 72 images were replaced with 60 new illustrations due to error in the segmentation method.

Liu *et al.* [34] used the brain MRI dataset to diagnose PD. This dataset was accepted by Ethics committee, which was earlier associated with the Hospital of Jinzhou Medical University and then after the written consent was relinquish. This dataset contains the T2-weighted MRI images of patients, which were captured on a 3.0T MRI machine. They also used axial T1 weighted, axial Fluid Attenuated Inversion Recovery (FLAIR) and sagittal T1 weighted for anatomical references for Regions of interest (ROI) delineation and placement. Around 138 patients' MRI images, out of which 69 were healthy and 69 were PD-classified, were used as input for the model.

Yasaka *et al.* [67] enrolled patients with PD and HC who paid a visit to Juntendo University Hospital between February 2017 and October 2018 with the help of specialised neurologists who made the criteria for PD and examined them with a 3-T MRI unit using a 64-channel head coil. Multi-shell dMRI, Magnetization transfer saturation images and as result of which we got T1-weighted images for 115 PD and 115 healthy patients.

Wang *et al.* [64] used Quantitative susceptibility mapping (QSM) and T1-weighted MRI in their investigation to detect PD. They recruited 92 PD patients and 287 healthy people from Ruijin Hospital through QSM. The First Affiliated Hospital of Zhengzhou University obtained 83 PD and 72 healthy control data from T1-weighted MRI. The results from QSM were obtained utilising a 3T scanner equipped with a array head coil having 15 phased channel and a 3D Gradient echo sequences (GRE) imaging sequence. The data was gathered utilising a 3T scanner with a 64-channel phased array head coil and a 3D GRE imaging sequence from T1-weighted MRI.

Talai *et al.* [55] evaluated 45 PD patients, 20 with progressive supranuclear palsy, and 38 healthy individuals from the University Medical Centre Hamburg-Eppendorf's movement disorder outpatient clinic. Between July 2009 and September 2010, patient data were used for this investigation. They acquired DTI, T1-weighted Magnetization-Prepared Rapid Acquisition Gradient Echo MRI (MPRAGE), and triple-echo T2-weighted MRI datasets of each patient using a 3T Siemens Skyra MR scanner.

Table 3.1: Early detection PD metadata

Author	**Year**	**Dataset Taken**	**Type of scan**	**Number of Subjects**
Pahuja *et al.* [42]	2016	PPMI	T1-weighted MRI	75 PD, 75 HC
Liu *et al.* [34]	2020	Hospital data	T2-weighted MRI	69 PD, 69 HC
Yasaka *et al.* [67]	2021	Hospital data	T1-weighted MRI, dMRI, MT saturation images	115 PD, 115 HC
Wang *et al.* [64]	2023	Hospital data	T1-weighted MRI, QSM	92 PD, 287 HC & 83 PD, 72 HC
Talai *et al.* [55]	2021	Medical Center	T1-weighted MRI, T2-weighted MRI, DTI	45 PD, 20 Progressive supranuclear palsy, 38 HC
Pahuja *et al.* [43]	2022	Heterogeneous	T1-weighted MRI	73 PD, 59 HC
Shinde *et al.* [52]	2019	Medical Center	Neuromelanin sensitive MRI	45 PD, 20 Atypical Parkinsonian Syndromes, 35 HC
Ramírez *et al.* [46]	2020	Open Access Database	DTI	129 PD, 57 HC

Pahuja *et al.* [43] used the heterogeneous dataset of about 73 PD patients and 59 healthy individuals, which includes MRI scans, cerebrospinal fluid (CSF) biomarkers, demographic information. It includes neuroimaging data that 3D T1-weighted MRI scans and SPECT images (used to extract the specific binding ratio), whereas biological markers data includes four CSF markers.

Shinde *et al.* [52] used MRI, specifically Neuromelanin sensitive MRI dataset using a 3T MRI scanner, that contains the result of 55 subjects, includes 30 patients having PD and 25 HC, which was arbitrarily distributed into training and holdout sets, 25 HC and 30 PD patients for training and cross-validating the models remaining 10 HC and 15 PD patients for testing models.

Ramírez *et al.* [46] used the PPMI dataset, used T1-weighted images to extract subcortical Volumes of Interests (VOIs) using the free surfer software package of 129 PD patients and 57 healthy patients, that includes eddy current correction, brain extraction and tensor fitting. The authors used the DTI model to compare Fractional Anisotrophy (FA) maps, which were used as inputs in DL models.

The Table 3.1 shows the overview of datasets available for PD detection.

3.2. Data Preprocessing. Bhan *et al.* [8] used the MRIcro tool for data preprocessing on T1-weighted MRI images from the PPMI dataset to detect PD. Here, the image data were labelled between Parkinson's and healthy patients for feeding to the DL model based on respective T1w scans. Then, they sliced the lower part of each image as they did not contain any significant information, leaving with 10548 total images. Further, they used the TensorFlow library to read images. Based on this, they labelled the images of PD patients as 0 and healthy patients as 1, then split the data into training(90%) and testing(10%). Lastly, they applied One Hot Encoding and converted these integer values into binary values.

Sangeetha *et al.* [51] used a median filtering approach for data preprocessing on the PPMI dataset for diagnosing PD. This non-linear approach effectively reduces the noise of the MRI image and helps in preserving the edges of the MRI image. In this technique, firstly, each pixel is arranged in numerical order, and then every pixel of the image is replaced by the middle of its adjacent pixels. This filtering technique is so efficient that it can filter out minimal noise.

Zhang *et al.* [69] applied two data preprocessing techniques on their PPMI datasets to detect PD. Firstly, due to less availability of data samples, which may lead to underfitting or overfitting in DL models, they applied data generation by using Wasserstein generative adversarial network (WGAN) technology, and when the training epochs reached to the specific limit, after some interval, a new image MRI is generated and added to the original dataset. Secondly, to achieve diversity in the dataset to get a more accurate answer, they used the ImageDataGenerator API of the Keras model. They applied some basic rotations and inversion to the images at the end.

Bhan *et al.* [7] used image enhancement techniques on the PPMI dataset for diagnosing PD. This technique

Table 3.2: Data cleaning and transformation for early PD detection

Author	**Year**	**Output**	**Process**
Bhan *et al.* [8]	2021	Binary	MRIcro tool One Hot Encoding
Sangeetha *et al.* [51]	2023	Image	Median Filtering
Zhang *et al.* [69]	2019	Image	WGAN Technology ImageDataGenerator API
Bhan *et al.* [7]	2021	Image	CLAHE Gaussian blur Histogram Equilizer
Camacho *et al.* [12]	2023	Image	HDBET ANTs log-Jacobian maps
Rubbert *et al.* [48]	2019	Image	FMRIB BET2 FIX
Pereira *et al.* [44]	2023	Image	Visual Rhythm approaches
Noor *et al.* [40]	2020	Image	Motion correction Spatial normalization Scaling Feature extraction
Veetil *et al.* [62]	2023	Image	Skull stripping Bias field correction Normalization Data augmentation
Erdaş *et al.* [23]	2023	Image	Image registration using python code
Chakraborty *et al.* [13]	2020	Image	Image registration

removed unwanted noise, colour, brightness, etc. This was done by applying image filtering techniques and using a histogram equaliser to increase the contrast and quality of images. For enhancing the image, the RGB(red, green and blue) images were converted to the most commonly used colour encoding scheme, that is, luma (brightness (Y), blue projection (U) and red projection (V))(YUV) channels. After the conversion, they used the Gaussian blur function to filter extra noise and pixels. To improve contrast in images the Contrast-limited adaptive histogram equalization (CLAHE) technique was used. Finally, the luma (brightness (Y), blue projection (U) and red projection (V))(YUV) channels were combined and converted into red, green and blue(RGB) colour space.

Camacho *et al.* [12] used the following steps to process T1-weighted pictures. HD-BET was used to eliminate all non-brain tissues from original MR images for brain extraction. The resulting images were then aggregated using linear interpolation to an isotropic resolution of 1 mm. The bias field was corrected in the second step using the non-parametric non-uniform intensity normalisation technique from the Advanced Normalisation Tools (ANTs) toolkit version 2.3.1. The T1-weighted MRI data were then non-linearly registered to the MNI PD25-T1-Magnetization-Prepared Rapid Acquisition Gradient Echo MRI (MPRAGE) 1 mm brain atlas (fixed image). After aligning the data to the atlas transformation, they employed ANTs to initialise the non-linear registration in the second step.

Rubbert *et al.* [48] used whole brain resting state resting state fMRI (rs-fMRI) to discriminate PD patients from healthy patients. To pre-process the rs-fMRI data, they took the help of the Oxford Centre for FMRIB Software Library(FSL) version 5.0. Brain extraction was performed by Brain Extraction Tool of fMRI of the

Brain (FMRIB) . After smoothing the image with a Gaussian kernel and normalisation, They used FMRIB's ICA Xnoisifier to automatically denoise rs-fMRI data.

Chakraborty *et al.* [13] acquired their research data from the PPMI dataset. So, to solve the differences, as the data has been collected from multiple centres worldwide, the images needed to be in the same space as their references, so they performed an image registration procedure. The image registration procedure was performed on the source images(in this study, the PPMI dataset), and the atlas, such as MNI and Individual Brain Atlases using Statistical Parametric Mapping (IBASPM), were identified as target images. They executed the registration of the MRI scan using symmetric normalisation using a tool known as Advanced Normalization Tools Python (ANTsPy).

Erdaş *et al.* [23] also used the image registration procedure to integrate the PPMI dataset to convert unseen or unknown images into adequately aligned fixed images. With the source image as PPMI and target image as Montreal Neurological Institute(MNI), the procedure was executed with 152 T1-weighted linear 1 millimetre (mm) atlas by making completely automated code in python based upon FLIRT registration tool linked with the FMRIB Software Library(FSL) using BET method to remove unnecessary tissues, bones, skin, fat and other bodily structures to increase the performance of the method to be applied on the given MRI dataset.

Pereira *et al.* [44] used an intelligent pen to extract information from handwritten dynamics and applied a normalisation step to input signals. In this, they used a filter bank to remove features such as edges and corners, and they then developed a CNN to learn pen-based features; they also mentioned that the CNN is composed of several layers, each of which is responsible for learning a different and finer representation of the data and textures from an image from input signals: helps to improve the performance of DL algorithms by ensuring that the input data is on a similar scale.

Veetil *et al.* [62] used several Data pre-processing techniques to prepare the MRI data for analysis. These techniques include skull stripping, bias field correction, and normalisation. Skull stripping eliminates non-brain tissue from the MRI images, while bias field correction is used to correct intensity inhomogeneities in the images. Normalisation is used to standardise the intensity values of the images. Moreover, the authors used Data augmentation techniques like rotation, flipping and scaling to increase the dataset's size and improve the models' robustness. The experiment progessed by taking an 80-20 training-testing, spliting augmented images. Then normalising the image intensities to the range of 0 to 1 for gaining monotonicity in the intensity and thus helps in contrasting across all images.

Noor *et al.* [40] used techniques that included motion correction, Spatial normalisation, and scaling. They also mentioned the importance of feature selection and extraction in achieving accurate results and summarised the features used in each study. Overall, the paper enlighten the importance of data preprocessing in careful way and selection of feature in achieving accuracy and reliability in DL applications for PD.

Pahuja *et al.* [43] used three primary image data preprocessing techniques before applying VBM, mainly Spatial normalisation (SPM8), Unified Segmentation (SPM8), used to segment images into various tissue types such as grey matter, cerebrospinal fluid that allow for identification of differences in tissues between individual, Smoothing used to reduce noise in images and additionally for feature extraction smoothed modulated GM volumes are used.

The Table 3.2 shows overview of various data preprocessing approaches that can be applied.

3.3. DL Models. Kumaran *et al.* [32] used modified Visual Geometry Group (VGG) Net Architecture, the standard CNN, to detect PD. They compared four different architectures and found the modified VGG Net architecture to be the best in accuracy. This architecture contains many blocks, and each block has a 2D convolution. They took the swallow tail sign from the brain MRI as an input variable to that model. The VGG Net model can detect around 1000 object categories like keyboard, person, mouse, animals, birds, pencils, etc. Through this model, they achieved approximately 93% of efficiency in detecting PD.

Bhan *et al.* [8] used LeNet-5 architecture for the early detection of PD. LeNet-5 architecture is a type of CNN. This model includes two Conv2D, two pool layers and a hidden layer. Their baseline model has 2 dense layers. The first layer includes of ReLu activation along with 128 neurons, and the second layer includes of a Sigmoid activation function followed by two neurons at each thick layer. They used a batch size of 32, and the number of epochs was 30. This model gave approximately 96.6% accuracy on training dataset and 97.6% accuracy on testing dataset with a loss of 0.07 percentage with no batch normalisation. When the same model

Table 3.3: Performance stats of DL model for early detection of PD

Author	**Year**	**Objective**	**Pros**	**Cons**
Kumaran *et al.* [32]	2022	Early and accurate detection of PD from brain MRI scans	Accomplished 93% accuracy in PD detection	Cannot detect the stages of PD
Bhan *et al.* [8]	2021	To increase the chance of curing through early detection	Attained 97.63% accuracy in PD detection	Can't analize medical or neuro images
Sangeetha *et al.* [51]	2023	To precisely diagonise PD using CNN model	Acheived 95% accuracy in PD detection	Medical picture analysis are less efficient
Kollia *et al.* [31]	2019	To enhance PD detection using CNN-RNN model on MRI	Reached 98% accuracy in PD detection	Absence of diagnosing neurodegenerative disease
Shinde *et al.* [52]	2019	Create diagnostic biomarkers of PD using NeuroMelanin-sensitive (NMS-MRI)	Obtained 80% testing accuracy compared to Radiomics based classifier (RA-ML) with its accuracy being 60%	Dependent on the testing ability of the method to differentiate between PD and other parkinsonian disorders
Sivaranjini *et al.* [53]	2019	Diagnosis of PD using CNN	An accuracy of 88.9% is achieved through proposed method	Low Performance level beacause of less tuned AlexNet model
Kaplan *et al.* [27]	2022	Early Diagnose of PD	An Accuracy of 99.53, 99.22,98.70 % is obtained from various datasets	the accuracy of the classification may be limited by the number and quality of features used
Choi *et al.* [16]	2017	Diagnosis of PD via DAT imaging interpretition	Improved performance in image classification:DAT(98%)	Carefull analysis and evaluation of performance and limitation are required
Yasaka *et al.* [67]	2021	Investigate the use of DL techniques to differentiate PD patient from Healthy patients	Accuracy: 67%-89%, AUC of 0.895,0.800,0.761 for (RK-weighted matrix, AK-weighted, ICVF&AVF-weighted matrices)	relative small sample size, lack of eternal validation, single-centre dataset, lack of generalizability of CNN
Khairnar *et al.* [29]	2023	Development of deep-learning method for early prediction of PD	Proposed the more reliable method for PD detection using CNN	Haven't used CNN & Artificial Neural Networks (ANN) model to give cost effective prediction of PD
Sahu *et al.* [50]	2021	Efficient detection of PD using DL technique	Obtained 93.46% of accuracy comparing other existing approaches	Less DL tools are used

is used with batch normalisation, the accuracy of training dataset is 95.4%, and the accuracy of testing dataset is 97.9% with a loss of 0.05%.

Kollia *et al.* [31] used Deep Neural Network (DNN) based CNN and CNN-RNN for diagnosing PD. The Residual Network-50 (ResNet) structure was used for the pooling and convolutional part. The Gated Recurrent Units (GRU) was utilised in the RNN part of the CNN-RNN architecture. Firstly, the CNN-RNN model was trained with training MRI datasets. It gave around 70.6% performance on the testing dataset, then after the performance was enhanced by changing the value of the modified Loss Function. The CNN model with two completely connected layers and no hidden layer gave around 94% accuracy. The CNN-RNN architecture with a single fully connected layer and two hidden layers gave around 98% accuracy.

Sangeetha *et al.* [51] applied CNN on brain MRI images for PD detection. The main reason for applying this model was the spatial nature of the model, because of which the number of hyper-parameters was reduced. They used five convolutional levels with ReLu activation at each layer; at the first level, there were 16 filters. At the second level, there were 32 filters, and the third, fourth and fifth layers contained 64 filters. Their model includes five max pooling layers and a flattened layer in the middle of the first dense layer and last pooling layer. Moreover, there was ReLu stimulation in all 128 primary levels and SoftMax activation in 2 layers, making 130 dense layers. The model reached to 95% accuracy with specificity and sensitivity of approximately 97%.

Shinde *et al.* [52] advised a computer-based analysis algorithm that uses a CNN to produce (NeuroMelanin-sensitive(NMS)) NMS-MRI diagnostic biomarkers for PD. They employed a standard CNN architecture acquired from ResNet design, which is considered good in image classification related to the medical field. They also compared Contrast Ratio Classifier (CR-ML) and Regression Analysis (RA) with their proposed method. They obtained higher accuracy, sensitivity and specificity compared to radiomics with their novel approach, with the accuracy of RA being 81.8% and CNN-DL's accuracy being 83.6% in cross-validation.

Chowdhary *et al.* [18] used a computer vision method to make the process of detection of PD more refined. They used histogram of oriented gradients (HoG) as a feature extraction method and used CNN, which is based on sequential model, for a lightweight model. Using this proposed model they achieved 94% accuracy with specificity and sensitivity as 92% and 80%. The proposed model can be used on embedded and hand-held devices for a quick self-analysis.

Sivaranjini *et al.* [53] analysed T2-weighted MR images of the brain for detecting the PD using CNN pre-trained model named AlexNet. AlexNet, which comprises many layers like the input layer, convolution layer, pooling, dropout layer and fully connected layer, helps to classify the input data images in PD and healthy patients using mandatory operations. The presented method's performance of pre-trained AlexNet CNN model is determined by measuring its accuracy, specificity and sensitivity. They obtained these parameters of their proposed approach and compared them with other methods. They achieved 88.9% accuracy, 89.3% sensitivity and 88.4% specificity with their approach.

Khairnar *et al.* [29] presented a CNN and ANN based method using MRI and SPECT scans. They applied data pre-processing techniques on the MRI dataset like image resizing, augmentation, normalisation and noise removal. And then pre-processed data were then given to already trained CNN model. While applying the pre-processing techniques like on the MRI scans with the difference of data cleaning instead of image resizing and feature selection instead of noise removal on the SPECT dataset, they trained the ANN model on the SPECT dataset.

Sahu *et al.* [50] used a mixture of RA and ANN to detect the disease by probability estimation using these DL tools. They also estimated the predefined edge of the neurons to the patients' vocal recognition, content of iron, and pulse rate data. RA is used to pre-process the data, and then the pre-processed data is fed to the trained ANN model. The calculated probability values and the five attributes obtained by RA are stored in a file and then used to produce the probability of PD with ANN. After evaluating the final output, they compared it with other approaches and found their accuracy of 93.46% with a specificity of 67.34% and sensitivity of 95.64% of their proposed approach. They executed their proposed algorithm's stimulation with the help of C language with the Scientific Laboratory (SCILAB) program for graph plotting.

Kaplan *et al.* [27] employed two classifiers and a combination of feature extraction and classification techniques to achieve accuracy. The k-nearest neighbour (kNN) algorithm produces the best results for clinical staging and PD motor symptom classification, and k is a hyperparameter that may be adjusted to improve classification accuracy. Whereas Support Vector Machine (SVM), a type of supervised learning algorithm that finds the optimal hyperplane that distinguishes data points of different classes with the most significant margin, is effective for high-dimensional data and can handle noisy data, it is sensitive to the kernel function and hyperparameters used. As a result, the classifier produced the best classification results for dementia status categorization.

Choi *et al.* [16] used a DL-based system, a type of ANN particularly well suited for image classification tasks. Mainly, a CNN was used to train the system on SPECT images of PD patients and healthy patients. Once the CNN is trained, it can accurately classify SPECT images as PD or non-PD. Compared to the traditional methods followed by humans for the interpretation of FP-CIT-SPECT imaging, the DL-based model system

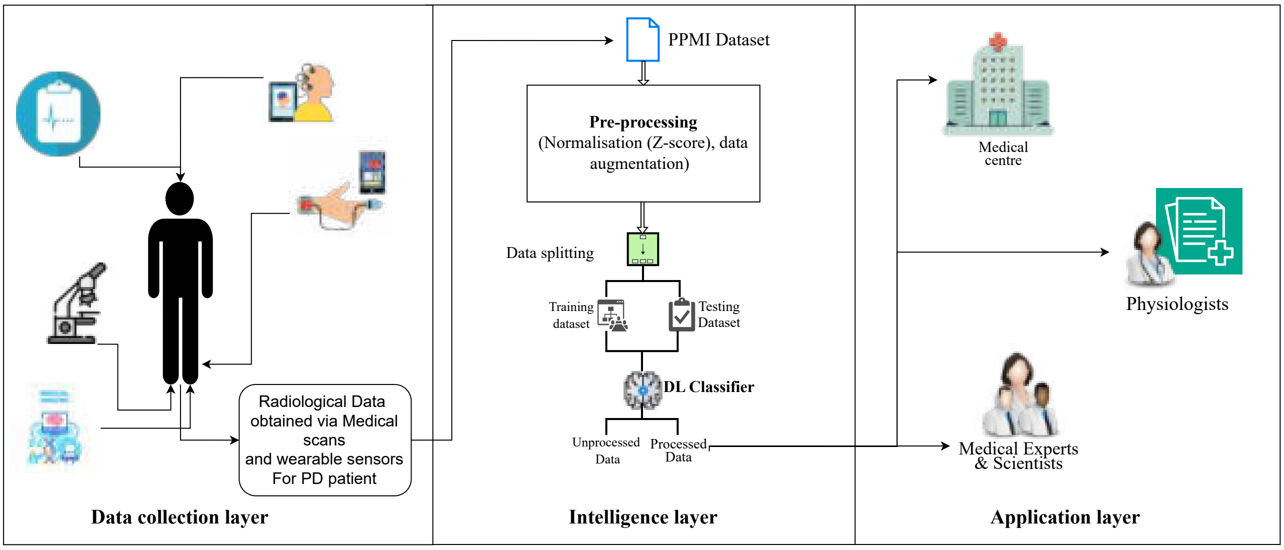


Fig. 4.1: Case study overview

can overcome the variability of human evaluation and provide more objective patient group classification. A particular audience for the same are patients with uncertain Parkinsonism and for the classification of atypical subgroups, for example, SWEDD.

Shivangi *et al.* [26] used the first model, VGER Spectrogram Detector, which takes spectrogram images as input and then used a CNN to categorize the data into one of three classes: healthy,early-stage PD. The next one or second model is the Voice Impairment Classifier, as input we using elabrated features of speech images and uses a stacked autocoder (sAE) to classify the images into severe or not severe PD. They were trained and tested for the balanced datasets that contained almost equal proportions of all the classes.

Yasaka *et al.* [67] used a DL method called CNN for the classification of PD patients and healthy patients. Input data was parameter weighted, and the structural connectome matrices vary according to the number of streamlines was thus evaluated from dMRI and was trained with high accuracy. They also used gradient-weighted class activation mapping (Grad-CAM) to visualise the regions of the connectome matrices, which were essential during CNN's decision making process. The Table 3.3 shows the brief of various DL models for early detection of PD.

4. Case study.

4.1. Data collection layer. We gathered a PPMI dataset by setting specific options; for the machine, we chose the Seimen machine, and the images were grey. This dataset included 970 PD patients and 210 healthy individuals. These individuals underwent 3D MRI scans as part of the study. We used preprocessed MRI images that were converted into 3D Numpy arrays. Numpy arrays are a well-known and feasible data structure for numerical operations in Python, which makes it suitable for feeding the data into DL models. Fig. 4.1 demonstrates the overview of case study we performed.

4.2. Intelligence layer. Data is being processed and prepared for training and testing purposes for the CNN model.

4.2.1. Data Preprocessing. In our study, we used Z-Score Normalisation, which was then applied to the MRI data so that the pixel has a mean of zero and ensures a standard deviation of one. This aids in lowering data variance and qualifies it for machine learning model training.

4.2.2. CNN model. We chose CNN model for the analysis. CNNs are particularly well-suited for image data because they can automatically learn features from the images. The preprocessed 3D Numpy arrays served

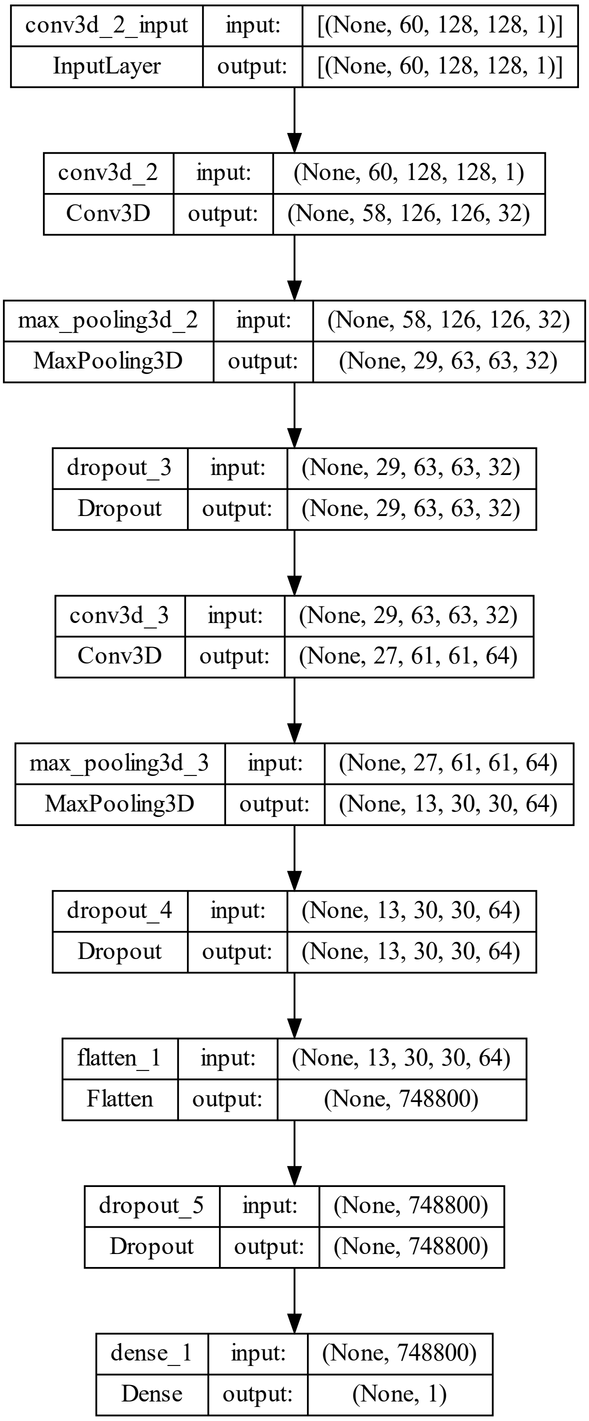

conv3d_2_input | input: | [(None, 60, 128, 128, 1)]
InputLayer | output: | [(None, 60, 128, 128, 1)]
conv3d_2 | input: | (None, 60, 128, 128, 1)
Conv3D | output: | (None, 58, 126, 126, 32)
max_pooling3d_2 | input: | (None, 58, 126, 126, 32)
MaxPooling3D | output: | (None, 29, 63, 63, 32)
dropout_3 | input: | (None, 29, 63, 63, 32)
Dropout | output: | (None, 29, 63, 63, 32)
conv3d_3 | input: | (None, 29, 63, 63, 32)
Conv3D | output: | (None, 27, 61, 61, 64)
max_pooling3d_3 | input: | (None, 27, 61, 61, 64)
MaxPooling3D | output: | (None, 13, 30, 30, 64)
dropout_4 | input: | (None, 13, 30, 30, 64)
Dropout | output: | (None, 13, 30, 30, 64)
flatten_1 | input: | (None, 13, 30, 30, 64)
Flatten | output: | (None, 748800)
dropout_5 | input: | (None, 748800)
Dropout | output: | (None, 748800)
dense_1 | input: | (None, 748800)
Dense | output: | (None, 1)

Fig. 4.2: CNN_Layered_Architecture

as input to the CNN, and we used data augmentation to generate more samples of healthy patients artificially. Standard data augmentation techniques for images include rotation, flipping, cropping, and changing brightness or contrast. The researchers created variations of the original MRI images by applying these transformations, effectively by expanding the dataset. The detailed architecture is being shown in the Fig. 4.2 The input grey scale MRI images had 60 layers in which each layer of size 128 × 128, whereas "None" in the input shape (None, 60, 128, 128, 1) means that the network can accept variable batch sizes for input data while keeping the other dimensions fixed. The number of PD patient images was 970, and the number of healthy patient's images was 210 before and 840 after augmentation. There were a total of 100 epochs with the batch size of 32. So, the total trainable parameters were 1553861. The CNN model was initially trained on the preprocessed data without data augmentation, which resulted in a 65% accuracy rate in distinguishing between PD and healthy subjects, as shown in the figure. This accuracy level indicates the model's performance before data augmentation. After data augmentation, the model's accuracy remarkably improved to 88%, as shown in the figure. This depicts that data augmentation substantially enhanced the model's capability for distinguishing between PD and healthy patients.

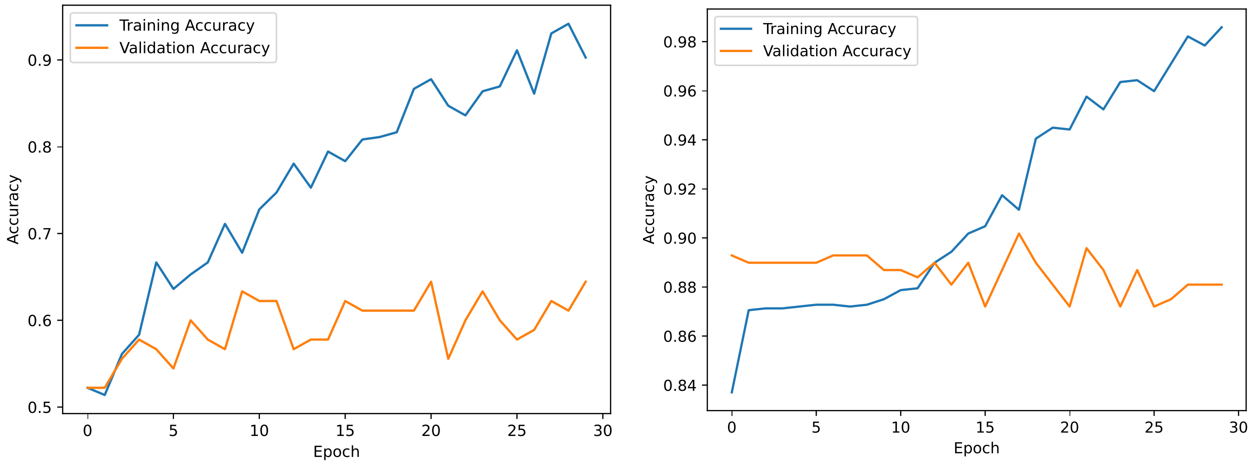
4.3. Application layer. If the result demonstrates high accuracy and reliability, it can be integrated into clinical workflow. Surgeons and healthcare experts can access the result/model through a user-friendly interface or software application. These results give surgeons a broad perspective for deciding the supporting tool or aid in diagnosing or treating PD. By analysing patient data, genetic information and biomarker data (that are critical for tracking disease progression) that is used for evaluating the effectiveness of potential vaccines and treatments, this information can guide researchers or scientists in developing vaccines that target species, aspects of the disease such as abnormal proteins aggregation or neuro-inflammation. Also, the insights gained from the results are helpful for the development of drugs or therapies that might alleviate symptoms or slow down disease progression. These results too provide valuable information for tailoring vaccine or treatment strategies to individual patients based on their specific disease profiles since results are divided into subgroups of patients with varying disease characteristics. Researchers and vaccine developers can use the results of machine learning models to review existing literature, clinical studies, and relevant datasets more efficiently, potentially uncovering previously overlooked insights or connections.

5. Performance Analysis of presented case study. We have used MobaXterm supercomputer as our virtual environment for coding in Python. The supercomputer has installed Ubuntu 20.04.6 LTS as its operating system and has RAM of 250 Gigabytes and Virtual RAM of 32 Gigabytes. With these computer parameters, we began fitting MR scans of PD and HC patients to our model and achieved the following results. Fig. 5.1b shows the accuracy of our model on the scans of PD and HC patients; after we applied data augmentation as a pre-processing technique, it suddenly increased by about 20 %, as shown in Fig. 5.1a. The model was trained through trial and error method, So it took around 1 week to completely train the model.

Fig. 5.2a demonstrates the function of loss in training of data on the chosen model without applying data augmentation, but after referring to Fig. 5.2b, it becomes apparent that the decreasing loss in the validation process is due to data augmentation.

Fig. 5.3b represents a confusion matrix for our model, which shows the numerical data of medical scan identifications for true PD and true Healthy patients after applying data augmentation. These results are significantly higher than the results obtained before using data augmentation as a pre-processing technique, as shown in Fig. 5.3a.

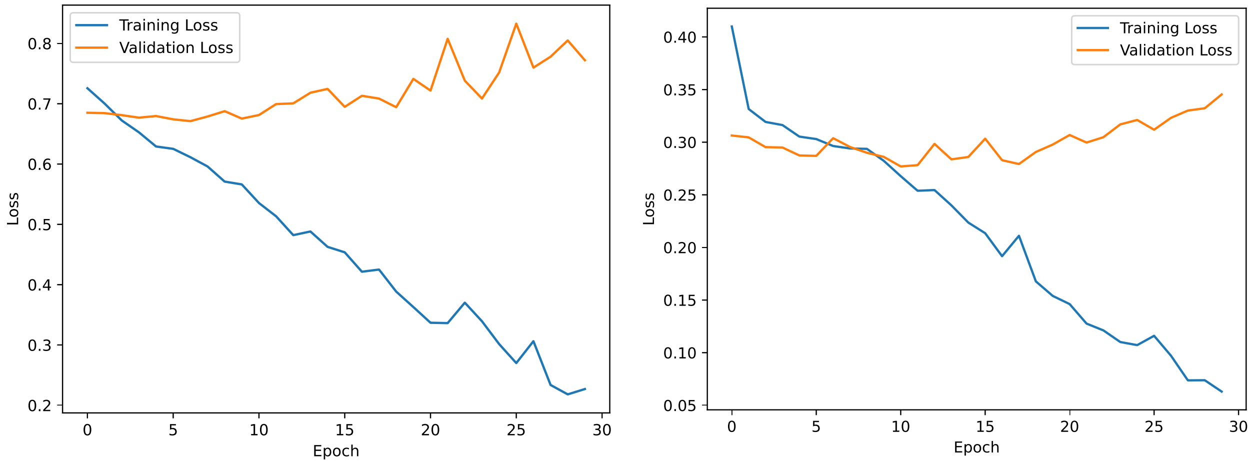
6. Research Challenges. Early detection for PD using DL on MRI datasets is an active area to go for research. However, it does come with some difficulties. Limited availabilities of high-quality datasets on 3D-MR images can make whole research work in vain [1]. The quality and quantity depends on various sources and scanning factors, which is the ultimate reason for biasing. One major bottleneck is feature selection that is identifying the features from MRI data that helps distinguish early-stage PD patients from healthy patients. Also, the models like CNN are known as "Black Boxes" analysing the results and their decisions are quite challenging [47]. Moreover, hardware inconsistencies may affects the performance and accuracy results of DL that is considerable point that makes it more challenging to go for [60]. Collecting the longitudinal data that is important for early diagnosis that is used to track PD progression, which is a cumbersome task [21]. Also,



(a) Training and validation accuracy graph before data augmentation

(b) Training and validation accuracy graph after data augmentation

Fig. 5.1: Accuracy Graphs of Model training



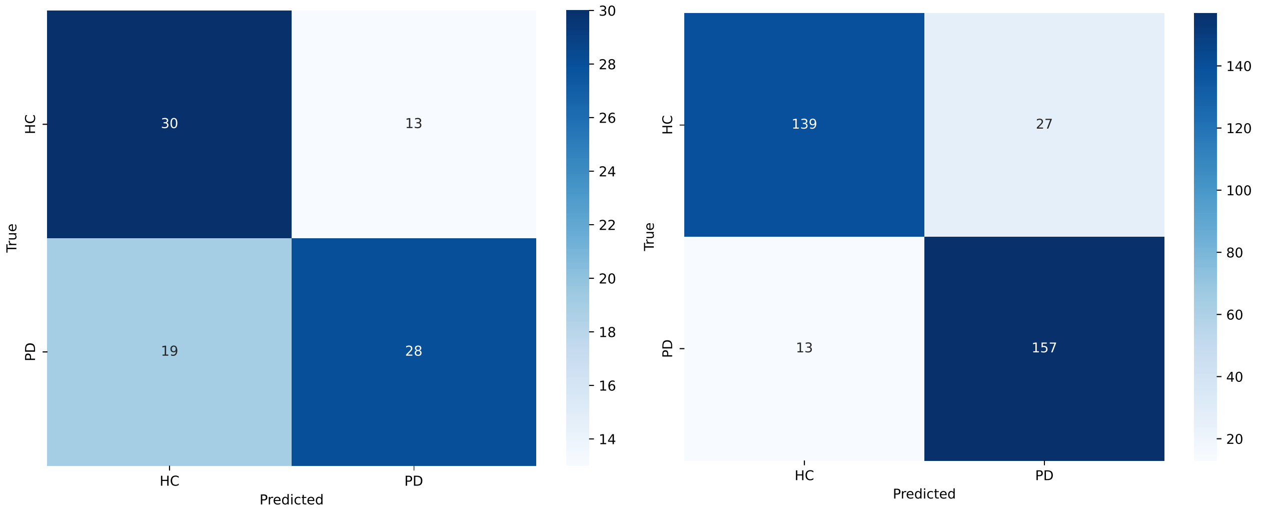
(a) Training and validation loss graph before data augmentation

(b) Training and validation loss graph after data augmentation

Fig. 5.2: Loss graph of model training

to gather diverse MRI datasets and then segmenting them and applying model architectures in order to ensure the quality, performance and accuracy of results would require collaborations between doctors, clinical experts, data scientists and researchers, which is a crucial and more challenging task [38].

Biomarkers and AI will be critical in the future of PD diagnosis. While genetics remains important, the majority of cases include complicated genetic factors interacting with environmental circumstances. This understanding might open the way for the application of AI to predict disease progression and features. Radio-logical imaging studies and genetic markers may play important roles in early diagnosis, making PD a biomarker-supported disorder. Nonetheless, a rising number of medications for disease modification are being devel-



(a) Confusion matrix for CNN model before data augmentation

(b) Confusion matrix for CNN model after data augmentation

Fig. 5.3: Confusion matrix of model training

oped, but obstacles remain, particularly for asymptomatic patients who may not have access to preventive medicine [57].

7. Conclusion. In this paper, we conducted a detailed analysis of the disease covering ts causes, symptoms, and conventional treatment methods. We have also covered a broad range of various datasets, explored different data preprocessing techniques, and a wide range of DL models that was being practised by authors for early detection of PD. The approach of this paper is kept simple and straightforward, so that it can be easy to understand .In our case study, we used 3D brain MRI images of 840 healthy control and PD from the PPMI dataset on which we have applied various data preprocessing techniques and employed CNN for the model. Through this, we have achieved an eminent accuracy of 88%. Our study underscores the potential of advanced technologies to revolutionise early detection and remedial approaches, provides a hope to those who are struggling with the ailment.

REFERENCES

[1] F. Albrecht, J. B. Pereira, M. Mijalkov, M. Freidle, H. Johansson, U. Ekman, E. Westman, and E. Franzén, *Effects of a highly challenging balance training program on motor function and brain structure in parkinson's disease*, Journal of Parkinson's Disease, 11 (2021), pp. 2057–2071.

[2] L. Ali, C. Zhu, M. Zhou, and Y. Liu, *Early diagnosis of parkinson's disease from multiple voice recordings by simultaneous sample and feature selection*, Expert Systems with Applications, 137 (2019), pp. 22–28.

[3] M. S. Alzubaidi, U. Shah, H. Dhia Zubaydi, K. Dolaat, A. A. Abd-Alrazaq, A. Ahmed, and M. Househ, *The role of neural network for the detection of parkinson's disease: a scoping review*, in Healthcare, vol. 9, MDPI, 2021, p. 740.

[4] M. J. Armstrong and M. S. Okun, *Diagnosis and treatment of parkinson disease: a review*, Jama, 323 (2020), pp. 548–560.

[5] Y. J. Bae, J.-M. Kim, C.-H. Sohn, J.-H. Choi, B. S. Choi, Y. S. Song, Y. Nam, S. J. Cho, B. Jeon, and J. H. Kim, *Imaging the substantia nigra in parkinson disease and other parkinsonian syndromes*, Radiology, 300 (2021), pp. 260–278.

[6] M. Bergamino, E. G. Keeling, V. R. Mishra, A. M. Stokes, and R. R. Walsh, *Assessing white matter pathology in early-stage parkinson disease using diffusion mri: A systematic review*, Frontiers in neurology, 11 (2020), p. 314.

[7] A. Bhan, S. Kapoor, and M. Gulati, *Diagnosing parkinson's disease in early stages using image enhancement, roi extraction and deep learning algorithms*, in 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM), 2021, pp. 521–525.

[8] A. Bhan, S. Kapoor, M. Gulati, and A. Goyal, *Early diagnosis of parkinson's disease in brain mri using deep learning algorithm*, in 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile

Networks (ICICV), 2021, pp. 1467–1470.

[9] K. Bharti, A. Suppa, S. Tommasin, A. Zampogna, S. Pietracupa, A. Berardelli, and P. Pantano, *Neuroimaging advances in parkinson's disease with freezing of gait: a systematic review*, NeuroImage: Clinical, 24 (2019), p. 102059.

[10] N. S. Bidesi, I. Vang Andersen, A. D. Windhorst, V. Shalgunov, and M. M. Herth, *The role of neuroimaging in parkinson's disease*, Journal of neurochemistry, 159 (2021), pp. 660–689.

[11] B. R. Bloem, M. S. Okun, and C. Klein, *Parkinson's disease*, The Lancet, 397 (2021), pp. 2284–2303.

[12] M. Camacho, M. Wilms, P. Mouches, H. Almgren, R. Souza, R. Camicioli, Z. Ismail, O. Monchi, and N. D. Forkert, *Explainable classification of parkinson's disease using deep learning trained on a large multi-center database of t1-weighted mri datasets*, NeuroImage: Clinical, 38 (2023), p. 103405.

[13] S. Chakraborty, S. Aich, and H.-C. Kim, *Detection of parkinson's disease from 3t t1 weighted mri scans using 3d convolutional neural network*, Diagnostics, 10 (2020), p. 402.

[14] B. Chen, M. Xu, H. Yu, J. He, Y. Li, D. Song, and G. G. Fan, *Detection of mild cognitive impairment in parkinson's disease using gradient boosting decision tree models based on multilevel dti indices*, Journal of Translational Medicine, 21 (2023), p. 310.

[15] S. J. Cho, Y. J. Bae, J.-M. Kim, D. Kim, S. H. Baik, L. Sunwoo, B. S. Choi, and J. H. Kim, *Diagnostic performance of neuromelanin-sensitive magnetic resonance imaging for patients with parkinson's disease and factor analysis for its heterogeneity: a systematic review and meta-analysis*, European radiology, 31 (2021), pp. 1268–1280.

[16] H. Choi, S. Ha, H. J. Im, S. H. Paek, and D. S. Lee, *Refining diagnosis of parkinson's disease with deep learning-based interpretation of dopamine transporter imaging*, NeuroImage: Clinical, 16 (2017), pp. 586–594.

[17] C. L. Chowdhary, N. Khare, H. Patel, S. Koppu, R. Kaluri, and D. S. Rajput, *Past, present and future of gene feature selection for breast cancer classification–a survey*, International Journal of Engineering Systems Modelling and Simulation, 13 (2022), pp. 140–153.

[18] C. L. Chowdhary and R. Srivatsan, *Non-invasive detection of parkinson's disease using deep learning*, International Journal of Image, Graphics and Signal Processing, 13 (2022), p. 38.

[19] F. C. Church, *Treatment options for motor and non-motor symptoms of parkinson's disease*, Biomolecules, 11 (2021).

[20] C. Clarke, S. Patel, N. Ives, C. Rick, R. Woolley, K. Wheatley, M. Walker, S. Zhu, R. Kandiyali, G. Yao, et al., *Uk parkinson's disease society brain bank diagnostic criteria*, NIHR Journals Library, (2016).

[21] J. P. Devarajan, V. R. Sreedharan, and G. Narayanamurthy, *Decision making in health care diagnosis: Evidence from parkinson's disease via hybrid machine learning*, IEEE Transactions on Engineering Management, 70 (2021), pp. 2719–2731.

[22] F. Doná, C. Aquino, J. Gazzola, V. Borges, S. Silva, F. Ganança, H. Caovilla, and H. Ferraz, *Changes in postural control in patients with parkinson's disease: a posturographic study*, Physiotherapy, 102 (2016), pp. 272–279.

[23] Ç. B. Erdaş and E. Sümer, *A fully automated approach involving neuroimaging and deep learning for parkinson's disease detection and severity prediction*, PeerJ Computer Science, 9 (2023), p. e1485.

[24] P. Feraco, C. Gagliardo, G. La Tona, E. Bruno, C. D'angelo, M. Marrale, A. Del Poggio, M. C. Malaguti, L. Geraci, R. Baschi, et al., *Imaging of substantia nigra in parkinson's disease: a narrative review*, Brain Sciences, 11 (2021), p. 769.

[25] M. Gilat, B. W. Dijkstra, N. D'Cruz, A. Nieuwboer, and S. J. Lewis, *Functional mri to study gait impairment in parkinson's disease: a systematic review and exploratory ale meta-analysis*, Current neurology and neuroscience reports, 19 (2019), pp. 1–12.

[26] A. Johri, A. Tripathi, et al., *Parkinson disease detection using deep neural networks*, in 2019 Twelfth international conference on contemporary computing (IC3), IEEE, 2019, pp. 1–4.

[27] E. Kaplan, E. Altunisik, Y. E. Firat, P. D. Barua, S. Dogan, M. Baygin, F. B. Demir, T. Tuncer, E. Palmer, R.-S. Tan, et al., *Novel nested patch-based feature extraction model for automated parkinson's disease symptom classification using mri images*, Computer Methods and Programs in Biomedicine, 224 (2022), p. 107030.

[28] Z. Karapinar Senturk, *Early diagnosis of parkinson's disease using machine learning algorithms*, Medical Hypotheses, 138 (2020), p. 109603.

[29] S. Khairnar and P. Yawalkar, *Early-stage detection of parkinson's disease*, Industrial Engineering Journal, 52 (2023).

[30] Y. F. Khan, B. Kaushik, C. L. Chowdhary, and G. Srivastava, *Ensemble model for diagnostic classification of alzheimer's disease based on brain anatomical magnetic resonance imaging*, Diagnostics, 12 (2022), p. 3193.

[31] I. Kollia, A.-G. Stafylopatis, and S. Kollias, *Predicting parkinson's disease using latent information extracted from deep neural networks*, in 2019 International Joint Conference on Neural Networks (IJCNN), 2019, pp. 1–8.

[32] R. Kumaran and S. Shanthini, *A hospital application involving deep learning architecture for detecting parkinson's disease*, in 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), 2022, pp. 1–5.

[33] R. Lamba, T. Gulati, H. F. Alharbi, and A. Jain, *A hybrid system for parkinson's disease diagnosis using machine learning techniques*, International Journal of Speech Technology, (2021), pp. 1–11.

[34] P. Liu, H. Wang, S. Zheng, F. Zhang, and X. Zhang, *Parkinson's disease diagnosis using neostriatum radiomic features based on t2-weighted magnetic resonance imaging*, Frontiers in neurology, 11 (2020), p. 248.

[35] H. W. Loh, W. Hong, C. P. Ooi, S. Chakraborty, P. D. Barua, R. C. Deo, J. Soar, E. E. Palmer, and U. R. Acharya, *Application of deep learning models for automated identification of parkinson's disease: a review (2011–2021)*, Sensors, 21 (2021), p. 7034.

[36] M. Löhle, A. Storch, and H. Reichmann, *Beyond tremor and rigidity: non-motor features of parkinson's disease*, Journal of neural transmission, 116 (2009), pp. 1483–1492.

[37] S. Mohana Devi, I. Mahalaxmi, N. P. Aswathy, V. Dhivya, and V. Balachandar, *Does retina play a role in parkinson's*

disease?, Acta Neurologica Belgica, 120 (2020), pp. 257–265.

[38] T. Mortezazadeh, H. Seyedarabi, B. Mahmoudian, and J. P. Islamian, *Imaging modalities in differential diagnosis of parkinson's disease: opportunities and challenges*, Egyptian Journal of Radiology and Nuclear Medicine, 52 (2021), pp. 1–12.

[39] A. A. Moustafa, S. Chakravarthy, J. R. Phillips, A. Gupta, S. Keri, B. Polner, M. J. Frank, and M. Jahanshahi, *Motor symptoms in parkinson's disease: A unified framework*, Neuroscience & Biobehavioral Reviews, 68 (2016), pp. 727–740.

[40] M. B. T. Noor, N. Z. Zenia, M. S. Kaiser, S. A. Mamun, and M. Mahmud, *Application of deep learning in detecting neurological disorders from magnetic resonance images: a survey on the detection of alzheimer's disease, parkinson's disease and schizophrenia*, Brain informatics, 7 (2020), pp. 1–21.

[41] V. Oppo, M. Melis, M. Melis, I. Tomassini Barbarossa, and G. Cossu, *"smelling and tasting" parkinson's disease: Using senses to improve the knowledge of the disease*, Frontiers in Aging Neuroscience, 12 (2020), p. 43.

[42] G. Pahuja and T. Nagabhushan, *A novel ga-elm approach for parkinson's disease detection using brain structural t1-weighted mri data*, in 2016 Second International Conference on Cognitive Computing and Information Processing (CCIP), 2016, pp. 1–6.

[43] G. Pahuja and B. Prasad, *Deep learning architectures for parkinson's disease detection by using multi-modal features*, Computers in Biology and Medicine, 146 (2022), p. 105610.

[44] C. R. Pereira, S. A. Weber, C. Hook, G. H. Rosa, and J. P. Papa, *Deep learning-aided parkinson's disease diagnosis from handwritten dynamics*, in 2016 29th SIBGRAPI conference on graphics, patterns and images (SIBGRAPI), Ieee, 2016, pp. 340–346.

[45] S. Raj, R. Sarvankar, L. Filipe, V. Benedetto, N. Mason, J. Dawber, J. Hill, and A. Clegg, *Cost-effectiveness of levodopa-carbidopa intestinal gel in treating people with advanced parkinson's disease*, British Journal of Neuroscience Nursing, 19 (2023), pp. 140–144.

[46] V. M. Ramírez, V. Kmetzsch, F. Forbes, and M. Dojat, *Deep learning models to study the early stages of parkinson's disease*, in 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 2020, pp. 1534–1537.

[47] M. I. Razzak, S. Naz, and A. Zaib, *Deep learning for medical image processing: Overview, challenges and the future*, Classification in BioApps: Automation of Decision Making, (2018), pp. 323–350.

[48] C. Rubbert, C. Mathys, C. Jockwitz, C. J. Hartmann, S. B. Eickhoff, F. Hoffstaedter, S. Caspers, C. R. Eickhoff, B. Sigl, N. A. Teichert, et al., *Machine-learning identifies parkinson's disease patients based on resting-state between-network functional connectivity*, The british journal of radiology, 92 (2019), p. 20180886.

[49] M. Ryan, C. V. Eatmon, and J. T. Slevin, *Drug treatment strategies for depression in parkinson disease*, Expert opinion on pharmacotherapy, 20 (2019), pp. 1351–1363.

[50] L. Sahu, R. Sharma, I. Sahu, M. Das, B. Sahu, and R. Kumar, *Efficient detection of parkinson's disease using deep learning techniques over medical data*, Expert Systems, 39 (2022), p. e12787.

[51] S. Sangeetha, K. Baskar, P. Kalaivaani, and T. Kumaravel, *Deep learning-based early parkinson's disease detection from brain mri image*, in 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), 2023, pp. 490–495.

[52] S. Shinde, S. Prasad, Y. Saboo, R. Kaushick, J. Saini, P. K. Pal, and M. Ingalhalikar, *Predictive markers for parkinson's disease using deep neural nets on neuromelanin sensitive mri*, NeuroImage: Clinical, 22 (2019), p. 101748.

[53] S. Sivaranjini and C. Sujatha, *Deep learning based diagnosis of parkinson's disease using convolutional neural network*, Multimedia tools and applications, 79 (2020), pp. 15467–15479.

[54] S. Sveinbjornsdottir, *The clinical symptoms of parkinson's disease*, Journal of neurochemistry, 139 (2016), pp. 318–324.

[55] A. S. Talai, J. Sedlacik, K. Boelmans, and N. D. Forkert, *Utility of multi-modal mri for differentiating of parkinson's disease and progressive supranuclear palsy using machine learning*, Frontiers in Neurology, 12 (2021), p. 648548.

[56] M. Thomas, A. Lenka, and P. Kumar Pal, *Handwriting analysis in parkinson's disease: current status and future directions*, Movement disorders clinical practice, 4 (2017), pp. 806–818.

[57] E. Tolosa, A. Garrido, S. W. Scholz, and W. Poewe, *Challenges in the diagnosis of parkinson's disease*, The Lancet Neurology, 20 (2021), pp. 385–397.

[58] T. Tuncer, S. Dogan, and U. R. Acharya, *Automated detection of parkinson's disease using minimum average maximum tree and singular value decomposition method with vowels*, Biocybernetics and Biomedical Engineering, 40 (2020), pp. 211–220.

[59] M. Ugrumov, *Development of early diagnosis of parkinson's disease: Illusion or reality?*, CNS neuroscience & therapeutics, 26 (2020), pp. 997–1009.

[60] A. ul Haq, J. P. Li, B. L. Y. Agbley, C. B. Mawuli, Z. Ali, S. Nazir, and S. U. Din, *A survey of deep learning techniques based parkinson's disease recognition methods employing clinical data*, Expert Systems with Applications, 208 (2022), p. 118045.

[61] J. C. Vásquez-Correa, T. Arias-Vergara, J. R. Orozco-Arroyave, B. Eskofier, J. Klucken, and E. Nöth, *Multimodal assessment of parkinson's disease: a deep learning approach*, IEEE journal of biomedical and health informatics, 23 (2018), pp. 1618–1630.

[62] I. K. Veetil, E. Gopalakrishnan, V. Sowmya, and K. Soman, *Parkinson's disease classification from magnetic resonance images (mri) using deep transfer learned convolutional neural networks*, in 2021 IEEE 18th India Council International Conference (INDICON), IEEE, 2021, pp. 1–6.

[63] X. Wang, Y. Zhang, C. Zhu, G. Li, J. Kang, F. Chen, and L. Yang, *The diagnostic value of snpc using nm-mri in parkinson's disease: meta-analysis*, Neurological Sciences, 40 (2019), pp. 2479–2489.

[64] Y. Wang, N. He, C. Zhang, Y. Zhang, C. Wang, P. Huang, Z. Jin, Y. Li, Z. Cheng, Y. Liu, et al., *An automatic*

interpretable deep learning pipeline for accurate parkinson's disease diagnosis using quantitative susceptibility mapping and t1-weighted images, Tech. Report 12, Wiley Online Library, 2023.

[65] Y.-H. Weng, T.-C. Yen, M.-C. Chen, P.-F. Kao, K.-Y. Tzen, R.-S. Chen, S.-P. Wey, G. Ting, and C.-S. Lu, *Sensitivity and specificity of 99mtc-trodat-1 spect imaging in differentiating patients with idiopathic parkinson's disease from healthy subjects*, Journal of Nuclear Medicine, 45 (2004), pp. 393–401.

[66] R. Xia and Z.-H. Mao, *Progression of motor symptoms in parkinson's disease*, Neuroscience bulletin, 28 (2012), pp. 39–48.

[67] K. Yasaka, K. Kamagata, T. Ogawa, T. Hatano, H. Takeshige-Amano, K. Ogaki, C. Andica, H. Akai, A. Kunimatsu, W. Uchida, et al., *Parkinson's disease: Deep learning with a parameter-weighted structural connectome matrix for diagnosis and neural circuit disorder investigation*, Neuroradiology, (2021), pp. 1–12.

[68] T. A. Zesiewicz, *Parkinson disease*, CONTINUUM: Lifelong Learning in Neurology, 25 (2019), pp. 896–918.

[69] X. Zhang, Y. Yang, H. Wang, S. Ning, and H. Wang, *Deep neural networks with broad views for parkinson's disease screening*, in 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019, pp. 1018–1022.

Edited by: Chiranji Lal Chowdhary
Special issue on: Scalable Machine Learning for Health Care: Innovations and Applications
Received: Oct 31, 2023
Accepted: Dec 29, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1424–1441, DOI 10.12694/scpe.v25i3.2777

EHEALTH INNOVATION FOR CHRONIC OBSTRUCTIVE PULMONARY DISEASE: A CONTEXT-AWARE COMPREHENSIVE FRAMEWORK

ANAM IQBAL[*], SHAIMA QURESHI[†], AND MOHAMMAD AHSAN CHISHTI[‡]

Abstract. Chronic Obstructive Pulmonary Disease (COPD) poses a significant global healthcare challenge. It is a progressive lung disease that causes breathing difficulties and can significantly impact a person's quality of life. COPD is primarily caused by smoking, but other factors, such as air pollution and genetic predisposition, can also contribute to its development. This paper introduces a novel Context-Aware Framework for the Diagnosis and Personalized Management of COPD. We discuss the limitations of traditional COPD management, highlighting the importance of early detection and remote monitoring. Early detection and remote monitoring are crucial in managing COPD as they allow for timely interventions and better disease management. In this paper, we propose a framework based mostly on contextual data and other parameters of COPD as put forth by the World Health Organization (WHO) in the form of the International Classification of Functioning, Disability, and Health.

Ontologies drive this architecture and incorporate dynamic contextual information from patient environments, user profiles, and sensor data. In addition to the various obvious data items like patient personal details (gender, contact, medical history) and COPD risks and symptoms, the COPD ontology also considers the details about the caregiver and healthcare professional. This is in addition to the contextual data processed separately using the Context Ontology. The ontology we constructed using Protégé serves as the framework for the structured representation and logical inference of contextual information. By harnessing dynamic contextual data, our ontology enables real-time decision-making tailored to individual patient requirements. It empowers healthcare professionals to make informed choices and deliver timely interventions, enhancing healthcare services by offering proactive care to detect early signs of health deterioration and suggest preventive measures. This approach improves patient experiences and optimizes resource allocation within the healthcare system. To uphold ethical standards and prioritize the needs of patients, we emphasize the significance of safeguarding data, obtaining informed permission, and recognizing data ownership. The ontology-based approach presented in this study offers a scalable and flexible framework that can be readily incorporated into existing healthcare systems, redefining the management of COPD in response to evolving demands. Security poses one of the biggest threats in context-based environments due to the different data formats acquired by the diverse sensors. Another essential consideration is confidentiality because the data in hand is sensitive patient information.

Key words: COPD, Ontology, Context-Aware, Healthcare Management, Sensor Data, Personalization, Ethical Considerations, Remote Monitoring.

1. Introduction.

1.1. Understanding COPD. COPD is an acronym that stands for Chronic Obstructive Pulmonary Disease. It is a lower respiratory tract disease [1],[2]. It is a broad term to describe a widespread and relentless chronic pulmonary disease characterized by irreversible airflow limitation and inflammatory disorders. It is an umbrella term for certain chronic lung diseases [3]. It includes chronic bronchitis, emphysema, and refractory asthma. COPD is life-limiting and irreversible; It can be prevented and some of its symptoms treated [4]. The repercussions of COPD are not limited to the individual patient but substantially burden the healthcare system. COPD is a worldwide problem, affecting people of all ages and backgrounds [5],[6]. It ranks among the leading causes of mortality globally [7],[8]. COPD is primarily caused by long-term exposure to irritants that damage the lungs and airways [9]. The most common risk factors are tobacco consumption and smoking, but environmental factors like air pollution and occupational dust exposure can also contribute [10]. Notably, COPD primarily affects individuals in mid-life or later, with a disproportionately higher incidence among

[*] Department of Computer Science and Engineering, National Institute of Technology, Hazratbal, Srinagar, Jammu and Kashmir, India (anam_03phd19@nitsri.net)

[†] Department of Computer Science and Engineering, National Institute of Technology, Hazratbal, Srinagar, Jammu and Kashmir, India (shaim@nitsri.net)

[‡] Department of Computer Science and Engineering, National Institute of Technology, Hazratbal, Srinagar, Jammu and Kashmir, India (ahsan@nitsri.net)

Table 1.1: Limitations of Traditional COPD Management Approaches.

Limitations	**Details**	**Research Questions**	**Solution**
Lack of Personalization	The traditional treatment plan does not consider the unique characteristics of every individual patient.	The benefit of personalization, parameters considered for personalization, and technical support required.	Gather patient-specific and context data.
Dependency of Occurrence of Symptoms	The traditional treatment plan relies on intervention when the symptoms worsen.	Shifting from symptom-based to proactive COPD management, biomarkers to identify symptoms.	Use predictive analysis for early intervention in diagnosis and treatment.
Dependency on Healthcare professionals and Lack of Self-management functionalities	Patients may need to gain the knowledge or tools to manage their condition effectively and depend on support from healthcare professionals.	Technology assistance required for self-management.	Introduce remote support options to patients.
Restricted Use of Technology	Conservative treatment and diagnosis methods do not utilize the full potential of newer available technologies.	Identify technological support that can revolutionize COPD management and identify the privacy and ethical considerations when using patient data for COPD management.	Make IoMT an integral part of our healthcare system while taking due care of ethical considerations, patient consent, and data protection.

males [11]. The insidious influence of risk factors, such as tobacco, accelerates the decline of pulmonary capacity, obstructs airways, and exacerbates symptoms [12]. These symptoms, including chronic cough, sputum production, and dyspnea, introduce substantial functional limitations, disrupting patients' day-to-day lives [13]. Health professionals use various tools, including questionnaires and lung function tests, to identify those at risk or in the early stages of COPD. This early detection is vital for timely intervention [14]. Exacerbations, defined as acute worsening of symptoms and lung function, are pivotal events in the trajectory of COPD [15]. They Yield a significant influence, negatively affecting patients' quality of life, hastening the rate of lung function decline, and exacting a substantial socioeconomic toll [16]. The frequency of exacerbations escalates as the disease progresses, increasing hospital visits and healthcare resource utilization.

1.2. The Need for Early Detection and Remote Management of COPD.. Highlighting the significance of early detection and management of COPD, particularly exacerbations, cannot be emphasized enough. Early intervention at the onset of COPD prevents disease progression, preserves lung function, minimizes symptoms, and plays a pivotal role in reducing exacerbations [19]. These critical events, often triggered by infections or environmental factors, can lead to severe health deterioration if not addressed promptly [26]. By detecting exacerbations early, healthcare professionals can initiate timely treatments, such as antibiotics and steroids, averting exacerbation-related hospital admissions and the associated decline in health status. Adopting such guidelines will enable us to improve the patient's quality of life by improving their health and reducing the costs of maintaining a healthy life. During the last four years, since the onset of Coronavirus 2019 (COVID-19), a lot of research has been conducted on respiratory disorders. In [27], authors discuss how, for conducting research for COVID-19, a huge number of research laboratories have been set up and dedicated to pulmonary disorders. Due to the pandemic, the interdisciplinary sciences saw a boost, with Artificial Intelligence (AI) and Machine Learning (ML) rapidly becoming integrated into biomedical systems. Due to lockdowns worldwide, researchers have collaborated over the cloud, utilising pooled resources. All these factors have immensely impacted the research related to COPD as well. The E-learning tools that were not very prominent before COVID-19 were adopted;hence, researchers from medical fields were also exposed to these tools. This enables them to use these better to collaborate and share the data regarding COVID-19 across domains. [28]

1.3. Significance. The main requirement of a sustainable healthcare system nowadays is to be more personalized toward patient needs and require less intervention from the patient/caregiver. Both these can be achieved by putting forth an architecture for COPD management that is contextually aware. It can be a novel approach for real-time monitoring and personalization of COPD management. We have designed the architecture to integrate dynamic contextual information from the environment and details about the patient with actual patient data to make timely and more informed decisions specific to the particular patient's needs. The implementation will also alleviate the burden on healthcare professionals and the healthcare system.

1.4. Research Objectives. The primary objective of this research setup is to put forth a context-aware framework for real-time monitoring and personalized management of COPD. This requires comprehensively detailing the architecture's components, methodologies, and potential implications for COPD diagnosis and risk assessment. The objectives include understanding context-aware systems and their application in healthcare, particularly in COPD management; examining the structure and components of the scalable context-aware architecture; and evaluating the effectiveness of ontology-based methods for reasoning and modeling in enhancing COPD diagnosis and risk assessment. The forthcoming sections of this paper will delve deeper into the key components and methodologies that underpin the scalable context-aware architecture. The following section is the literature review, which explores the context awareness applications in COPD management and the details of the International Classification of Functioning, Disability and Health (ICF) model. The Methodology Section includes an in-depth examination of the ontology, implemented using Protégé in Web Ontology Language (OWL) format. This project seeks to illuminate the architecture's potential impact on COPD diagnosis and risk assessment through systematic analysis.

2. Literature Review.

2.1. Comparative Analysis of Some Existing Works. Table 2.1 illustrates some of the existing works on Ontologies for COPD.

2.2. The International Classification of Functioning, Disability and Health. The ICF is a comprehensive framework developed by the WHO[39]. It is used to classify various health conditions and understand the disabilities related to such conditions. It is a universal language and conceptual basis for understanding and measuring health and disability across various disciplines.

2.2.1. Framework Overview. The ICF emphasizes a shift from negative connotations like disability and focuses on an individual's function and positive abilities. It does not classify people but provides a framework for assessing functioning, promoting better communication, facilitating data comparison, and serving as a coding system for health information.

2.2.2. Fundamental Principles and Components. Four fundamental principles underlie the ICF: universality, parity, neutrality, and environmental influence. It categorizes functioning and disability into four main components: Body Functions and Structures, Activities and Participation, and Personal and Environmental Factors [41]. Figure 2.1 shows one of the representations of the disability model for any disease.

a. Body Functions and Structures: This component addresses the physiological functions of body systems, including psychological functions and anatomical parts of the body [43].
b. Activities and Participation: Activities refer to the execution of tasks or actions by an individual [44].
c. Personal and Environmental Factors: Personal factors are considered but not classified within the ICF framework. Environmental factors encompass the physical, social, and attitudinal environment in which people live and conduct their lives [45].

2.2.3. ICF Core Sets. ICF Core Sets are practical tools developed for clinical practice to comprehensively describe functioning in specific patient populations. They help healthcare professionals better understand the needs of patients, particularly those with chronic diseases, and aid in clinical assessment and treatment planning. In summary, the ICF is a versatile framework that promotes a positive, holistic approach to assessing health and disability by considering various aspects of an individual's functioning within their unique context [46]. It plays a crucial role in healthcare, disability services, education, social policy, and more, emphasizing the

Table 2.1: Comparison of some existing Ontology Based Research's.

Paper	Problem Statement	Approach	Data	Evaluation (if any)	Conclusion
[29]	Need for personalized COPD monitoring and recommendations	Rule-based ontology for COPD patients	Rules based on biomarkers, indoor/outdoor conditions, simulated dataset	Confusion matrix analysis, potential for telemonitoring enhancement	Context-aware system for COPD
[30]	Enhancing COPD patient management	COPDology for proactive management	Reuse of Systematized Nomenclature of Medicine Clinical Terminology (SNOMED CT) and Global Medical Device Nomenclature (GMDN), extensive ontology	Not provided	A significant tool for COPD management
[93]	Inaccurate self-identification of Acute exacerbations of chronic obstructive pulmonary disease (AECOPDs)	A machine learning algorithm for AECOPD prediction	Data collection, machine learning algorithm	Exceptional accuracy, compared to pulmonologists	A promising tool for COPD triage
[28]	Reducing AECOPD-related hospitalization	Remote telemonitoring for AECOPDs	Data collection, ontology development, structural aspects	Mixed evidence on effectiveness	Potential for AECOPD prediction and management
[31]	Handling diverse Internet of Things (IoT) healthcare data	Ontology-based approach for IoT healthcare	Ontology development, SPARQL Protocol and RDF Query Language (SPARQL) queries	Framework for data heterogeneity	Framework for Cardiovascular Disease Diagnosis
[32]	Achieving semantic interoperability in IoT healthcare	Resource Description Framework (RDF) and RDF Mapping Language (RML) for IoT Healthcare Semantic Interoperability	Data collection, RDF mapping, SPARQL queries	Experiments using RDF mapping	RML-based approach for IoT Healthcare Semantic interoperability
[33]	Autonomous model for predicting COPD exacerbations	Machine learning with Bayesian networks, attribute selection, discretization	61 attributes and 1985 COPD patients dataset	Area under the Receiver Operating Characteristic (ROC) curve Area under the ROC Curve (AUC)	Promising autonomous model for COPD exacerbation prediction
[34]	Ontology, telemonitoring, COPD, physiological data, environmental parameters	Efficient self-management and early detection of exacerbations	Ontology-based telemonitoring for COPD patients	Design, experiments, evaluation, and validation	Importance of validation, performance measurement metrics
[35]	Dynamic detection model, ontology, COVID-19 symptoms, COPD patients	Early detection of COVID-19 symptoms in COPD patients	Ontology-based dynamic detection model	Data from questionnaire answers, simulation, and implementation	Prototype results compared to actual patient outcomes
[36]	IoT, Healthcare Information Systems, Semantic Web, Ontology	Semantic Web in IoT-based healthcare information systems	Development of ontologies for medical devices and health domain [37]	Description and processing of data using ontologies, semantic rules	Not provided

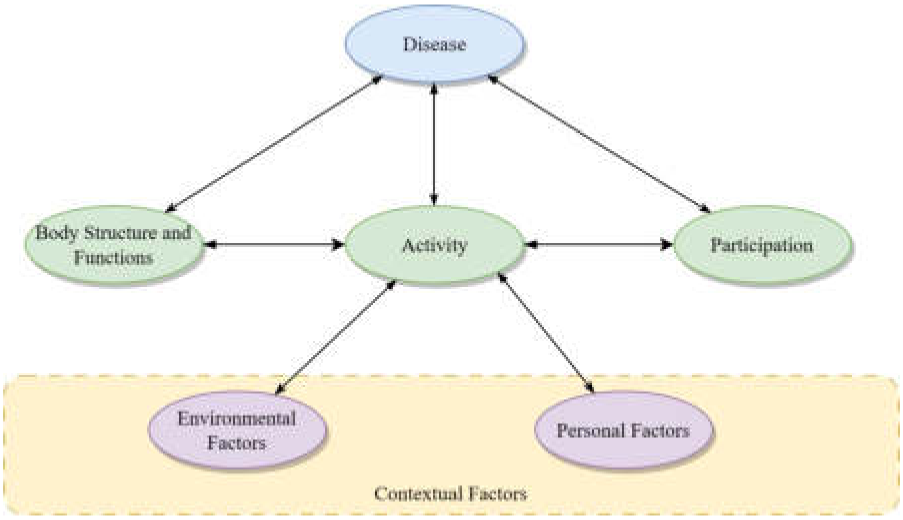


Fig. 2.1: International Classification of Functioning, Disability and Health Model by World Health Organization.

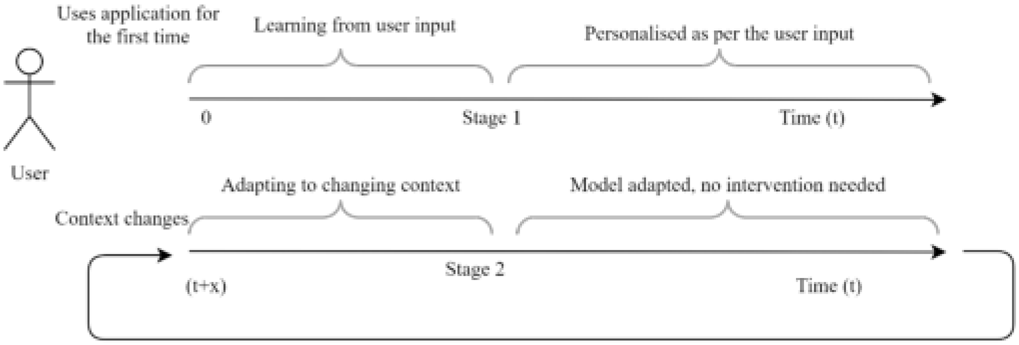

Uses application for the first time
Learning from user input
Personalised as per the user input
User
0
Stage 1
Time (t)
Adapting to changing context
Model adapted, no intervention needed
Context changes
(t+x)
Stage 2
Time (t)

Fig. 3.1: User interaction/intervention with a Generalised Self Adaptive Context Based System.

importance of ethical and patient-centered application. We have developed the COPD ontology based on the ICF identified for COPD.

3. Proposed Framework. Figure 3.1 depicts a sequence of steps in operating a context-based personalized human activity recognition system. The main stages of the process are highlighted, showing the user's interaction with the application and how it learns and adapts over time. It shows a generalized process flow of the user's interaction with the application that a context-aware personalized human activity recognition system can follow. It explains how the application adapts over time.

1. Initialisation (Time 0): We assume this is a first-time user who has never interacted with the system in the scenario we consider. This means no context or any other data about the user is present in the system. On interacting with the application, the user provides specific input, such as preferences, behavior patterns, or other relevant data, which serves as the primary input, and other details like the user profile and environmental details act as the contextual data.
2. Learning and Personalisation: The application processes this input and builds a personalized model based on the user's information, preferences, habits, and activity patterns. The responses provided by the application are tailored to the user's input and preferences.

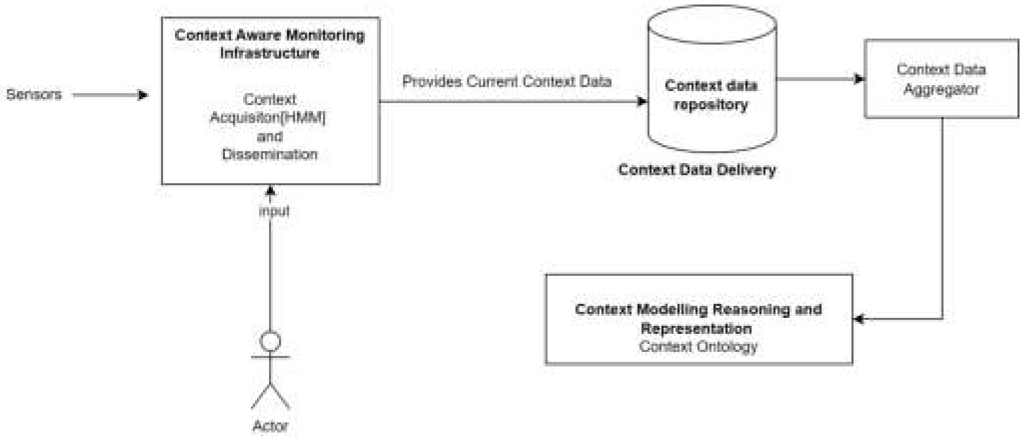

Sensors
Context Aware Monitoring Infrastructure
Context Acquisiton[HMM] and Dissemination
Provides Current Context Data
Context data repository
Context Data Delivery
Context Data Aggregator
input
Actor
Context Modelling Reasoning and Representation
Context Ontology

Fig. 3.2: Proposed Contextual Information Engine.

3. Adaptation to Changing Context (Time t+x): As time progresses, the user's context changes. This might include changes in location, behavior, or preferences. We assume that at the time (t+x), the user engages in a new activity or behavior that differs from the initial input.
4. Model Adaptation The application's model recognizes the changing context and adapts to the new information. The personalized model is updated to incorporate the latest user behavior and preferences. This adaptation ensures that the application's responses remain relevant and accurate. Once the model adapts to the changing context, the application provides accurate and personalized responses that align with their current context and preferences without requiring further intervention from the user.

3.1. Context-Aware Model for COPD Management. The architectural diagram in Figure 3.2 represents the proposed Contextual Information Engine, the main processing system of the Context-Aware COPD Model, and the workflow given in Figure 3.3. Context data is acquired from user data as well as sensed data. This enables us to provide more personalized, dynamic, and effective healthcare services. This context data is fed to a Context-Aware Monitoring Infrastructure.

1. Inputs from sensors and users/actors are sent to the Context-Aware Monitoring Infrastructure.
   a. Sensors: These are data-gathering devices that capture data from the environment, like health-related information, such as vital signs, environmental data, and patient activity.
   b. User/Actor: These individuals interact with the healthcare system, such as patients, doctors, and caregivers. They act as an important source of contextual data.
2. Context-Aware Monitoring Infrastructure: The Context-Aware Monitoring Infrastructure manages the context information. It consists of two main components: Context Acquisition and Context Dissemination.
   a. Context Acquisition: This component receives input from sensors and users/actors. It gathers sensor data and captures user interactions, creating a holistic context.
   b. Context Dissemination: The acquired context is processed and prepared for further analysis. This component ensures that relevant contextual information is efficiently disseminated for subsequent stages.

      Current Context Data: The processed context data from the monitoring infrastructure is considered the "current context data." This is stored in the Context Data Repository. This information is essential for making informed decisions and providing context-aware services.
3. Context Data Repository (Context Data Delivery): The context data collected over time, called the "current context," is stored here, forming a repository of historical context information. This repository is a centralized storage for the contextual information gathered from sensors and users. It ensures data integrity and accessibility for downstream processes. The "Context Data Repository" feeds the

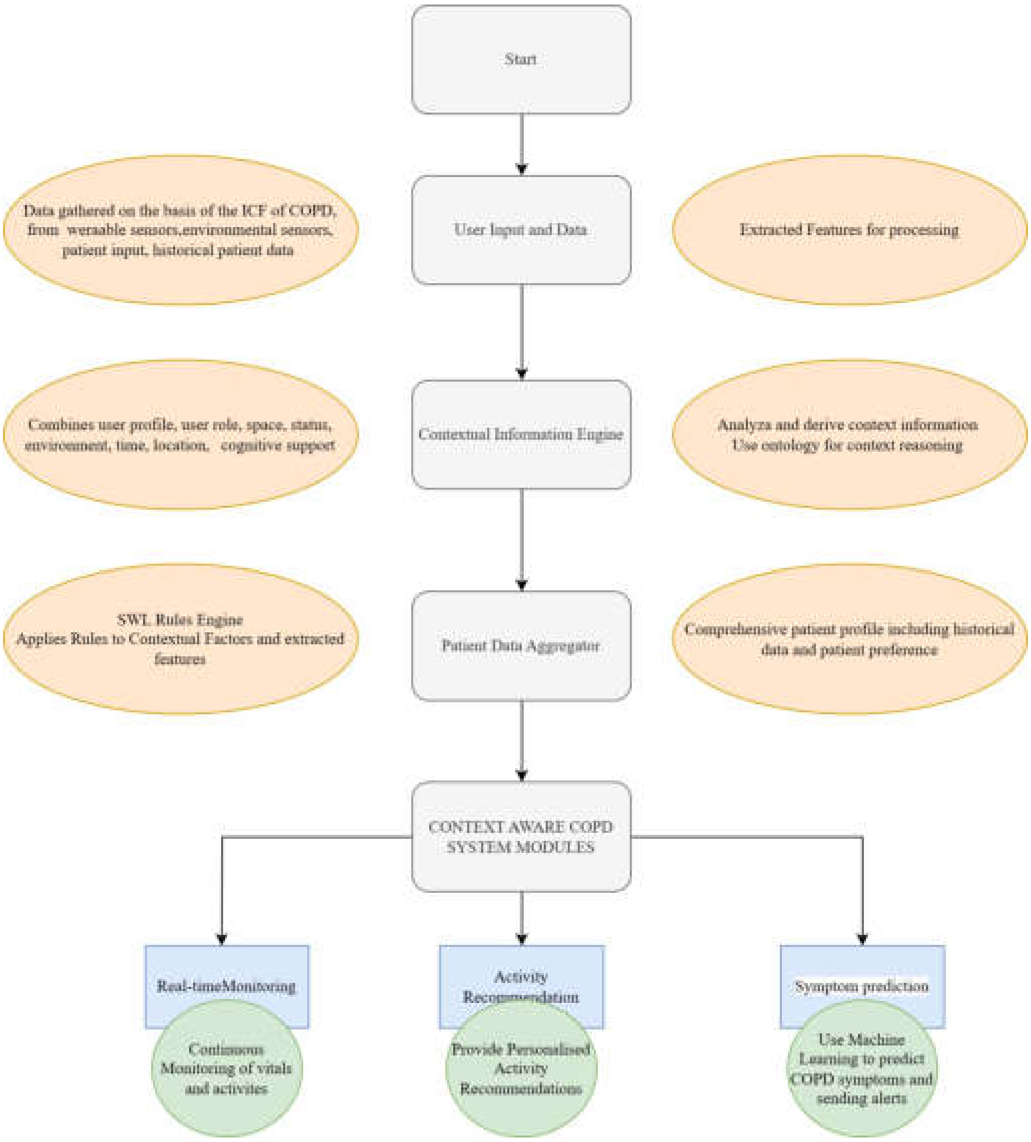

Start
Data gathered on the basis of the ICF of COPD, from weraable sensors,environmental sensors, patient input, historical patient data
User Input and Data
Extracted Features for processing
Combines user profile, user role, space, status, environment, time, location, cognitive support
Contextual Information Engine
Analyza and derive context information Use ontology for context reasoning
SWL Rules Engine Applies Rules to Contextual Factors and extracted features
Patient Data Aggregator
Comprehensive patient profile including historical data and patient preference
CONTEXT AWARE COPD SYSTEM MODULES
Real-timeMonitoring
Continuous Monitoring of vitals and activites
Activity Recommendation
Provide Personalised Activity Recommendations
Symptom prediction
Use Machine Learning to predict COPD symptoms and sending alerts

Fig. 3.3: Workflow of the Proposed Context-Aware COPD Model

contextual information to the ”Context Data Aggregator.”

4. Context Data Aggregator: This component processes and compiles context data, possibly from multiple sources, before sending it forward. It helps consolidate information to view the patient’s state and environment comprehensively. The compiled context data is forwarded to the ”Context Modeling, Reasoning, and Representation System for advanced processing.
5. Context Modeling, Reasoning, and Representation System: This crucial system processes the aggregated context data. It performs modeling, reasoning, and representation using ontology methodology, enabling a structured and organized way to interpret the context.

3.2. Ontology Methodology. This methodology is the foundation for modeling and organizing the context information in a standardized and meaningful manner. It defines classes, properties, and relationships among various contextual factors [47]. It creates a semantic model that captures the relationships between different context elements. Semantic interoperability enables data from different sources to be seamlessly integrated and understood [26], enabling effective reasoning and inference. This becomes crucial in healthcare, where diverse data types and sources must be harmonized for comprehensive analysis. This involves structuring the context data into meaningful entities, relationships, and attributes [48]. The output of the Context Modeling, Reasoning, and Representation System can be used for various purposes. It can support decision-making, provide insights, and facilitate personalized healthcare services based on the current context [49]. This system transforms raw data into actionable information. This system does not operate in isolation. It feeds its insights and conclusions to various parts of the healthcare service system, enabling dynamic adjustments and improvements. This closed-loop approach ensures that the system continually refines its understanding of context, leading to more accurate and relevant outcomes [50].

3.3. Advantages of Context-Aware COPD Model.

1. Intelligent Decision-Making and Personalisation: The gathered context data lays the basis for an intelligent decision-making system since each patient's context is very specific [51]. Ontological modeling allows us to infer relationships between the context data gathered from different sources [53]. Collaboration between healthcare experts and technology professionals is essential for designing an architecture that aligns with medical best practices. Their combined expertise ensures that the context data collected and processed aligns with clinical needs and priorities [52].
2. Improved Healthcare Services and Outcomes: Healthcare providers can offer proactive, personalized care by incorporating real-time context data [53]. This approach leads to improved patient experiences and better clinical outcomes. The healthcare system can allocate resources more efficiently by understanding the patient's context [54]. This includes optimizing bed utilization, staffing, and medical supplies based on anticipated needs [51].
3. Continuous Learning and Optimisation: The architecture also supports continuous learning and optimization through data feedback loops [55]. As the system interacts with more patients and accumulates data, it can refine its understanding of context and improve its decision-making capabilities [52]. Also, feedback from healthcare professionals and system performance data is invaluable for refining the system's functionality and ensuring its relevance over time.

   Machine Learning Integration [56]: As the architecture accumulates more data over time, the system can integrate machine learning algorithms to recognize patterns and trends within the context data. This can lead to better prediction of health events and further optimization of care plans. Telemedicine Enhancement: The architecture can significantly enhance telemedicine capabilities. Patients can securely share context data with remote healthcare providers, enabling accurate diagnoses and treatment recommendations [57].
4. Long-Term Sustainability: By focusing on long-term sustainability, we ensure that the architecture remains relevant, adaptable, and effective in addressing the changing needs of healthcare systems. This includes planning for hardware and software upgrades, scalability, and maintenance.
5. User-Centered Design: The architecture should prioritize user needs and experiences. User-centered design principles can lead to intuitive, user-friendly interfaces aligned with the diverse requirements of patients, caregivers, and healthcare professionals [59]

   The "Context-Aware Architecture for COPD" presents a transformative approach to healthcare delivery. It offers tailored, real-time, and data-driven healthcare solutions by harnessing the power of context data from sensors and users. While implementation challenges exist, a well-designed, user-centric, and ethically sound approach can lead to a future where healthcare is truly context-aware, improving patient outcomes and overall well-being.

3.3.1. Implementation Challenges.

1. Interoperability: Integrating diverse sensors, devices, and data sources may present challenges in terms of standardization and interoperability. Ensuring seamless communication and data exchange is crucial for the architecture's success.
2. Data Quality: The accuracy and reliability of the collected data impact the architecture's effectiveness. Measures to handle noisy or erroneous data, as well as calibration of sensors, must be considered.
3. Contextual Complexity: Contextual information can encompass a wide range of factors. Developing

Table 4.1: Different Ontology models and our identified contexts for COPD Ontology

Ontology	**R**eference	**P**rofile	**R**ole	**S**pace	**S**tatus	**E**nvironment
Context Broker Ontology (COBRAONT)	[70]	Yes	Yes	Yes	No	Yes
Context Ontology Language (CoOL)	[71]	Yes	No	No	No	No
Ontology server	[72]	Yes	Yes	Yes	Yes	Yes
Mobile sensor context ontology	[73]	Yes	No	No	Yes	Yes
Context-Driven Adaptation of Mobile Services (CoDAMos)	[75]	Yes	Yes	Yes	No	Yes
OWL encoded context ontology (CONoN)	[75]	Yes	No	Yes	Yes	No
Standard ontology for ubiquitous and pervasive applications (SOUPA)	[76],[77]	Yes	No	Yes	No	No
Situation Ontology	[78]	Yes	No	Yes	No	Yes
Delivery Context Ontology	[79]	No	No	Yes	No	Yes
Multidimensi- onal Integrated Ontologies (mIO!)	[80]	Yes	Yes	Yes	No	Yes
PiVOn(n.a)	[81]	Yes	No	No	No	Yes
Health context ontology	[82]	Yes	Yes	Yes	Yes	Yes
PaISPOT(n.a)	[83]	Yes	No	Yes	Yes	Yes
Rover Context Model Ontology (RoCoMo)	[84]	Yes	Yes	Yes	Yes	Yes
Meta Context Ontology Model (McOnt)	[85]	Yes	No	Yes	No	Yes
Smarton tosensor	[86]	Yes	Yes	Yes	No	Yes
Context awareness meta ontology modeling (CAMeOnto)	[87]	Yes	Yes	Yes	Yes	Yes
Extensible Context Ontology for Persuasive Physical-Activity Applications (ECOPPA)	[88]	Yes	No	Yes	Yes	No
Multimedia Semantic Sensor Network Ontology (MSSN-Onto)	[89]	Yes	No	Yes	No	No

sophisticated algorithms and models for context representation and reasoning is essential to capture this complexity accurately.

4. Ethical, Security, and Privacy Considerations: Amid the sensitive nature of healthcare data, the architecture strongly emphasizes ethical regulations, security, and privacy [60] Robust security measures ensure that patient information remains confidential and protected from unauthorized access [62].
   - Informed Consent: Users should be informed about the data collection and usage practices and give their consent. Transparent communication builds trust between the healthcare system and its users. [64].
   - Data Ownership [65].
   - Data Encryption [61], [64], [66]
   - Access Control [67].
   - Data Privacy [65],[68]
   - Unauthorized Access to Patient's Data [69].

4. Methodology Used. As discussed in the previous subheading, ”Context-Aware Model for COPD Management,” we use Ontology Modeling to implement the Context-Aware COPD Model. In most ontology-based models, the context features considered are tabulated in Table 4.1. We identify an additional contextual feature, which is cognitive support. The two contextual features always taken into account are location and time. We devise certain competency questions to retrieve desirable information after the ontology's development [32]. For example, can our ontology provide insights about a patient's COPD progression, what risk factors the ontology can detect, or what emergency guide is available for addressing critical COPD patients?

Figure 4.1 represents the taxonomy of our Context-Aware COPD model. It gives a snapshot of the different classes related to COPD and the context classes we have identified, i.e., profile, role, status, space,

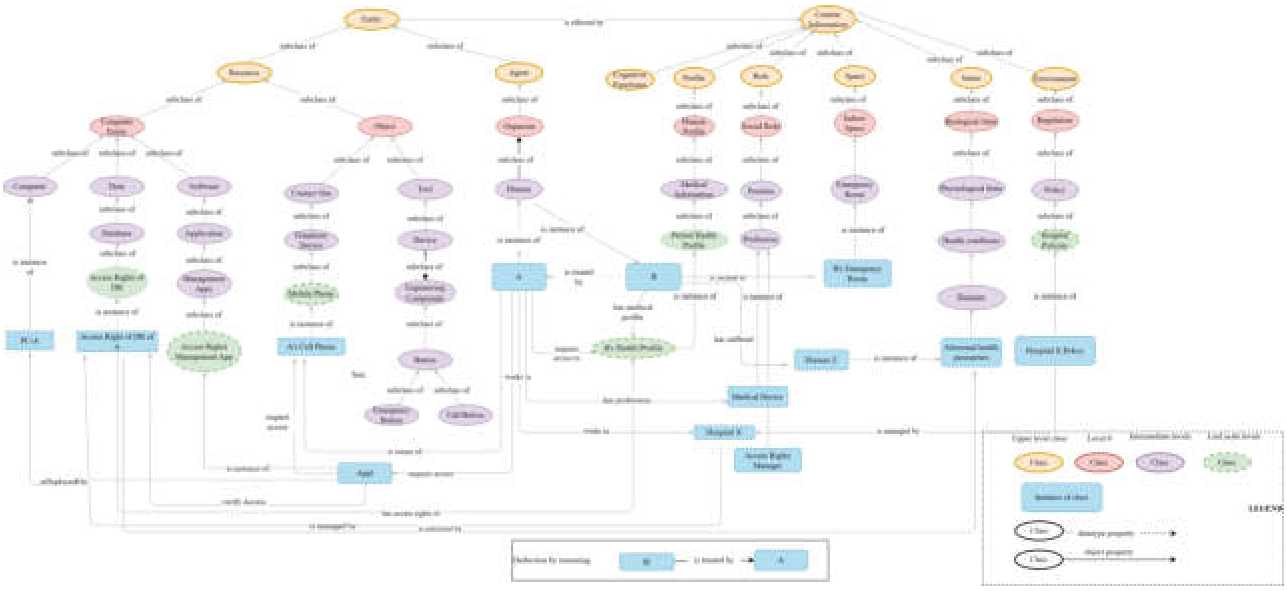


Fig. 4.1: Taxonomy of Context-Aware COPD Ontology Model

and environment.

4.1. Dataset. For our application, setting up an environment for data acquisition from sensors is difficult due various practical limitations like ethical and privacy issues, cost actor, and deployment time. The alternative method is to use an intelligent simulation method. This simulation is based on the ICF designed for COPD by the WHO. Even though a data availability statement is not required, [97],[98], and [99] helped us to design the COPD ontology. We also used [100], for designing the Context-aware ontology.

4.2. Ontology Development.

4.2.1. Purpose of the Developed Ontology. We identify why we are developing this ontology and identify its primary functions. We also point out the focus of the ontology. Purpose Implemented: Patient Self-Management The idea of patient self-management came to light during COVID-19 when there were quarantines, and there was no access to caregivers for the patients. [95] The purpose is to allow the patients and caregivers to take an active role in managing their COPD. So, a context-aware ontology is prepared to encompass two ontologies based on the ICF of COPD, categorizing raw data and context data inputs separately. By enabling the patients to know their condition better, we allow them to make informed decisions about their choice of activities and self-care strategies. Other purpose scenarios to be considered:

1. Healthcare Professional Support: The ontology's function is to support the clinician in tailoring a treatment plan based on the patient's context.
2. Research Analysis: The ontology is research-friendly and used to identify and analyze the patterns in COPD data management; the focus is interventions required in different contexts.
3. Education and Awareness: This is also a patient-centered and patient-friendly ontology, focusing on proactive self-management under changing contextual factors.

4.2.2. Goals of the Developed Ontology. We identify the outcome within the context of our ontology's purpose. Goals associated with our purpose are:

1. Providing Personalized Recommendations: Since the purpose is self-management, the ontology must provide patient-centered recommendations based on their context.
   - Treatment recommendations
   - Activity recommendations

Table 4.2: COPD Ontology Development.

Development Stage	**Details**
Data Source	[90], [91], [92], [93], [94].
Ontology Structure	Figure 4.1
Ontology Tools	Protégé
Class Definitions	Figure 4.3, 4.4, 4.5
Semantic Relationships	Figure 4.7

Table 4.3: Context-Aware Ontology Development.

Development Stage	**Details**
Ontology Structure	Figure 4.8
Ontology Tools	Protégé
Class Definitions	Figure 4.9
Semantic Relationships	Figure 10

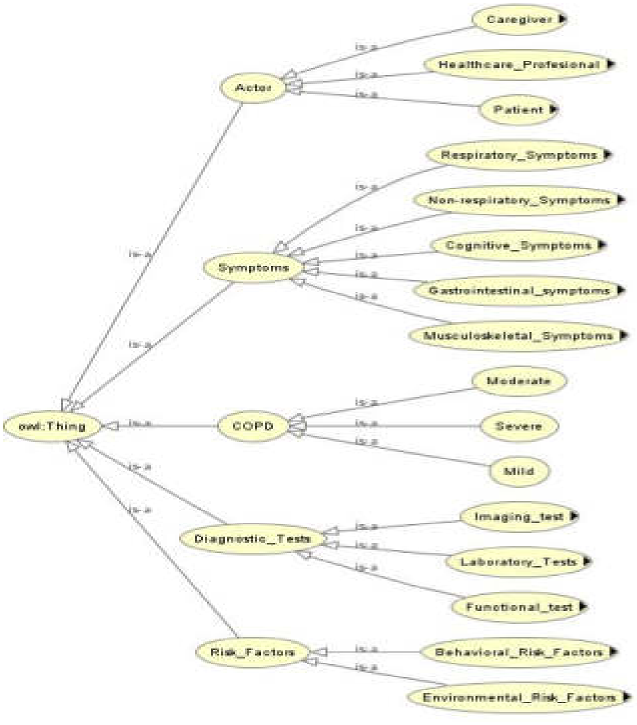


Fig. 4.2: Class Hierarchy in COPD Ontology.

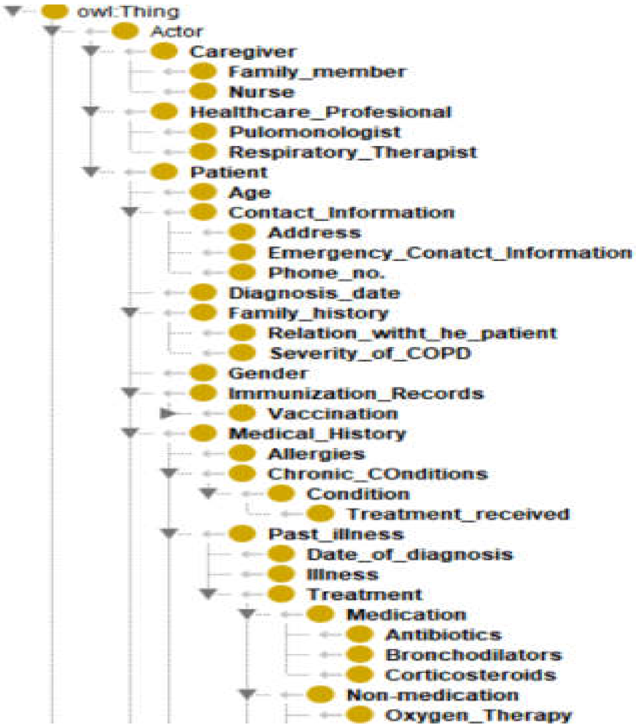


Fig. 4.3: Classes in COPD Ontology.

2. Symptom Prediction: For self-management, it is crucial to anticipate the potential symptoms based on historical data, current patient information, and contextual factors.
3. Administration and monitoring of medicines: This needs to consider the daily activities and medication schedules.
4. Response Planning in Critical Conditions / Emergency Response: We design rules to address COPD-related emergencies, checking available resources, including medications, caregiver support, healthcare professionals' availability, and emergency treatment preparedness.

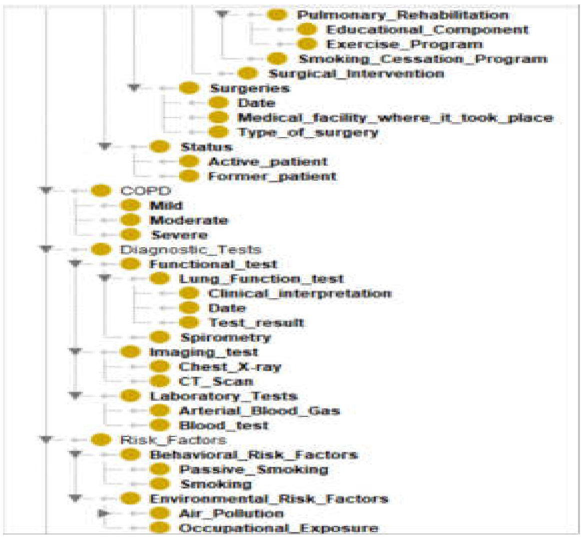


Fig. 4.4: Classes in COPD Ontology.

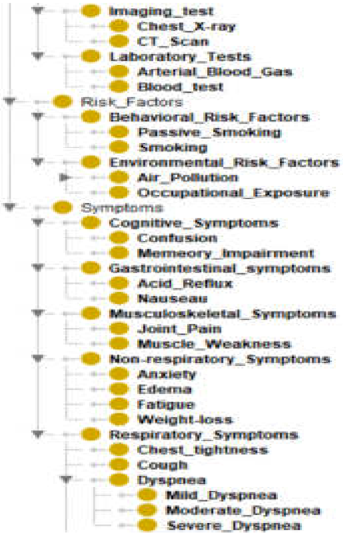


Fig. 4.5: Classes in COPD Ontology.

5. Seamless Lifestyle Integration: The lifestyle recommendations are closely related to the activity recommendations. This includes incorporating lifestyle choices and social and professional status as contextual factors.
6. Long-term Monitoring: We monitor the patients' symptoms over a long time to make informed decisions about any treatments required or modifications in the medication, treatment, or lifestyle.
7. Cognitive Support: We aim to support psychosocial well-being like anxiety management, and behavioral changes, like smoking cessation.

5. Conclusion. In conclusion, our research addresses the pressing challenges of managing COPD by developing a Context-Aware Ontology. COPD, a complex and pervasive pulmonary condition, demands innovative solutions beyond traditional approaches. Our paper has elucidated the limitations of conventional COPD management, emphasizing the critical need for personalized and proactive healthcare interventions. The core contribution of our work is the creation of a sophisticated Context-Aware Ontology designed to revolutionize COPD management. This ontology harnesses dynamic contextual data from patient environments, user profiles, and sensor inputs, enabling real-time decision-making that caters to individual patient requirements. Developed using Protégé, this ontology provides a structured framework for representing and reasoning about the multifaceted contextual elements influencing COPD care. The advantages of our context-aware COPD model are manifold. It facilitates intelligent decision-making by considering the unique context of each patient, empowering healthcare professionals to make informed choices and deliver timely interventions. Moreover, it enhances healthcare services by offering proactive care to detect early signs of health deterioration and suggest preventive measures. This approach improves patient experiences and optimizes resource allocation within the healthcare system.

Our architecture is not static; it supports continuous learning and optimization. As the system accumulates more data, it can integrate machine learning algorithms to recognize patterns and trends within the context data, leading to better predictions and further care plan optimization. Additionally, the architecture enhances telemedicine capabilities, enabling secure context data sharing with remote healthcare providers for accurate diagnoses and treatment recommendations. We underscore the importance of data security, informed consent, and ownership to ensure our context-aware COPD model's ethical and patient-centred application. These considerations are paramount in the sensitive realm of healthcare data, where maintaining patient privacy and security is non-negotiable. Our Context-Aware Ontology for COPD offers a trans-formative path forward in healthcare delivery. By embracing the power of context data from various sources, our approach provides

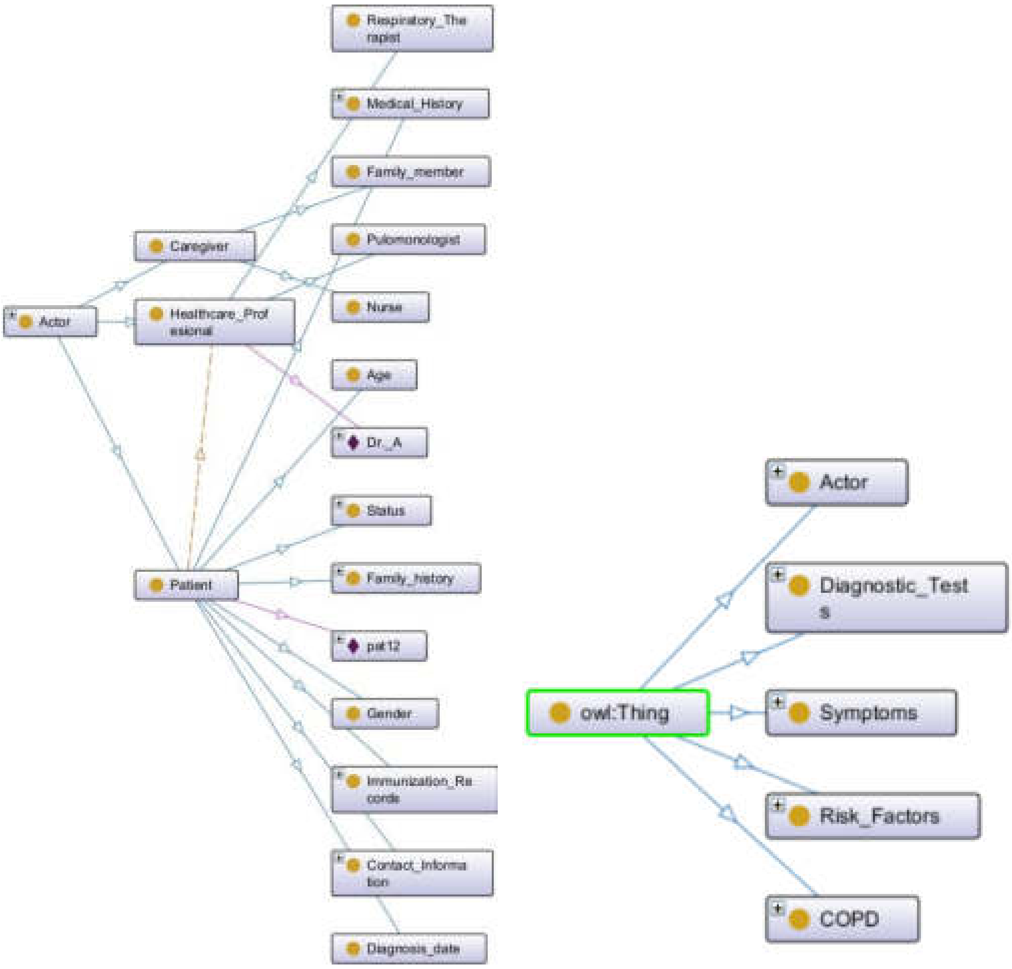


Fig. 4.6: Semantic Relationships in the COPD Ontology

tailored, real-time, and data-driven healthcare solutions. At the same time, we acknowledge the challenges associated with interoperability, data quality, and ethical concerns; a well-designed, user-centered, and ethically sound implementation can propel healthcare into a future where patient care is genuinely context-aware, leading to improved patient outcomes and overall well-being. Our research contributes significantly to the evolving landscape of COPD management, introducing an innovative ontology-based solution that aligns seamlessly with the evolving needs of modern healthcare systems. For future research, this ontology will be validated through real-life implementation. Comparison with existing techniques is to be considered in future papers. Additionally, this framework can be expanded by implementing it for other chronic diseases.

REFERENCES

[1] V.T. Anju, S. Busi, M. S. Mohan, and M. Dyavaiah, *Bacterial infections: Types and pathophysiology.* In Antibiotics-Therapeutic Spectrum and Limitations,. Academic Press, (2023), pp. 21-38

[2] H. Hutton, K., H.J. Zar, and A. C. Argent, *Clinical features and outcome of children with severe lower respiratory tract infection admitted to a pediatric intensive care unit in South Africa.*, Journal of Tropical Pediatrics,65, no. 1, (2019), pp. 46–54.

[3] N. Murgia, and A. Gambelunghe, *Occupational COPD—The most under-recognized occupational lung disease?.*, Respirology, 27, no. 6 (2022), pp. 399-410.

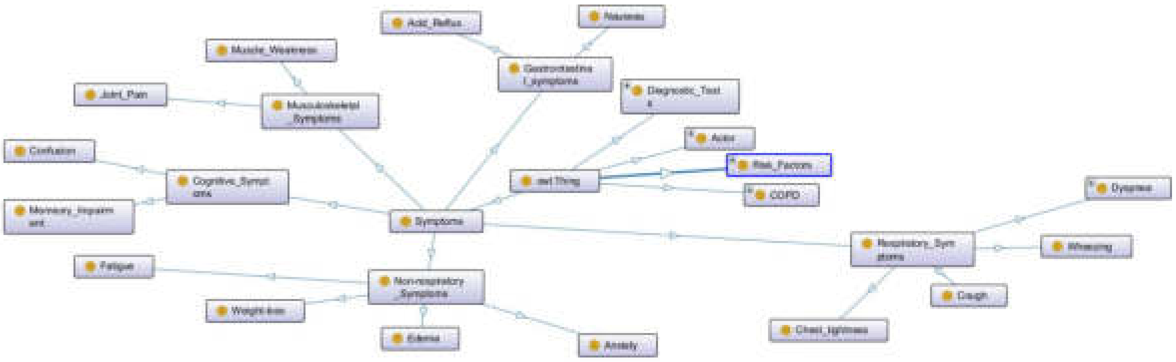


Fig. 4.7: Semantic Relationships in the COPD Ontology

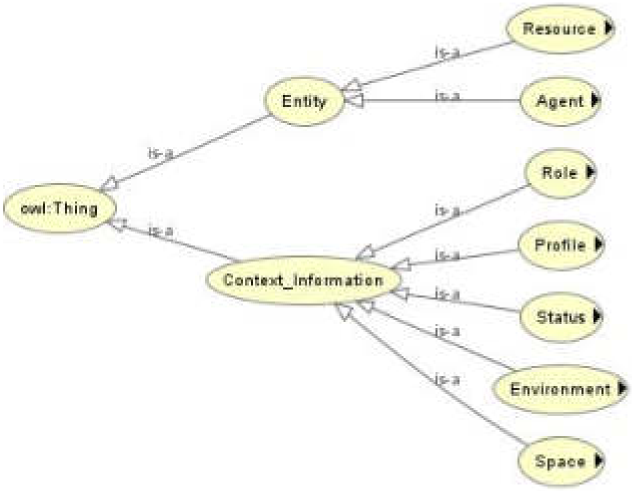


Fig. 4.8: Class Hierarchy in Context Ontology .

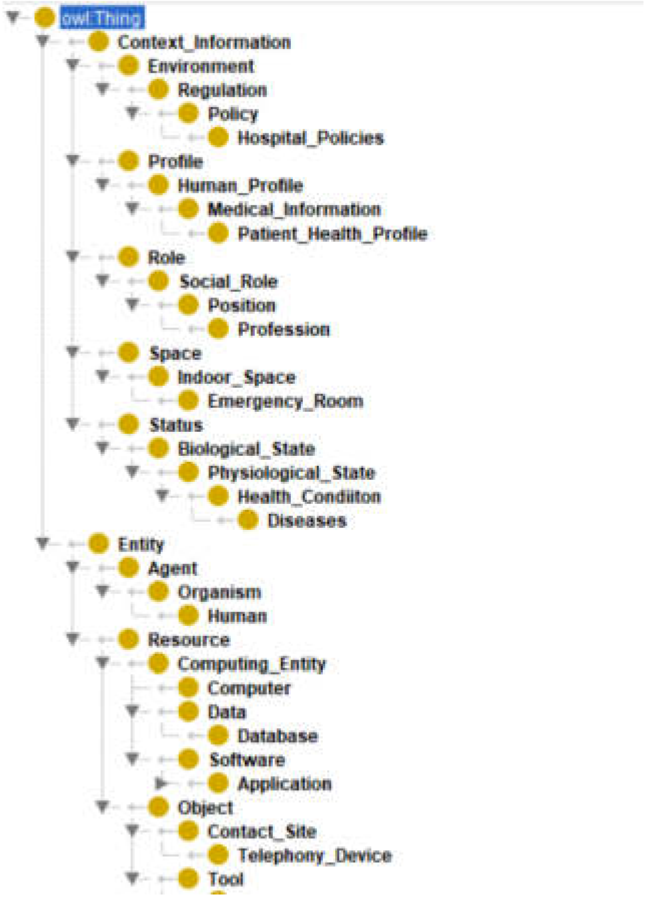

owl:Thing
Context_Information
Environment
Regulation
Policy
Hospital_Policies
Profile
Human_Profile
Medical_Information
Patient_Health_Profile
Role
Social_Role
Position
Profession
Space
Indoor_Space
Emergency_Room
Status
Biological_State
Physiological_State
Health_Condiiton
Diseases
Entity
Agent
Organism
Human
Resource
Computing_Entity
Computer
Data
Database
Software
Application
Object
Contact_Site
Telephony_Device
Tool

Fig. 4.9: Classes in Context Ontology.

[4] J. S. Alqahtani, C. M. Njoku, B. Bereznicki, B. C. Wimmer, G. M. Peterson, L. Kinsman, Y. S. Aldabayan, A.M.

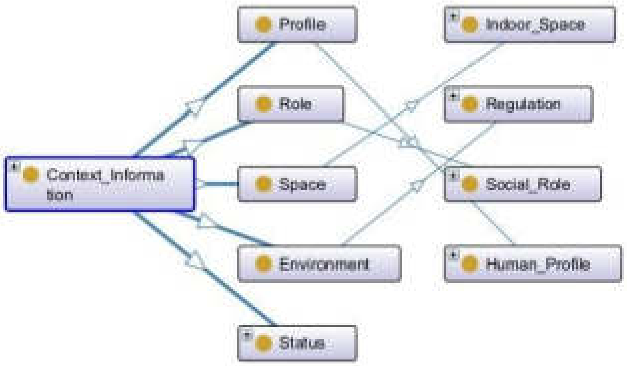


Fig. 4.10: Semantic Relationships in the Context Ontology.

Alrajeh, A.M. Aldhahir, S. Mandal, S. and J.R. Hurst, *Risk factors for all-cause hospital readmission following exacerbation of COPD: a systematic review and meta-analysis*, European Respiratory Review, 29, no. 156 (2020).

[5] A. Mollalo, B. Vahedi, S. Bhattarai, L. C. Hopkins, S. Banik, and B. Vahedi,*Predicting the hotspots of age-adjusted mortality rates of lower respiratory infection across the continental United States: Integration of GIS, spatial statistics and machine learning algorithms*, International Journal of Medical Informatics, 142 (2020): 104248.

[6] P. Hanlon, X. Guo, E. McGhee, J. Lewsey, Dav. McAllister, and F. S. Mair,*Systematic review and meta-analysis of prevalence, trajectories, and clinical outcomes for frailty in COPD*, NPJ Primary Care Respiratory Medicine 33, no. 1 (2023): 1.

[7] N. Murad, and E. Melamud, *Global patterns of prognostic biomarkers across disease space*, Scientific Reports, vol. 12, no. 1, 21893 (2022).

[8] P. Venkatesan, *GOLD COPD report: 2023 update*, The Lancet Respiratory Medicine, vol. 1, no. 1, p. 18 (2023).

[9] A. Fazleen, and T. Wilkinson, *Early COPD: current evidence for diagnosis and management*, Therapeutic advances in respiratory disease, vol. 14, p. 1753466620942128 (2020).

[10] A.U. Rehman, M.A.A. Hassali, S.A. Muhammad, S.N. Harun, S. Shah, and S. Abbas, *The economic burden of chronic obstructive pulmonary disease (COPD) in Europe: results from a systematic review of the literature*, The European Journal of Health Economics 21(2020): 181–194

[11] Q. Song, P. Chen, and X.M. Liu, *The role of cigarette smoke-induced pulmonary vascular endothelial cell apoptosis in COPD*, Respiratory research, vol. 22 (2021), pp. 1–15.

[12] A.I. Ritchie, and J.A. Wedzicha, *Definition, causes, pathogenesis, and consequences of chronic obstructive pulmonary disease exacerbations*, Clinics in chest medicine, vol. 41, no. 3 (2020), pp. 421–438.

[13] E.L. Axson, K. Ragutheeswaran, V. Sundaram, C.I. Bloom, A. Bottle, M.R. Cowie, and J.K. Quint, *Hospitalisation and mortality in patients with comorbid COPD and heart failure: a systematic review and meta-analysis*, Respiratory Research, vol. 21 (2020), pp. 1–13.

[14] I.A.Ratiu, T. Ligor, V. Bocos-Bintintan, C.A. Mayhew, and B. Buszewski, *Volatile organic compounds in exhaled breath as fingerprints of lung cancer, asthma, and COPD*, Journal of Clinical Medicine, vol. 1, no. 1 (2020), p. 32.

[15] L. Ruvuna, and A. Sood, *Epidemiology of chronic obstructive pulmonary disease*, Clinics in Chest Medicine, vol. 41, no. 3 (2020), pp. 315–327.

[16] B. A. Parris, H.E. O'Farrell, K.M. Fong, and I.A. Yang, *Chronic obstructive pulmonary disease (COPD) and lung cancer: common pathways for pathogenesis*, Journal of Thoracic Disease, vol. 11, Suppl 17 (2019), pp. S2155.

[17] A. Watson, and T.M. Wilkinson, *Digital healthcare in COPD management: a narrative review on the advantages, pitfalls, and need for further research*, Therapeutic Advances in Respiratory Disease, vol. 16 (2022), p. 17534666221075493.

[18] S. Pimenta, H. Hansen, H. Demeyer, P. Slevin, and J. Cruz, *Role of digital health in pulmonary rehabilitation and beyond: shaping the future*, ERJ Open Research, vol. 9, no. 2 (2023).

[19] F.M. Franssen, P. Alter, N. Bar, B.J. Benedikter, S. Iurato, D. Maier, M. Maxheim, F.K. Roessler, M.A. Spruit,C.F. Vogelmeier, and E.F. Wouters, *Personalized medicine for patients with COPD: where are we?*, International Journal of Chronic Obstructive Pulmonary Disease (2019), pp. 1465–1484.

[20] S. Wollenstein-Betech, C.G. Cassandras, and I.C. Paschalidis, *Personalized predictive models for symptomatic COVID-19 patients using basic preconditions: hospitalizations, mortality, and the need for an ICU or ventilator*, International Journal of Medical Informatics, vol. 142 (2020), p. 104258.

[21] T. Bonnevie, P. Smondack, M. Elkins, B. Gouel, C. Medrinal, Y. Combret, J.F. Muir, A. Cuvelier, G. Prieur, and F.E. Gravier, *Advanced telehealth technology improves home-based exercise therapy for people with stable chronic obstructive pulmonary disease: a systematic review*, Journal of Physiotherapy, vol. 67, no. 1 (2021), pp. 27–40.

[22] M. Tsutsui, F. Gerayeli, and D.D. Sin, *Pulmonary rehabilitation in a post-COVID-19 world: telerehabilitation as a new standard in patients with COPD*, International journal of chronic obstructive pulmonary disease (2021), pp. 379–391.

[23] M.T. Dransfield, G. J. Criner, D. M. Halpin, M. K. Han, B. Hartley, R. Kalhan, P. Lange, D.A. Lipson, F.J. Martinez, D. Midwinter, and D.Singh, *Time-dependent risk of cardiovascular events following an exacerbation in patients with chronic obstructive pulmonary disease: post hoc analysis from the IMPACT trial*, Journal of the American Heart Association, vol. 11, no. 18 (2022), p. e024350.

[24] Q. Zhan, J. Zhang, Y. Lin, W. Chen, X. Fan, and D. Zhang, *Pathogenesis and treatment of Sjogren's syndrome: Review and update*, Frontiers in Immunology, vol. 14 (2023), p. 1127417.

[25] H. Ding, F. Fatehi, A. Maiorana, N. Bashi, W. Hu, and I. Edwards, *Digital health for COPD care: the current state of play*, Journal of Thoracic Disease, vol. 11, Suppl 17 (2019), p. S2210.

[26] H. Kwon, S. Lee, E. J. Jung, S. Kim, J. K. Lee, D. K. Kim, T.H.Kim, S.H. Lee, M.K. Lee, S. Song, and K. Shin, *An mHealth management platform for patients with chronic obstructive pulmonary disease (efil breath): randomized controlled trial*, JMIR mHealth and uHealth, vol. 6, no. 8 (2018), p. e10502.

[27] V. Kumar, H. Alshazly, S. A. Idris, and S. Bourouis,*Evaluating the Impact of COVID-19 on Society, Environment, Economy, and Education*. Sustainability 13, no. 24: 13642 (2021).

[28] A. Agarwal, S. Sharma, V. Kumar and M. Kaur,*Effect of E-learning on public health and environment during COVID-19 lockdown*.In Big Data Mining and Analytics, vol. 4, no. 2,(2021): 104–115.

[29] H. Ajami, H. Mcheick, and K. Mustapha, *Ubiquitous healthcare systems and medical rules in COPD Domain*, In How AI Impacts Urban Living and Public Health: 17th International Conference, ICOST 2019, New York City, NY, USA, October 14-16, 2019, Proceedings, vol. 17, pp. 97–108. Springer International Publishing.

[30] H. Ajami and H. Mcheick, *Ontology-based model to support ubiquitous healthcare systems for COPD patients*, Electronics, vol. 7, no. 12 (2018), p. 371.

[31] J. Ahamed and M. A. Chishti, *Ontology-based semantic interoperability approach in the Internet of Things for healthcare domain*, Journal of Discrete Mathematical Sciences and Cryptography, vol. 24, no. 6 (2021), pp. 1727–1738.

[32] J. Ahamed, R. N. Mir, and M. A. Chishti, *RML based ontology development approach in internet of things for healthcare domain*, International Journal of Pervasive Computing and Communications, vol. 17, no. 4 (2021), pp. 377–389.

[33] K.M. Kouamé, and H. Mcheick, *An ontological approach for early detection of suspected COVID-19 among COPD patients*, Applied System Innovation, vol. 4, no. 1 (2021), p. 21.

[34] H. Ajami, H. Mcheick, and C. Laprise, *First Steps of Asthma Management with a Personalized Ontology Model*, Future Internet, vol. 14, no. 7 (2022), p. 190.

[35] E. Sezer, O. Bursa, O. Can, and M.O. Unalir, *Semantic Web Technologies for IoT-Based Health Care Information Systems*, In The Tenth International Conference on Advances in Semantic Processing, pp. 45–48 (2016).

[36] H.B. Elhadj, F. Sallabi, A. Henaien, L. Chaari, K. Shuaib, and M. Al Thawadi, *Do-Care: A dynamic ontology reasoning based healthcare monitoring system*, Future Generation Computer Systems, vol. 118 (2021), pp. 417–431.

[37] J.L.Pépin, B. Degano, R. Tamisier, and D. Viglino, *Remote monitoring for prediction and management of acute exacerbations in chronic obstructive pulmonary disease (AECOPD)*, Life, vol. 12, no. 4 (2022), p. 499.

[38] *International Classification of Functioning, Disability and Health (ICF)*, n.d.

[39] World Health Organization, *International Classification of Functioning, Disability, and Health: Children and Youth Version: ICF-CY*, World Health Organization (2007).

[40] N.A. Marotta, A. Ammendolia, C. Marinaro, A. Demeco, L. Moggio, and C. Costantino, *International classification of functioning, disability and health (ICF) and correlation between disability and finance assets in chronic stroke patients*, Acta Bio Medica: Atenei Parmensis, vol. 91, no. 3 (2020), p. e2020064.

[41] P. J., *Video 1 NA: What is the International Classification of Functioning, Disability and Health (ICF)?*, 2106, `https://www.youtube.com/watch?v=uoEIc4wBaIo`

[42] World Health Organization, *International Classification of Functioning, Disability and Health (ICF) Beginners Guide*, (2002)

[43] P. J., (2021), *Video 2 NA: What is Body Structure and Function?*, `https://youtu.be/O2pRqr-THMs`

[44] P.J., (2021), *Video 3 NA: What are Activity and Participation?*, `https://youtu.be/mwYxs567Cg0`

[45] P.J., (2021), *Video 4 NA: What are contextual factors?*, `https://youtu.be/-j0495iwCX0`

[46] World Health Organization, (2023), *ICF Core Sets*, `https://www.icf-core-sets.org/`

[47] N. S. Raj, and V. G. Renumol, *A systematic literature review on adaptive content recommenders in personalized learning environments from 2015 to 2020*, Journal of Computers in Education, vol. 9, no. 1 (2022), pp. 113–148.

[48] P. Pradeep, and S. Krishnamoorthy, *The MOM of context-aware systems: A survey*, Computer Communications, vol. 137 (2019), pp. 44–69.

[49] X. Li, C. H Chen, P. Zheng, Z. Jiang, and L. Wang, *A context-aware diversity-oriented knowledge recommendation approach for smart engineering solution design*, Knowledge-Based Systems, vol. 215 (2021), p. 106739.

[50] P. Pradeep, S. Krishnamoorthy, R. K. Pathinarupothi, and A. V. Vasilakos, *Leveraging Context-Awareness for Internet of Things Ecosystem: Representation, Organization, and Management of Context*, Computer Communications, vol. 177 (2021), pp. 33–50.

[51] H. S. Jim, A. I. Hoogland, N. C. Brownstein, A. Barata, A. P. Dicker, H. Knoop, B. D. Gonzalez, and Perkins, R., *Innovations in research and clinical care using patient-generated health data*, CA: a cancer journal for clinicians, vol. 70, no. 3 (2020), pp. 182–199.

[52] M. Menear, M. A. Blanchette, O. Demers-Payette, and D. Roy, *A Framework for Value-Creating Learning Health Systems*, Health Research Policy and Systems, vol. 17, no. 1 (2019), pp. 1–13.

[53] C. A. Low, *Harnessing Consumer Smartphone and Wearable Sensors for Clinical Cancer Research*, Digital Medicine, vol. 3, no. 1 (2020), p. 140.

[54] A. Adikari, D. D. Silva, H. Moraliyage, D. Alahakoon, J. Wong, M. Gancarz, S. Chackochan, B. Park, R. Heo,

and Y. Leung, *Empathic Conversational Agents for Real-Time Monitoring and Co-Facilitation of Patient-Centered Healthcare*, Future Generation Computer Systems, vol. 126 (2022), pp. 318–329.

[55] T. T. Khuat, D. J. Kedziora, and B. Gabrys, *The Roles and Modes of Human Interactions with Automated Machine Learning Systems*, arXiv preprint arXiv:2205.04139 (2022).

[56] Jayatilake, S. M. D. A. Chinthaka., and G. U. Ganegoda, *Involvement of Machine Learning Tools in Healthcare Decision Making*, Journal of Healthcare Engineering, vol. 2021 (2021).

[57] M. Zon, G. Ganesh, M. J. Deen, and Q. Fang, *Context-Aware Medical Systems within Healthcare Environments: A Systematic Scoping Review to Identify Subdomains and Significant Medical Contexts*, International Journal of Environmental Research and Public Health, vol. 20, no. 14 (2023), p. 6399.

[58] N. Upadhyay, A. Kamble, and A. Navare, *Virtual Healthcare in the New Normal: Indian Healthcare Consumers' Adoption of Electronic Government Telemedicine Service*, Government Information Quarterly, vol. 40, no. 2 (2023), p. 101800.

[59] E. Tsekleves, and J. Keady, *Design for People Living with Dementia: Interactions and Innovations*, Routledge (2021).

[60] J.N. Nyakina, and B. H. Taher, *A Survey of Healthcare Sector Digitization Strategies: Vulnerabilities, Countermeasures, and Opportunities*, World Journal of Advanced Engineering Technology and Sciences, vol. 8, no. 1 (2023), pp. 282-301.

[61] M. AbdulRaheem, J. B. Awotunde, C. Chakraborty, E.A. Adeniyi, I. D. Oladipo, and A. K. Bhoi, *Security and privacy concerns in smart healthcare system*, In Implementation of Smart Healthcare Systems using AI, IoT, and Blockchain, pp. 243–273. Academic Press (2023).

[62] Ferreira, R. C. Correia, *COVID-19 and Cybersecurity: Finally, an Opportunity to Disrupt?* JMIRx Med 2 (2021): e21069.

[63] A. Sardi, A. Rizzi, E. Sorano and A. Guerrieri, *Cyber Risk in Health Facilities: A Systematic Literature Review.* Sustainability 12, no. 17 (2020): 7002.

[64] M. Maksimović, and V. Vujović, *Internet of Things Based E-health Systems: Ideas, Expectations and Concerns.* In Handbook of Large-Scale Distributed Computing in Smart Healthcare, 241–280. 2017.

[65] S. Ahmed, and A. Rajput, *Threats to Patients' Privacy in Smart Healthcare Environment.* In Innovation in Health Informatics, 375–393. (2020).

[66] H.K. Channi, and C.L. Chowdhary, *Blockchain-Based IoT E-Healthcare.* In Handbook of Research on Solving Societal Challenges Through Sustainability-Oriented Innovation, IGI Global, (2023): 56–73.

[67] P.E. Idoga, M. Agoyi, E. Y. Coker-Farrell, and O. L. Ekeoma, *Review of security issues in e-Healthcare and solutions.* In 2016 HONET-ICT, pp. 118–121. IEEE, 2016.

[68] M. Papaioannou, M. Karageorgou, G. Mantas, V. Sucasas, I. Essop, J. Rodriguez, and D. Lymberopoulos, *A survey on security threats and countermeasures in internet of medical things (IoMT).* Transactions on Emerging Telecommunications Technologies 33, no. 6 (2022): e4049.

[69] Tagliaferri, Luca, A. Budrukkar, J. Lenkowicz, M. Cambeiro, F. Bussu, J. L. Guinot, G. Hildebrandt, B. Johansson,J.E. Meyer,P. Niehoff, and A. Rovirosa, *ENT COBRA ONTOLOGY: the covariates classification system proposed by the Head and Neck and Skin GEC-ESTRO Working Group for interdisciplinary standardized data collection in head and neck patient cohorts treated with interventional radiotherapy (brachytherapy).* Journal of contemporary brachytherapy 10, no. 3 (2018): 260–266.

[70] G. Padinjappurathu, Shynu, C. L. Chowdhary, C. Iwendi, M. A. Farid, and L. K. Ramasamy, *An Efficient and Privacy-Preserving Scheme for Disease Prediction in Modern Healthcare Systems.* Sensors 22, no. 15: 5574, (2022)

[71] S. Liu, and L. Cheng, *A Context-Aware Reflective Middleware Framework for Distributed Real-Time and Embedded Systems.* Journal of Systems and Software 84, no. 2 (2011): 205–218.

[72] R. Ali, F. Dalpiaz, and P. Giorgini, *A Goal-Based Framework for Contextual Requirements Modeling and Analysis.* Requirements Engineering 15, no. 4 (2010).

[73] P. Brézillon, and J. C. Pomerol, *Contextual Knowledge Sharing and Cooperation in Intelligent Assistant Systems.* Le Travail Humain 62, no. 3 (1999): 223–246.

[74] P. Couderc, and A. M. Kermarrec, *Enabling Context-Awareness from Network-Level Location Tracking.* In International Symposium on Handheld and Ubiquitous Computing, 67–73. Springer, 1999.

[75] V. Arnaboldi, M. Conti, and F. Delmastro, *CAMEO: A Novel Context-Aware Middleware for Opportunistic Mobile Social Networks.* Pervasive and Mobile Computing 11 (2014): 97–108.

[76] H. Chen, T. Finin, and A. Joshi, *An Ontology for Context-Aware Pervasive Computing Environments.* Knowledge Engineering Review 18, no. 3 (2003): 197–207.

[77] F.A. Norki, R. Mohamad, and N. Ibrahim, *Context ontology in mobile applications.* Journal of Information and Communication Technology, 19(1), (2020): 21–44.

[78] B. Schilit, N. Adams, and R. Want, *Context-Aware Computing Applications.* In 1994 First Workshop on Mobile Computing Systems and Applications, 85–90. IEEE, 1994.

[79] International Data Corporation (IDC) Corporate USA, *Worldwide Smart Connected Device Shipments.* `http://www.idc.com/getdoc.jsp?containerId=prUS23398412`. Accessed on: 2012-08-01.

[80] M. Popova, L. Globa, and R. Novogrudska,*Multilevel ontologies for big data analysis and processing.*(2021)

[81] S. Rizou, K. Häussermann, F. Dürr, N. Cipriani, and K. Rothermel, *A System for Distributed Context Reasoning.* In 2010 Sixth International Conference on Autonomic and Autonomous Systems, 84–89. IEEE, March 2010.

[82] E. J. Y. Wei, and A. T. S. Chan, *Campus: A Middleware for Automated Context-Aware Adaptation Decision Making at Run Time.* Pervasive and Mobile Computing 9, no. 1 (2013): 35–56.

[83] T. M. Chiu, and B. P. Ku, *Moderating Effects of Voluntariness on the Actual Use of Electronic Health Records for Allied Health Professionals.* JMIR Medical Informatics 3, no. 1 (2015): e2548.

[84] C. Diamantini, A. Nocera, D. Potena, E. Storti, and D. Ursino, *Multi-Dimensional Contexts for Querying IoT Networks.* In SEBD, (2019)

[85] L. Zhong-Jun, L. Guan-Yu,and P. Ying, *A method of meta-context ontology modeling and uncertainty reasoning in swot.* In 2016 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC),(2016): 128–135.

[86] J. Gantz, *The Embedded Internet: Methodology and Findings.* In IDC. 2009.

[87] C. Diamantini, C., Nocera, A., Potena, D., Storti, E. and Ursino, D., 2019. Multi-Dimensional Contexts for Querying IoT Networks. In SEBD.

[88] S . Ali, S. Khusro, I. Ullah, A. Khan, and I. Khan, *Smartontosensor: Ontology for Semantic Interpretation of Smartphone Sensors Data for Context-Aware Applications.* Journal of Sensors 2017.

[89] M. Hoda, V. Montaghami, H. Al Osman, and A. El Saddik, *ECOPPA: Extensible Context Ontology for Persuasive Physical-Activity Applications.* In Proceedings of the International Conference on Information Technology & Systems (ICITS 2018), 309–318. Springer International Publishing, 2018.

[90] M. Hoda, M, V. Montaghami, H. Al Osman, and A. El Saddik, *ECOPPA: Extensible Context ontology for persuasive physical-activity applications.* In Proceedings of the International Conference on Information Technology and Systems ,Springer International Publishing. (2018): 309–318

[91] C. Angsuchotmetee, R. Chbeir, and Y. Cardinale, *MSSN-Onto: An ontology-based approach for flexible event processing in Multimedia Sensor Networks.* Future Generation Computer Systems, 108, (2020): 1140–1158.

[92] N. Gupta, N. Malhotra, and P. Ish, *GOLD 2021 Guidelines for COPD—What's New and Why.* Advances in Respiratory Medicine 89, no. 3 (2021): 344–346.

[93] T. L. Croxton, G. G. Weinmann, R. M. Senior, and J. R. Hoidal, *Future Research Directions in Chronic Obstructive Pulmonary Disease.* American Journal of Respiratory and Critical Care Medicine 165, no. 6 (2002): 838–844.

[94] H. A. H. Albitar, and V. N. Iyer, *Adherence to Global Initiative for Chronic Obstructive Lung Disease Guidelines in the Real World: Current Understanding, Barriers, and Solutions.* Current Opinion in Pulmonary Medicine 26, no. 2 (2020): 149–154.

[95] A. R. Patel, A. R. Patel, S. Singh, S. Singh, and I. Khawaja, *Global initiative for chronic obstructive lung disease: the changes made.* Cureus 11, no. 6 (2019).

[96] J. Peng, C. Chen, M. Zhou, X. Xie, , Y. Zhou, and C. H. Luo, *A machine-learning approach to forecast aggravation risk in patients with acute exacerbation of chronic obstructive pulmonary disease with clinical indicators*, Scientific reports,10(1), (2020): 3118.

[97] COPD risk factors AI table. https://www.kaggle.com/code/mlconsult/copd-risk-factors-ai-table

[98] The COPD Dataset. https://www.kaggle.com/code/mlconsult/summary-page-covid-19-risk-factors

[99] COPD Patients Dataset. https://www.kaggle.com/datasets/prakharrathi25/copd-student-dataset

[100] Context-Aware Recommender. https://www.kaggle.com/code/amiralisa/context-aware-recommender

Edited by: Chiranji Lal Chowdhary
Special issue on: Scalable Machine Learning for Health Care: Innovations and Applications
Received: Oct 31, 2023
Accepted: Mar 1, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1442–1454, DOI 10.12694/scpe.v25i3.2671

RESEARCH ON INTEGRATING BLOCKCHAIN AND MACHINE LEARNING LPP ALGORITHM IN ONLINE EDUCATION PLATFORM UNDER COVID-19 ENVIRONMENT

DAOJUN WANG, MEISHU WANG, AND XINLI XING*

Abstract. Learners confront the issue of navigating an enormous quantity of resources in the developing field of online education, which has been exacerbated by the COVID-19 pandemic. To solve this, our research proposes a novel Learner Path Planning (LPP) model that integrates blockchain and machine learning technologies to maximize the online learning experience. This model employs the ant colony optimization technique, which has been upgraded with blockchain for enhanced security and machine learning for intelligent path planning, to provide a more personalized and efficient learning experience. Our approach determines the extent of concept realization and interaction by examining the interaction degrees of knowledge points, establishing heuristic information and initial pheromone levels for the optimization process. This technique not only optimizes teaching duration based on instructional efficacy, but it also adapts dynamically to individual learner needs. Our empirical data reveal that goal success rates improve significantly across all learner levels. For example, elementary students in 2021 had the highest goal achievement rate of 0.5896. In 2019, intermediate and advanced learners attained rates of 0.7726 and 0.9058, with a significant association between course similarity and target achievement. Blockchain integration ensures secure and transparent processing of educational data, while machine learning algorithms successfully personalize learning routes to meet the various demands of learners. This study not only assists learners in effectively identifying suitable resources, but it also provides useful insights for instructors in improving online teaching approaches. The model's adaptability and scalability make it particularly applicable in the context of the COVID-19 pandemic's rapid developments and problems in the education sector.

Key words: COVID-19; Online education platform; Ant colony optimization algorithm; Learner behavior characteristics; Path planning

1. Introduction.

Corona Virus Disease 2019 (COVID-19), as an acute respiratory infectious disease, is a threat to the economic development and life safety of people all over the world. In the context of the rampant COVID-19, many industries such as education, retail, engineering and finance have been seriously impacted. This has led to many offline work not being carried out normally. With the wide application of information technology, online education platform came into being in this context, and has been vigorously developed. It not only enabled students to harvest massive teaching resources, but also fundamentally changed the way learners acquire knowledge. It will no longer be constrained by objective conditions such as place, time and space [1, 2]. In recent years, new educational forms such as Muke have changed the original teaching methods. Learning management system and other high-quality learning platforms provided new possibilities for learners [3, 4]. The learning behavior data in the online learning platform can be used to analyze the learning effect through big data technology, thus promoting the teaching process [5, 6].

In order to enhance learners' efficiency and effect, many education experts have discussed learning path planning (LPP) algorithms. But learners' learning abilities vary greatly. Therefore, finding a satisfactory learning path is a major and challenging task. At present, LPP algorithms can be divided into two types, namely, planning the learning path of a course and planning the learning path between courses. Ant colony optimization algorithm has strong adaptability and robustness. It shows good performance in dealing with many problems. A LPP algorithm based on multidimensional time series data analysis is proposed. It is expected to promote the in-depth integration of education and teaching and modern information technology, and provide technical support for the wide application of intelligent education.

The rapid expansion of online education, spurred by the COVID-19 pandemic, has given the educational sector with new prospects as well as obstacles. One of the most significant issues for learners utilizing online platforms is resource overload. While abundant learning tools are valuable, they might overwhelm learners,

*Department of Physical Education, Qingdao Agricultural University, Qingdao 266109, China (Corresponding author, Xinli_Xing2023@outlook.com)

impeding effective learning and route planning. This circumstance involves a novel strategy to streamlining the online learning process, guaranteeing that learners may efficiently access and use resources adapted to their specific needs and learning objectives. In this context, the incorporation of blockchain and machine learning technology into Learner Path Planning (LPP) algorithms is a game changer. Blockchain technology, which is well-known for its security, transparency, and decentralized nature, provides a solid framework for managing educational data. At the same time, machine learning delivers cognitive analytical skills, which are critical for personalizing the learning experience based on individual learner profiles and behaviors.

The study makes a significant contribution to the field of online education technology in various ways:

1. We created a model that blends blockchain and machine learning with the LPP algorithm. This connection provides secure and effective resource management while also personalizing the learning route for each learner.
2. Our model effectively navigates the enormous array of online resources by incorporating ant colony optimization techniques into the LPP algorithm. This strategy effectively directs learners to the most relevant and valuable content, enhancing their learning experience.

2. Related works . COVID-19 has had a huge impact on the education industry. Online education in colleges and universities has become a hot research topic for educators. Learner behavior data is of great significance for analyzing learning habits, learning status and cognitive level, and can also improve teaching quality. Su G et al. found that there is great variability between learners' learning behavior and test results by visualizing learners' learning behavior. This is helpful for teachers to monitor the course progress and learners' learning performance, and timely adjust teaching strategies according to the actual situation [7]. Zhao Y and other researchers proposed a learning habit determination and LPP method based on learning behavior analysis. This method planed and recommended the path of learning content according to learners' learning habits [8]. Liu Y, et al., studied the characteristics of learners' learning activities and learning habits, and compared the importance of these characteristics through experiments. The results showed that the characteristics related to learning habits play a more important role in predicting students' performance [9]. Li J proposed a new adaptive network learning model based on big data. First of all, the model used genetic algorithms to evaluate learners' relevant future education goals. Then, the adaptive personalized learning path was generated by combining the ant colony optimization algorithm. Finally, social network analysis was used to determine learners' motivation to assign learning rhythm to each learner [9].

Wang J et al. built an improved adaptive tutoring system model using ant colony optimization algorithm, which can find the best learning path according to the learning mode and performance of the improved adaptive tutoring system [11]. Zhang J scholars and researchers classified learners' learning styles. Ant colony optimization algorithm was used to help learners find adaptive learning objects to obtain the best learning path [21]. Cui Z proposed a learning path optimization method based on evolutionary algorithm. The context of each knowledge point, the learning interests of learners and the fields involved in learning resources were comprehensively considered to extract the relationship between knowledge. Based on the evolutionary algorithm, the objective function was optimized, and the learning path to meet the learning needs of learners was finally constructed [22]. Zhou X et al. proposed an adaptive learning model based on ant colony algorithm. This can meet learners' different preferences and knowledge levels, and help improve learners' academic performance and learning efficiency [14]. Yang Y et al. proposed the research method of online learning resource serialization by analyzing the characteristics of learning resource serialization, and modeling it at different stages. And based on the learning needs of learners, combined with particle swarm optimization algorithm, the intelligent seriation service system of learning resources was built [15].

From the relevant research status of online education platform and LPP algorithms, there are three kinds of LPP algorithms that are common and widely used at present. They are data mining, association rules and other algorithms, LPP algorithm based on graph theory, intelligent bionic algorithm, etc. However, the existing LPP algorithms are difficult to achieve good learning results. This is embodied in the interaction between learners, the role of interaction between teachers and learners in LPP, the failure to consider the review of knowledge points in the learning, and the failure to reflect the role of learning habits in LPP and effect. The research used multidimensional time series data analysis method to construct LPP algorithm, with a view to making corresponding contributions to the improvement of teaching effect in online education platform.

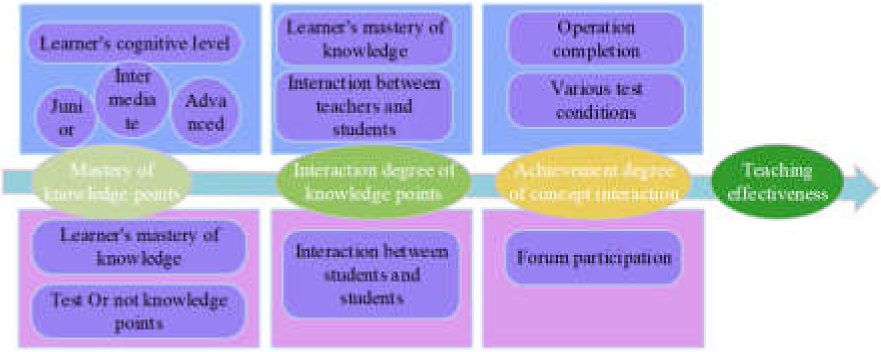

Learner's cognitive level
Juni or
Inter media te
Adva nced
Learner's mastery of knowledge
Interaction between teachers and students
Operation completion
Various test conditions
Mastery of knowledge points
Interaction degree of knowledge points
Achievement degree of concept interaction
Teaching effectiveness
Learner's mastery of knowledge
Test Or not knowledge points
Interaction between students and students
Forum participation

Fig. 3.1: Relevance Between Related Definitions

3. Construction of LPP Algorithm in Online Education Platform.

3.1. Definition of LPP Algorithm and Improvement of Ant Colony Optimization Algorithm. LPP algorithm plays a key role in improving teachers' teaching quality and students' learning effect. In order to obtain the optimal learning path of learners and in a relatively short time, a LPP algorithm based on ant colony optimization and multidimensional time series data is proposed. The algorithm defines the degree of realization of concept interaction by the degree of knowledge point interaction. This sets the heuristic information and initial pheromone of the ant colony optimization algorithm according to the degree of concept interaction and learning path. At the same time, it also optimizes the teaching duration according to the teaching effect to help learners obtain good learning effect [16-18]. To better describe the algorithm-related models, the research first describes the algorithm-related definitions. In view of the different cognitive levels of learners, they are divided into three levels: primary, intermediate and advanced. The definitions related to the algorithm include the mastery of knowledge points, the interaction degree of knowledge, the realization of concept interaction, and the teaching effect. Figure 3.1 refers to the relevance between related definitions.

After learners get corresponding scores through online test items, researchers can obtain the mastery of knowledge according to the test results. However, previous studies only considered untested and tested knowledge points. On this basis, the study considers the learners' mastery of knowledge points, and the detailed steps are as follows. The matrix S can be regarded as the test scores of m learners on the k test questions, that is, formula 3.1.

$$S = \begin{bmatrix} s_{11} & \cdots & s_{1m} \\ \vdots & \ddots & \vdots \\ s_{k1} & \cdots & s_{km} \end{bmatrix} \tag{3.1}$$

The matrix F is the correlation between$k$test questions and n knowledge points, which can then determine whether students' mastery of knowledge points can be tested in the test questions, that is, formula 3.2.

$$F = \begin{bmatrix} f_{11} & \cdots & f_{1k} \\ \vdots & \ddots & \vdots \\ f_{n1} & \cdots & f_{nk} \end{bmatrix} \tag{3.2}$$

In formula 3.2, the value of F in the research process is 0, 1, 2, and 3, and the degree of association is no association, partial association, indirect association, and direct association. After obtaining matrix S and matrix F , we can get the learner u's mastery of knowledge point i , which can be referred to by formula 3.3.

$$mkp_{u,i} = \frac{\sum_{j=1}^{J} f_{ij} \cdot s_{uj}}{\sum_{j=1}^{J} f_{ij} \cdot \mathrm{score}_j} \tag{3.3}$$

J is the number of exercises of test knowledge point$i$in formula 3.3. The correlation value of test question j and knowledge point i is f_{ij}. The test score of test question j is S_{uj} , and the test score of test question j is $score_j$. After the collaborative analysis of the interaction behavior data and learning behavior of learners at different levels, the research can obtain the interaction degree of knowledge points according to the interaction degree between students, the system interaction degree of learners, and the interaction of educators and students. Formula 3.4 refers to the interaction degree of learners to knowledge points $ikp_{u,j}$.

$$ikp_{u,j} = \alpha_1 \cdot SC_{u,i} + \beta_1 \cdot SS_{u,i} + \gamma_1 \cdot SS_{u,i} \tag{3.4}$$

In formula 3.4, the degree of systematic interaction between learner u and knowledge point i. The degree of interaction between learners, and the degree of interaction between teachers and students are denoted by $SC_{u,i}$, $SS_{u,i}$ and $SC_{u,i}$ respectively. According to the relevant introduction results of the references, the weight coefficient $(\alpha_1, \beta_1, \gamma_1) = (0.36, 0.33, 0.31)$ is shown below.

$$\begin{cases} SC_{u,i} & = \alpha_2 \cdot f_{SC_{u,j}} + \beta_2 \cdot t_{SC_{u,j}} + \gamma_2 \cdot p_{SC_{u,j}} \\ SS_{u,i} & = \frac{\sum_{v=1}^{m-1} w_{SS_{uv,i}} \cdot f_{SS_{uv,i}}}{m-1}, \quad u \neq v \\ SS_{u,i} & = \frac{w_{ST_{u,j}} \cdot f_{ST_{u,j}} + score_{u,j}}{\sum_{u=1}^{m} score_{u,j}} \end{cases} \tag{3.5}$$

In formula 3.5, the number and duration of the learner u to the knowledge point i are $f_{(SC_{(u,j)})}$ and $t_{(SC_{(u,j)})}$ respectively, and the number of times the learner u pauses and drags the progress bar to the knowledge point i is $P_{(SC_{(u,j)})}$. Research results of references, $(\alpha_2, \beta_2, \gamma_2) = (1, 5, 4)$. The interaction weight coefficient and interaction times of learner u and v to knowledge point i are and respectively, the weight coefficient and interaction times of learner to knowledge point are $f_{(ST_{(u,j)})}$ and $f_{(ST_{(u,j)})}$ respectively, and the homework test score of learner u to knowledge point i is $score_{u,j}$. Conceptual interaction attainment refers to the interaction between new and old concepts in learners' minds, which is difficult to obtain directly. According to the actual situation of online education development, conceptual interaction can be indirectly expressed through forum participation and other ways [19-20]. In order to accurately define the learners' understanding of knowledge points, the is defined by combining the degree of interaction between knowledge points and the learners' mastery of knowledge points, as shown in formula 3.6.

$$c_{kp_{(u,i)}} = \frac{mkp_{(u,i)}}{ikp_{(u,i)}} \tag{3.6}$$

In formula 3.6, the learner u's mastery of knowledge point i is $mkp_{(u,i)}$, and the learner u's interaction with knowledge point$i$is $ikp_{(u,i)}$. The learning effect of learners is directly related to the duration of video knowledge points and the degree of concept interaction. The learning effect of learners'u on the knowledge point i is referred to by formula 3.7.

$$TE_{(u,i)} = \frac{ckp_{(u,i)} \cdot T_i}{T} \tag{3.7}$$

In formula 3.7, the duration of video knowledge points i is T, and the total duration of n video knowledge points of the course is T . The teacher sets the teaching duration of learners at different levels according to the course objectives, and obtains the teaching effect of all learners in the U_c level on the knowledge points according to the learning effect, that is, formula 3.8.

$$TE_i = \frac{1}{m_c} \sum_{u=1}^{m_c} TE_{(u,i)} \tag{3.8}$$

In formula 3.8, the number of U_c learners is m_c.

As an intelligent bionic algorithm, ant colony optimization algorithm has many advantages, such as heuristic search, strong robustness, positive information feedback, self-organization, distributed computing, and so on. It is often used to find the best path. Figure 3.2 shows the principle of ant colony optimization algorithm.

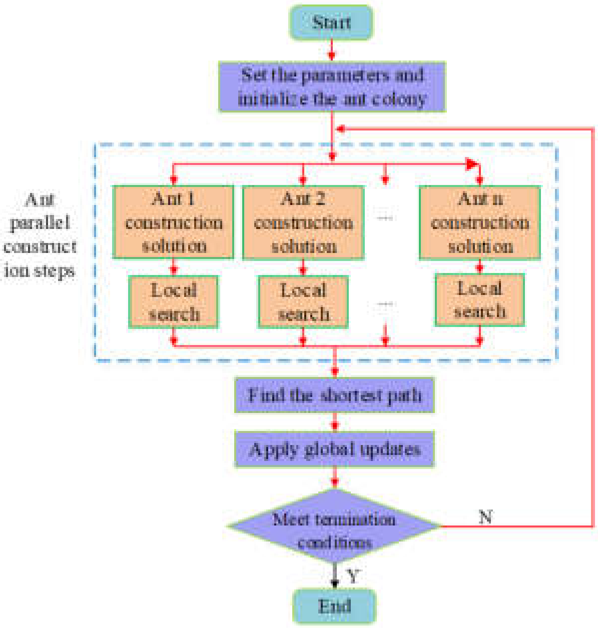

Start
Set the parameters and initialize the ant colony
Ant parallel construct ion steps
Ant 1 construction solution
Ant 2 construction solution
Ant n construction solution
Local search
Local search
Local search
Find the shortest path
Apply global updates
Meet termination conditions
N
Y
End

Fig. 3.2: The Principle of Ant Colony Optimization Algorithm.

Ant colony optimization algorithm actually simulates the ability of ants to find the shortest feeding path through information exchange. Ants secrete pheromones during foraging, which completes information exchange between ant groups. The calculation expression of pheromone increment is very important for the update of pheromone. There are three commonly used models for pheromone increment, namely, ant perisystem model, ant quantity system model, and ant density system model [21, 22, 23]. The first two pheromone increment models use local information to update pheromones after a node transfer. The latter pheromone increment model uses global information to update the pheromones passing through the path after each iteration. In order to further improve the performance of ant colony optimization algorithm, scholars in relevant fields have improved it. The ant colony optimization algorithm is improved by using optimization sequencing and elite strategy. Combining the advantages and disadvantages of the two ant colony optimization algorithms, the strategy to improve the ant colony optimization algorithm is optimized sorting. The algorithm sorts ants according to the path length to obtain the ranking order of each ant. The pheromone weighted update method is the ranking order of ants. The shorter the path length is, the higher the ranking order of the ants is, and the larger the weight value is. The pheromone update is required for the ants in front. The calculation formula is as follows 3.9.

$$\tau_{ij}(t+1) = (1-\rho)\tau_{ij} + \sum_{k=2}^{w} \Delta\tau_{ij}^{k}(t) + \Delta\tau_{ij}^{*}(t) \tag{3.9}$$

In formula (9), the initial pheromone volatilization factor is ρ, and its value range is (0,1). The ants from the 2nd to the w place updated the pheromone as $\sum_{k=2}^{w} \Delta\tau_{ij}^{k}(t)$, and the best ants updated the pheromone as $\Delta\tau_{ij}^{*}(t)$.

3.2. Learning Path Optimization Algorithm Using Multidimensional Time Series Data and Ant Colony Optimization Algorithm. From the definition of learning path optimization algorithm, the construction of learning path optimization algorithm is described. Because the carrier of knowledge points is course videos, the essence of learning path is to sort the learning order of learners' course videos according to their behavior characteristics. Figure 3.2 is the flowchart of the algorithm. First of all, the interaction degree

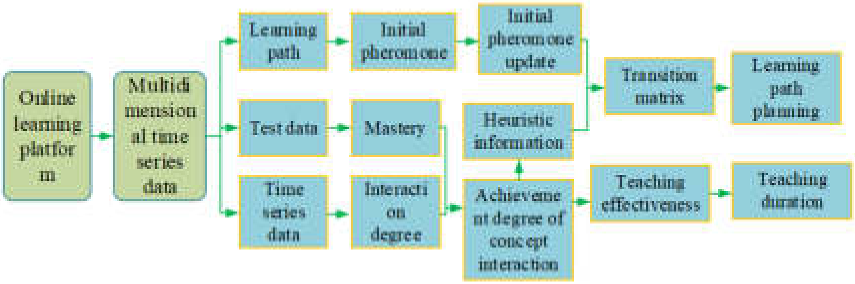


Fig. 3.3: Flow Chart of LPP Optimization Algorithm Based on Multidimensional Time Series Data and Ant Colony Optimization Algorithm

and mastery degree of knowledge points are obtained by analyzing the multi-dimensional time series data of learners in the online learning platform. According to the relevant calculation formula, the teaching effect and concept interaction attainment of learners at different levels are calculated. Then, the learning path of the prior learner is represented by the directed weight graph. The initial pheromone τ_{ij} is obtained by using ant colony optimization algorithm. The heuristic information η_{ij} is obtained according to the degree of realization of concept interaction. The initial information is updated by the learner's score ranking, the time to complete the knowledge points, and the length of the learning path. Heuristic information and pheromone are used to obtain the learner's transfer matrix P . According to the learner's transfer matrix P , it is need to select the next knowledge point to learn after learning the current knowledge point by comparing the transfer value. Finally, the learning paths planned by learners of different levels are obtained. At the same time, the teaching duration is optimized according to the effect that learners of different levels want to achieve.

The teacher cannot assure that the curriculum objectives are effectively achieved by presenting all students with a uniform video knowledge point learning length and learning path. As a result, the online education platform must incorporate earlier pupils' cognitive levels. This allows students to learn in the course based on the course objectives, learning needs, and personal advantages. The learning path of a priori learner can be expressed by formula 3.10.

$$W = [w_{ij}]_{n\times n} \tag{3.10}$$

In formula 3.10, the number of times the learner has learned i and j according to the sequence is w_{ij}. The initial pheromone between two knowledge points is related to the number of learning paths of learners. If the number of learning paths is more, the higher the initial pheromone left on the learning path can be considered, and the calculation is formula 3.11.

$$\tau_{ij} = w_{ij} \tag{3.11}$$

The heuristic information η_{ij} can reflect the heuristic preference of transferring from the current knowledge point i to the next knowledge point j. The heuristic information set in the study is related to the learning situation of learners, and the calculation is formula 3.12.

$$\eta'_{ij} = ckp_{ij} \tag{3.12}$$

In formula 3.12, the higher the value of η_{ij}, the better the effect is that learners continue to learn knowledge point j after completing knowledge point i. In the path planning learning process, the initial pheromone update depends on the appropriate parameter selection [24, 25]. The initial pheromone of learners mainly depends on the following three factors, namely, the score ranking of learners, the time spent in completing the learning of knowledge points, and the length of learning path. The relevance of knowledge points and compactness in the

Table 3.1: 2018-2021 Settings of Three Levels and Three Levels of Parameters

Grade	2019			2020			2021		
	p		k	p		k	p		k
Primary	0.15	1.60	0.65	0.45	2.50	1.70	0.05	0.75	0.40
Intermediate	0.15	1.45	3.00	0.25	1.50	1.70	0.20	0.95	2.80
Senior	0.15	0.50	2.60	0.30	0.65	3.00	0.25	1.15	1.50

teaching process should be considered. The calculation of $\Delta\tau^*_{ij}(t)$ is formula 3.13.

$$\Delta\tau^*_{ij}(t) = \begin{cases} \alpha_3 \cdot (m_c - R) \cdot w_{ij} \cdot TE_{(u,j)} + \beta_3 \cdot \frac{t_j}{T_j} + \gamma_3 \cdot \frac{l}{L} & \text{if } u \in (i,j) \\ 0 & \text{otherwise} \end{cases} \tag{3.13}$$

In formula 3.13, the score of the learner u is$R$, the individual learning effect and duration of the learner on the knowledge point are TE_{uj} and t_j , respectively. The video duration of the knowledge point is . The total path length of the learner is , and the original video path length is . According to the research results of many scholars, the value of $(\alpha_3, \beta_3, \gamma_3)$ is (0.34,0.33, 0.33). The transfer matrix can complete LPP for learners of different levels, which can be used in $P = (P_{{ij}_{n\times n}})$. According to the learning situation of learners, the transfer value of learning knowledge point after learning knowledge point i is equation 3.14.

$$p_{ij} = \begin{cases} [\tau'_{ij}]^\lambda \cdot [\eta'_{ij}]^k & \text{if } i \in (i,j) \\ 0 & \text{otherwise} \end{cases} \tag{3.14}$$

In formula 3.14, the heuristic information and pheromone of path(i,j) are referred to by $\tau_(ij)$ and, and the parameters λ and k refer to the influence of these two factors on the transfer value respectively. The optimization of teaching duration needs to be combined with the teaching effect and cognitive level of learners. The calculation formula is formula 3.15.

$$T'_i = \left(\frac{1}{TE_i}\right) \Bigg/ \left(\sum_{i=1}^{n} \frac{1}{TE_i}\right) \cdot T \tag{3.15}$$

For the optimization of teaching duration and learning path, the transfer matrix of each learner is calculated based on the multi-dimensional time series data and the learners' directed weight diagram. Then select the maximum transfer value of the knowledge point by comparing the transfer value of the unified level learners, and judge whether the selection of the next knowledge point is abnormal. The learning times of current knowledge points is judged by the number of knowledge points. Based on this, it needs to select the next knowledge point and observe the learners' transfer matrix to obtain the optimal learning path of learners' cognitive level. Finally, the teaching duration is optimized according to the teaching effect of learners. The data selected for the study is the multi-dimensional time series data generated by the Shanghai School during the learning process of 2018-2022 level-4 learners. The indicator of learners' learning is the degree of achievement of curriculum objectives. First of all, it needs to compare the degree of achievement of the overall objectives and sub-objectives of the courses for learners at different levels from 2019 to 2021. It selects the best grade from the same grade according to the engineering certification results. The optimal learning path is obtained according to the given algorithm. Then, it analyzes the learners' multidimensional time series data to optimize the duration of video knowledge points. The length of video knowledge points and the best learning path are recommended to the 2022 beginner, intermediate and advanced learners. Finally, the effectiveness of the algorithm is verified according to the learning situation of 2022 learners [26, 27, 28]. Table 3.1 refers to the settings of three levels and three levels of parameters in 2018-2021.

4. Performance and Effect Analysis of Learning Path Optimization Algorithm. There are three forms of assessment for the course "Logical Structure and Algorithm" selected in the study. It has a different

Table 4.1: Assigned Value of the Assessment Method Corresponding to the Course Sub-Objectives

Course sub-objectives	Assessment method 1 (70%)	Assessment method 2 (20%)	Assessment method 3 (10%)	Target score	Equivalent score
1	20	20	20	60	20
2	20	20	20	60	20
3	30	30	30	90	30
4	30	30	30	90	30

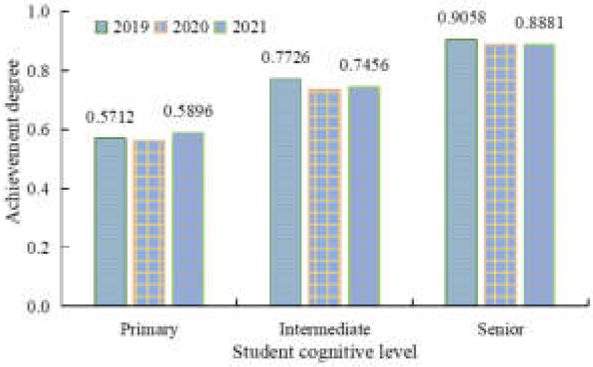


Fig. 4.1: Achievement of the Overall Objectives of Different Grades of Courses

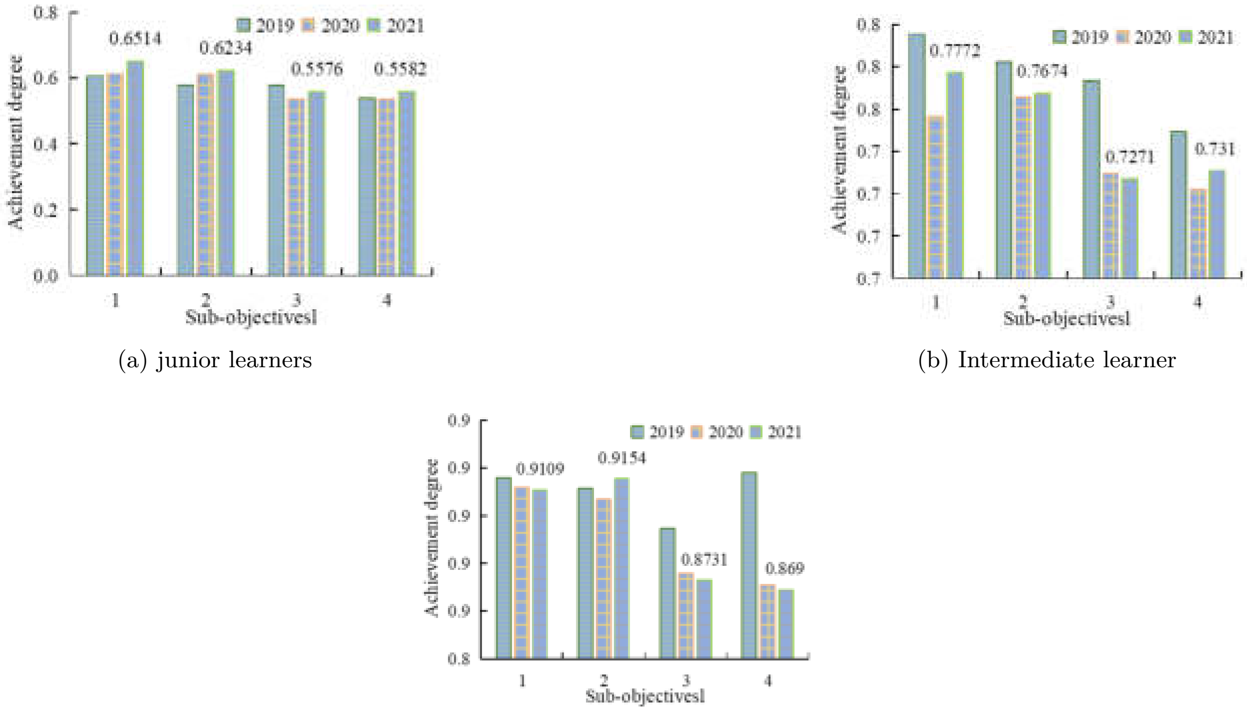
proportion in the achievement of the overall goal of the course, and is used as the basis to evaluate whether the learners have passed the assessment. The course has four objectives. They are to examine students' ability to apply basic theories and methods, judge students' ability to accurately describe the process of dealing with complex engineering problems, and evaluate students' comprehensive technical requirements. The ability to design modules and solutions to meet specific needs, and the ability to analyze students' ability to summarize and correlate complex engineering problems. Table 4.1 refers to the distribution value of the assessment method corresponding to the course objectives.

Figure 4.1 shows the achievement of the overall objectives of different grades of courses. For the same grade, the higher the cognitive level of learners, the higher the degree of achievement of the overall goal of the curriculum. For primary learners, the overall goal achievement rate of 2021 learners is the highest, with a value of 0.5896. For intermediate and advanced learners, the overall goal achievement of 2019 learners is the highest, with values of 0.7726 and 0.9058 respectively.

Figure 4.2(a) - (c) refers to the degree of achievement of the objectives of the primary, intermediate and advanced courses. For junior learners, except that the degree of achievement of sub-goal 3 in 2019 is higher than that in 2021, the degree of achievement of other sub-goals in 2021 is the highest; Intermediate learners have the highest degree of achievement of the four courses in 2019; Advanced learners achieved the highest level of sub-objectives in 2019, except that the sub-objectives in 2019 were lower than those in 2021.

Therefore, the research will analyze the time series data generated during the learning process of 2018 level junior, 2016 level intermediate and advanced learners. Figure 4.3(a) - (c) shows the optimal path for different learners. Green, yellow and red refer to learners appearing once, twice, three or more times in the path planning. The path planned by primary learners is short and simple; The path complexity of intermediate learners' planning is related to learners' ability in view of the relationship between junior and senior learners; The path planned by advanced learners is long and complex.

Figure 4.3(a) - (c) refers to the comparison between the optimized teaching duration and the original duration of primary, intermediate and advanced learners. The effect of some knowledge points in education is poor, and the duration has obvious changes. The teaching duration optimized by different levels of the



(a) junior learners

(b) Intermediate learner

(c) Advanced learners

Fig. 4.2: Achievement of Objectives of Primary, Intermediate and Advanced Courses)

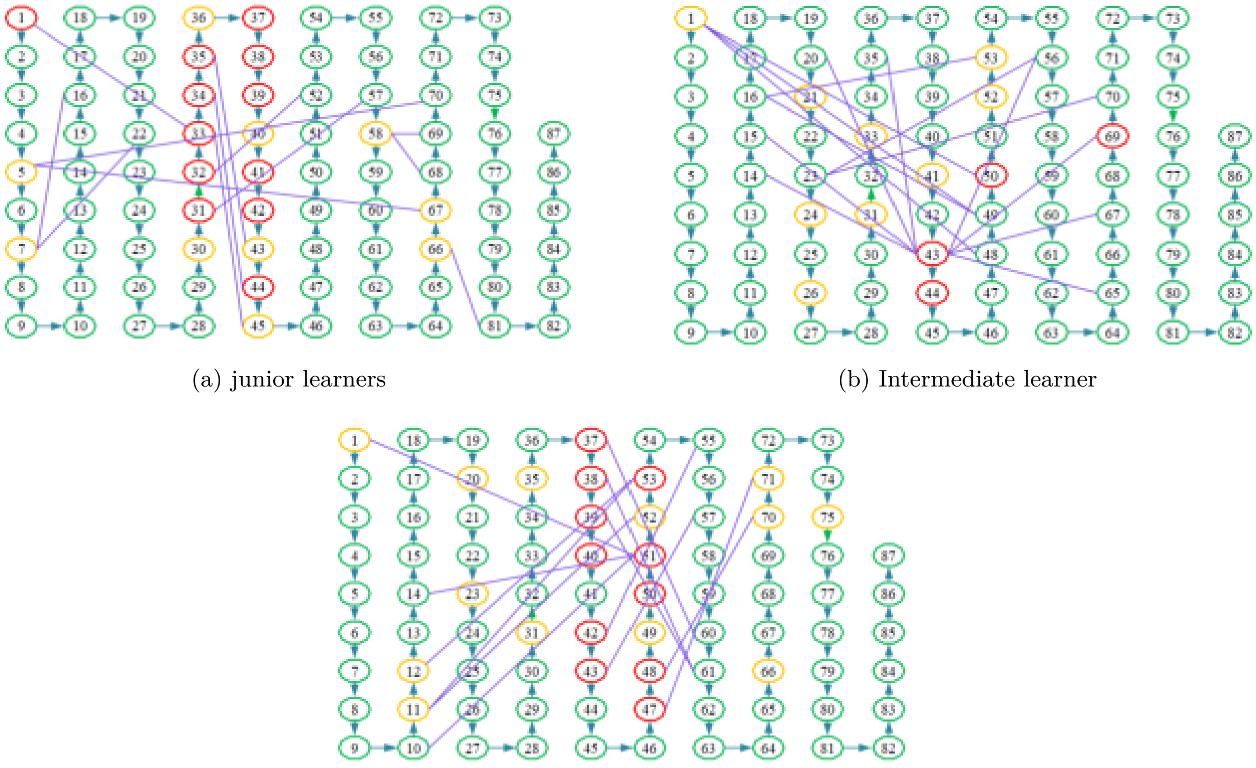
Table 4.2: Relationship Between the Original Video Duration of Some Knowledge Points and the Recommended Teaching Duration

Knowledge points	Original duration	Primary	Intermediate	Senior
26	500	152	147	177
28	371	191	185	229
31	157	1318	1391	1355
32	244	481	429	441
51	113	2131	1780	2152
57	462	117	130	152
89	149	1448	1376	1463

same knowledge points is different. The algorithm can optimize the teaching time of video knowledge points according to different cognitive levels of learning to meet learners' learning needs and improve learners' learning efficiency.

Figure 4.5 shows the relationship between the original video duration and the recommended teaching duration of some knowledge points. According to the actual teaching effect, there are obvious changes in the teaching duration planned by learners, such as knowledge points 26, 31, 89, etc. Different levels of learners have different learning abilities. Different from the teaching time planned by primary learners, the teaching time of most knowledge points planned by intermediate and advanced learners is shorter. The reduction ratio is about 15% and 20%.

Figure 4.5 shows the relationship between the achievement of the overall goal of the course and the similarity



(a) junior learners

(b) Intermediate learner

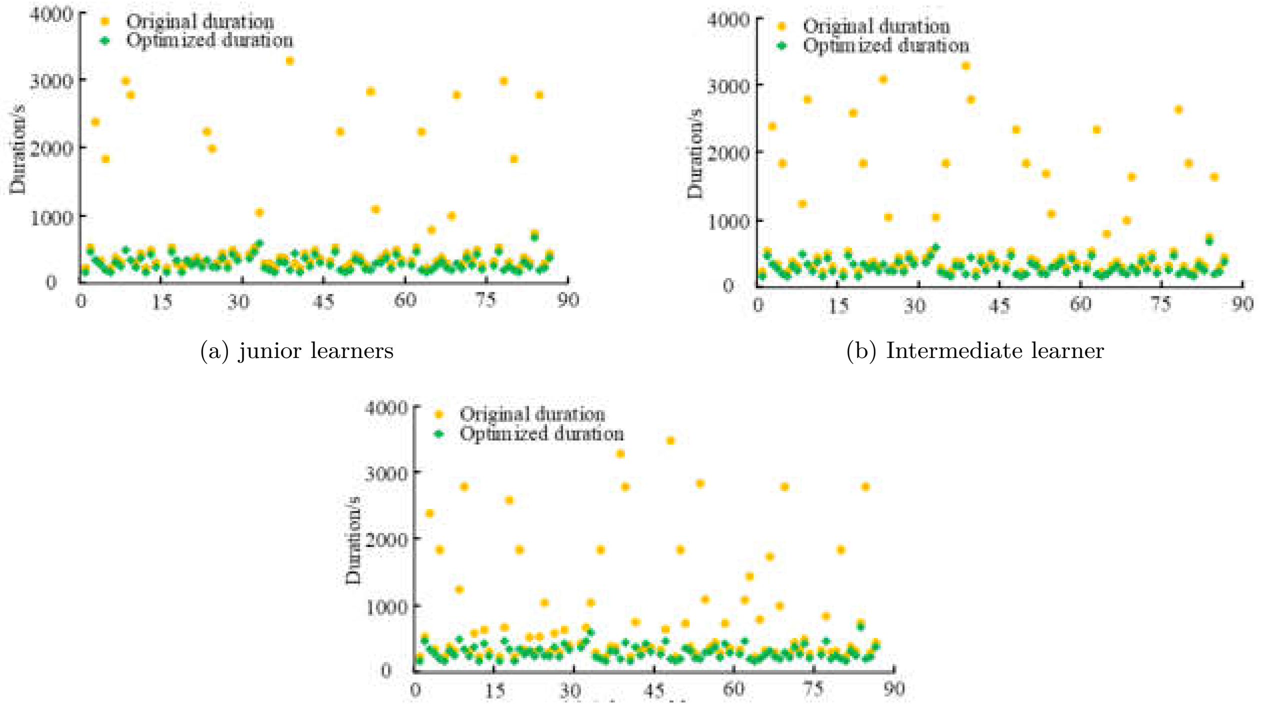
(c) Advanced learners

Fig. 4.3: The Best Learning Path for Different Learners)

of the learning path. For learners with the same cognitive level, the more similar the recommended path and learning path are, the better the overall goal of the course will be achieved. The algorithm given can improve the degree of achievement of the overall goal of the course and help learners to pass the course assessment. For intermediate learners, 40% 50% of the learners' overall course goal achievement degree slightly decreased, and the overall increase with the increase of similarity. The degree of achievement of the overall curriculum objectives of advanced learners increases with the increase of similarity.

Figure 4.6(a) - (c) refers to the relationship between the degree of achievement of curriculum objectives and the degree of similarity for primary, intermediate and advanced learners. Among the primary learners, the four course sub-objectives of the learners whose learning path similarity is less than 40% fail to meet the standard; Among the intermediate learners, except for the four courses with 40% 50% similarity, the degree of achievement of the sub-objectives has decreased, but the overall degree of achievement has increased with the improvement of the similarity; Among advanced learners, except the obvious decrease in sub-goal 1 of the curriculum, the achievement of sub-goal of the overall curriculum is in high degree.

5. Conclusion. To obtain the optimal learning path of learners in a relatively short time, a LPP algorithm based on multidimensional time series data and ant colony optimization was proposed. The path planned by primary learners was short and simple. The path complexity of intermediate learners' planning was related to learners' ability, considering the relationship between junior and senior learners. The path planned by advanced learners was long and complex. The teaching effect of some knowledge points was poor, and the duration had obvious changes. The teaching duration optimized by different levels of the same knowledge points was different. Different levels of learners had different learning abilities. Different from the teaching



(a) junior learners

(b) Intermediate learner

(c) Advanced learners

Fig. 4.4: Comparison Between Optimized Teaching Duration and Original Duration of Primary, Intermediate and Advanced Learners

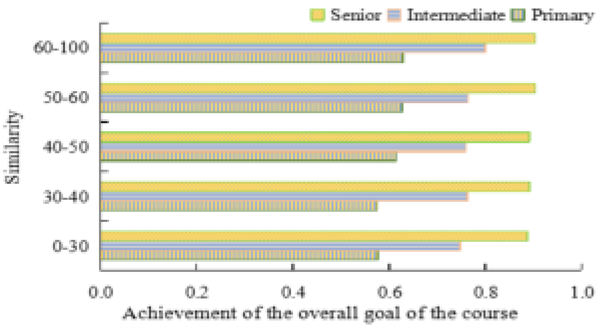
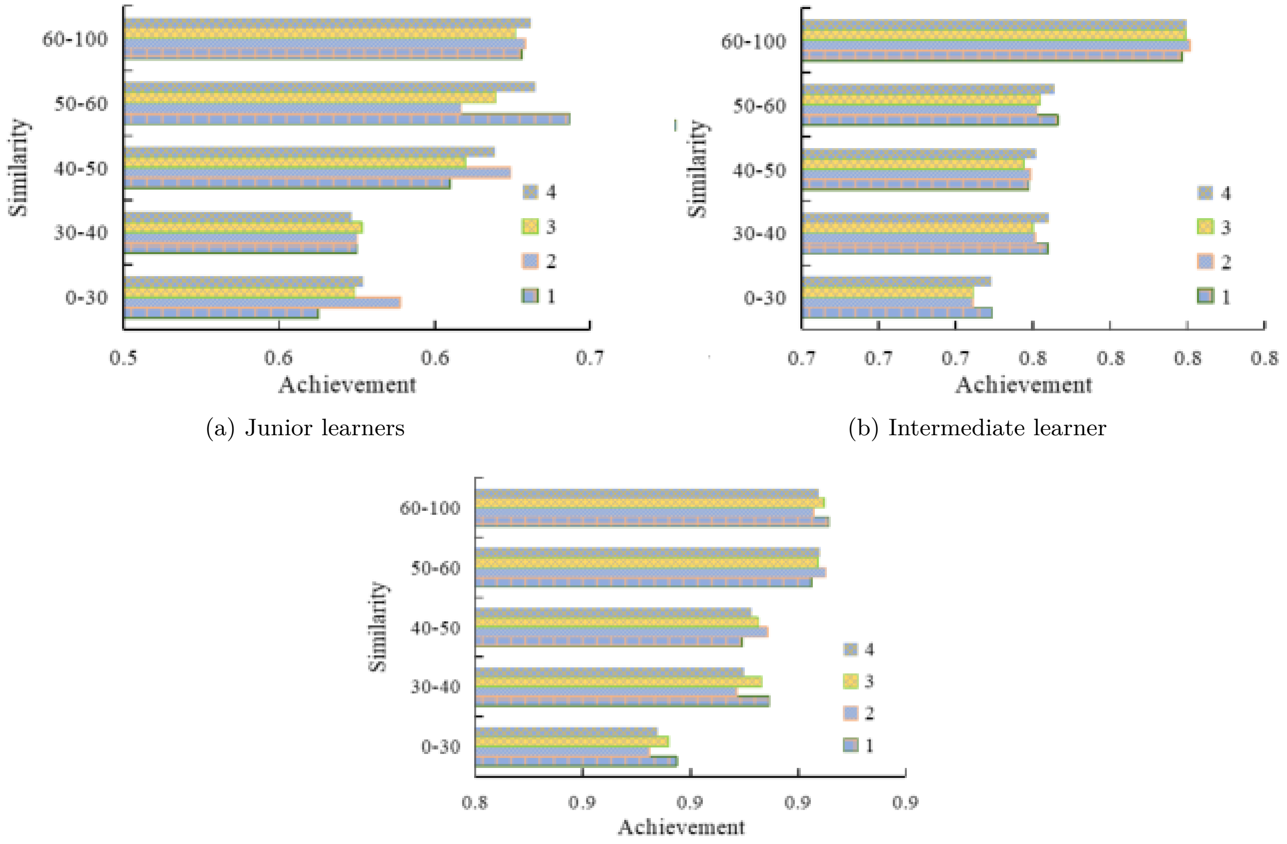


Fig. 4.5: the Relationship Between the Achievement of the Overall Goal of the Course and the Similarity of the Learning Path

time planned by primary learners, the teaching time of most knowledge points planned by intermediate and advanced learners was shorter, with a reduction of about 15% and 20%. Among the primary learners, the four course sub-objectives of the learners whose learning path similarity was less than 40% fail to meet the standard. Among the intermediate learners, except for the four courses with 40% 50% similarity, the degree of achievement of the sub-objectives has decreased, but the overall degree of achievement has increased with the improvement of the similarity. Among advanced learners, except for the obvious decrease in sub-goal 1 of



(a) Junior learners

(b) Intermediate learner

(c) Advanced learners

Fig. 4.6: the Relationship Between the Degree of Achievement of Curriculum Objectives and the Degree of Similarity for Learners

the curriculum, the degree of achievement of sub-goal of the overall curriculum was high. For learners with the same cognitive level, the more similar the recommended path and learning path were, the better the overall goal of the course was achieved. The algorithm given can improve the degree of achievement of the overall goal of the course and help learners to pass the course assessment. The LPP algorithm given by the research plays a key role in improving the teaching quality of teachers and the learning effect of students. Future research could look into more advanced blockchain uses, such as smart contracts, to automate other educational processes, such as assessments and certifications.

Fundings. The research is supported by the undergraduate teaching reform research project of Shandong Province in 2020: "The Construction and Application of College Aerobics online Teaching Resource Library in the post-Epidemic Era" (Project No. M2020098).

REFERENCES

[1] Mb, A., Bh, B. & Me, A. Towards an adaptive e-learning system based on q-learning algorithm - sciencedirect. *Procedia Computer Science*. **170** pp. 1198-1203 (2020)

[2] Zavolodko, H. & Kasilov, O. Interactive tools in online education. *Digital Platform Information Technologies In Sociocultural Sphere*. **3**, 11-21 (2020)

[3] Jing, L., Bo, Z., Tian, Q., Xu, W. & Shi, J. Network education platform in flipped classroom based on improved cloud computing and support vector machine. *Journal Of Intelligent And Fuzzy Systems.* **39**, 1-11 (2020)

[4] Shou, Z., Lu, X., Wu, Z., Yuan, H., Zhang, H. & Lai, J. On learning path planning algorithm based on collaborative analysis of learning behavior. *IEEE Access.* **8**, 19863-11987 (2020)

[5] Chen, H., Ji, Y. & Niu, L. Reinforcement learning path planning algorithm based on obstacle area expansion strategy. *Intelligent Service Robotics.* **13**, 289-297 (2020)

[6] Xie, R., Meng, Z., Wang, L., Li, H., Wang, K. & Wu, Z. Unmanned aerial vehicle path planning algorithm based on deep reinforcement learning in large-scale and dynamic environments. *IEEE Access.* **9** pp. 24884-24900 (2021)

[7] Su, G. Analysis of optimisation method for online education data mining based on big data assessment technology. *International Journal Of Continuing Engineering Education And Life-long Learning.* **29**, 321-335 (2019)

[8] Zhao, Y. & Shan, S. Online learning support service system architecture based on location service architecture. *Mobile Information Systems.* **2021**, 1-11 (2021)

[9] Liu, Y. Interactive system design of entrepreneurship education based on internet of things and machine learning. *Journal Of Intelligent And Fuzzy Systems.* **39**, 5761-5772 (2020)

[10] Li, J., Chen, Y., Zhao, X. & Huang, J. An improved DQN path planning algorithm. *The Journal Of Supercomputing.* **78**, 616-639 (2022)

[11] Wang, J., Chi, W., Li, C., Wang, C. & Meng, M. Neural RRT*: Learning-based optimal path planning. *IEEE Transactions On Automation Science And Engineering.* **17**, 1748-1758 (2020)

[12] Zhang, J., Xia, Y. & Shen, G. A novel learning-based global path planning algorithm for planetary rovers. *Neurocomputing.* **361** pp. 69-76 (2019)

[13] Cui, Z. & Wang, Y. UAV path planning based on multi-layer reinforcement learning technique. *IEEE Access.* **9** pp. 59486-59497 (2021)

[14] Zhou, X., Wu, P., Zhang, H., Guo, W. & Liu, Y. Learn to navigate: cooperative path planning for unmanned surface vehicles using deep reinforcement learning. *IEEE Access.* **7**, 65262-16527 (2019)

[15] Yang, Y., Juntao, L. & Lingling, P. Multi-robot path planning based on a deep reinforcement learning DQN algorithm. *CAAI Transactions On Intelligence Technology.* **5**, 177-183 (2020)

[16] Konar, A., Chakraborty, I., Singh, S., Jain, L. & Nagar, A. A deterministic improved Q-learning for path planning of a mobile robot. *IEEE Transactions On Systems, Man, And Cybernetics: Systems.* **43**, 1141-1153 (2013)

[17] Orozco-Rosas, U., Picos, K., Pantrigo, J., Montemayor, A. & Cuesta-Infante, A. Mobile robot path planning using a QAPF learning algorithm for known and unknown environments. *IEEE Access.* **10** pp. 84648-84663 (2022)

[18] Shou, Z., Lu, X., Wu, Z., Lai, J. & Chen, P. Learning path planning algorithm based on kl divergence and d-value matrix similarity. *ICIC Express Letters.* **15**, 49-56 (2021)

[19] Li, D., Yin, W., Wong, W., Jian, M. & Chau, M. Quality-oriented hybrid path planning based on a* and q-learning for unmanned aerial vehicle. *IEEE Access.* **10** pp. 7664-7674 (2021)

[20] Xiong, S., Zhang, Y., Wu, C., Chen, Z., Peng, J. & Zhang, M. Energy management strategy of intelligent plug-in split hybrid electric vehicle based on deep reinforcement learning with optimized path planning algorithm. *Proceedings Of The Institution Of Mechanical Engineers.* pp. 3287-3298 (2021)

[21] Low, E., Ong, P. & Cheah, K. Solving the optimal path planning of a mobile robot using improved Q-learning. *Robotics And Autonomous Systems.* **115** pp. 143-161 (2019)

[22] Wang, J., Hirota, K., Wu, X., Dai, Y. & Jia, Z. Hybrid bidirectional rapidly exploring random tree path planning algorithm with reinforcement learning. *Journal Of Advanced Computational Intelligence And Intelligent Informatics.* **25**, 121-129 (2021)

[23] Pan, Y., Yang, Y. & Li, W. A deep learning trained by genetic algorithm to improve the efficiency of path planning for data collection with multi-UAV. *Ieee Access.* **9** pp. 7994-8005 (2021)

[24] Xu, X., Cai, P., Ahmed, Z., Yellapu, V. & Zhang, W. Path planning and dynamic collision avoidance algorithm under COLREGs via deep reinforcement learning. *Neurocomputing.* **468** pp. 181-197 (2022)

[25] Chen, P., Pei, J., Lu, W. & Li, M. A deep reinforcement learning based method for real-time path planning and dynamic obstacle avoidance. *Neurocomputing.* **497** pp. 64-75 (2022)

[26] Singhal, V., Jain, S., Anand, D., Singh, A., Verma, S., Rodrigues, J., Jhanjhi, N., Ghosh, U., Jo, O., Iwendi, C. & Others Artificial intelligence enabled road vehicle-train collision risk assessment framework for unmanned railway level crossings. *IEEE Access.* **8** pp. 113790-113806 (2020)

[27] Humayun, M., Jhanjhi, N., Niazi, M., Amsaad, F. & Masood, I. Securing drug distribution systems from tampering using blockchain. *Electronics.* **11**, 1195 (2022)

[28] Kumar, M., Vimal, S., Jhanjhi, N., Dhanabalan, S. & Alhumyani, H. Blockchain based peer to peer communication in autonomous drone operation. *Energy Reports.* **7** pp. 7925-7939 (2021)

Edited by: Kumar Abhishek
Special issue on: Machine Learning and Block-chain based Solution for Privacy and Access Control in IoT
Received: Oct 7, 2023
Accepted: Jan 22, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1455–1465, DOI 10.12694/scpe.v25i3.2672

RESEARCH ON THE RECOMMENDATION SYSTEM OF MUSIC E-LEARNING RESOURCES WITH BLOCKCHAIN BASED ON HYBRID DEEP LEARNING MODEL

SHASHA JIN*AND LEI ZHANG†

Abstract. Learners are confronted with an ever-growing array of diverse and complex educational resources as music education increasingly moves to online platforms. Traditional resource curation methods, which rely heavily on educators, fall short of meeting the dynamic needs of modern students. To address this issue, we present a novel recommendation system for music e-learning resources that combines the power of blockchain technology with a hybrid deep learning model. Our model combines blockchain's robust security and transparency features with advanced deep learning algorithms, enhancing the personalization and efficiency of resource recommendations. A backpropagation neural network with K nearest neighbor classification, traditional collaborative filtering (CF), and an improved CF algorithm are used in the hybrid approach. For the back propagation neural network algorithm, K nearest neighbor classification algorithm, traditional collaborative filtering (CF) and improved CF algorithm, the accuracy rate of improved CF algorithm is higher, reaching 95%. Comparing the proposed model with the association rule-based recommendation model and the content-based recommendation model, the model constructed in this study received high evaluation from experts, with an average score of 98, and more than 97% of them gave a high score of 95 or more, and the evaluation of experts tended to be consistent. Overall, the model proposed in this study can make better recommendations for music education learning resources and bring users a good learning experience, so this study has some practical application value. This research demonstrates a highly effective, blockchain-enhanced recommendation system for music e-learning resources. Our model has significant practical value and potential for adoption in online music education platforms because it provides tailored educational content and an enhanced learning experience.

Key words: music education; recommendation algorithm; learning resources; CF; online education

1. Introduction. The emergence of online learning platforms has altered the landscape of music education in recent years. This digital transition has resulted in an extraordinary profusion of e-learning materials, providing students with a variety of material. The successful curation and suggestion of materials matched to individual learning needs and tastes, on the other hand, is a substantial difficulty. Due to the diverse types of music knowledge and the rich connotation of knowledge, offline music education can no longer satisfy the learning requirements of students, and there is an urgent need to develop online music education so that students can study independently at any time [1]. With the development of online music education, more and more platforms are offering music education resources, and the number of music education resources is growing extremely fast. The traditional way of screening is that teachers screen the resources in advance and recommend the screened resources to students, or students spend some time to screen them themselves [2]. However, the number of resources is too large and the quality of resources varies, and it is very time-consuming to find the right high-quality resources from the huge resource base [3]. In view of this, a large number of scholars have studied the recommendation algorithms of online education repositories, and common resource recommendation methods include content-based, Collaborative Filtering (CF), association rule-based, utility-based, knowledge-based, and hybrid recommendation algorithms [4].

In order to provide learners with high-quality resources that are more suitable for learners' own characteristics, this study chose hybrid recommendation algorithms for model design. In this study, the hybrid algorithm music education recommendation system structure will be designed by using LFM algorithm in the overall recommendation module; in the interest recommendation module, real-time recommendation algorithm based on learners' evaluation scores; in the similar resource recommendation module, CF and content-based recommendation algorithm will be used; in the high rating rate resource recommendation module, learner rating-based

*School of Education, South China Business College Guangdong University of Foreign Studies, Guangzhou, 510545, China (Corresponding Email: Shasha_Jin23@outlook.com)

†Shenzhen Information Technology and Industrialization Association, Shenzhen, 518100, China

The recommendation algorithm based on the quantity of times the learner rated the resource is used in the high rating rate resource recommendation module. Finally, the performance of the model constructed in this study will be evaluated and comparatively analyzed for validating the value of the proposed model in music education. Traditional techniques of resource selection, which mostly rely on instructors, are becoming increasingly insufficient in this dynamic and expanding digital environment.

To solve this issue, we provide an innovative, blockchain-enhanced recommendation engine created exclusively for music e-learning resources. This system uses developing blockchain and hybrid deep learning models to transform how learners access and engage with online music education information. Blockchain technology, which is well-known for its safe and transparent data processing, provides a strong framework for organizing and suggesting educational resources. At the same time, hybrid deep learning models offer advanced analytical capabilities for personalizing content recommendations, ensuring that learners obtain the most relevant and valuable materials. main contribution of the study rely on,

1. Our study is notable for incorporating blockchain technology into the recommendation system. This connection improves the resource suggestion process's security, transparency, and dependability, ensuring that learners receive trustworthy and high-quality content.
2. The model combines complex techniques such as backpropagation neural networks and K closest neighbor classification with collaborative filtering methods to create a hybrid deep learning model. When compared to standard models, our hybrid approach greatly enhances the accuracy and relevance of resource recommendations.

2. Related works. With the newer iterations of online technology, music education is becoming more and more important. Bath N believes that receiving music education is the right of every individual and should not be marginalized from setting courses [5]. Offline music education has emerged, and Kruse and Hill look to inform online education in popular music with a study that provides a detailed analysis of music videos. This study first analyzes the content of music videos, then extracts useful music techniques from them, and finally stores them in a library of music resource learning materials [6]. Camlin et al. explore the impact on learners of the shift from offline to online education models and predict a possible educational crisis, suggesting appropriate responses to this crisis [7]. Daubney with Fautley M, in their study of online music education, found problems with the assessment of student scores and, in light of this, made recommendations for specific teacher tasks and expected teachers to be trained to adapt to the online education model [8]. Cheng and Lam et al. found that the model of online education can also have an impact on teachers, who can be unable to adapt to online education, leading to anxiety and other psychological The teachers are unable to adapt to online education, which can lead to anxiety and other psychological problems [9].

Domestic online education is still in its preliminary stage and there are still a lot of problems that need settling. A large quantity of scholars has paid attention to the problem of recommending learning resources, and there has been a large amount of mature research in the field of network recommendation. Zhao P et al. constructed a new recommendation model based on the recurrent neural network algorithm mixed with the point-of-interest recommendation algorithm. The recurrent neural network assisted the ability of the point-of-interest recommendation method to link the context and predict the data more effectively [10]. Li et al. introduced a multi-objective optimization algorithm to improve the traditional recommendation model for solving the resource overload problem in an online educational system. The results showcased that the model was effective in improving the accuracy and novelty of resource recommendations [11]. Mou et al. presented a new model with the expectation of using it to explore the impact of recommendation algorithms, privacy protection, etc. of short video platforms on users. The research model uses equation modeling to analyze the questionnaire information. The results show that the recommendation algorithm has an impact on users' behaviors and affects their sustained engagement time [12]. Liang and Yin found that the quality of online educational resources is uneven and users' trust in the resources is low. In view of this, they proposed a new recommendation algorithm, expecting to improve users' trust. The algorithm first classifies educational resources and filters invalid educational resources. Then the Kalman filtering method is used to reduce the noise of educational resources and generate a list of highly similar resources for recommendation. The final experimental results show that this model improves users' trust in learning resources [13].

In summary, facing the shortcomings of music education online development, scholars need to conduct a

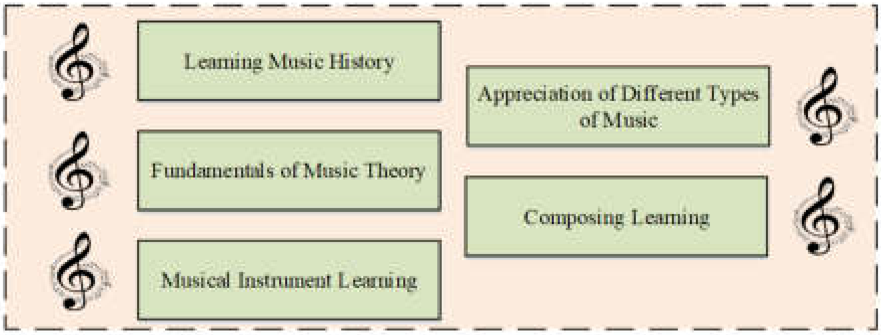

Learning Music History
Fundamentals of Music Theory
Musical Instrument Learning
Appreciation of Different Types of Music
Composing Learning

Fig. 3.1: Knowledge points of music education online course resources

lot of research to improve this. As for the deficiencies in music resource recommendation, a large number of recommendation algorithm studies have been relatively mature, so this study designs a model for music education based on hybrid algorithms, expecting to enhance the quality of resource recommendation and meet the characteristics of student learning, so as to improve students' music literacy.

3. Hybrid algorithm music education resource recommendation system design.

3.1. Hybrid algorithm music education recommendation system structure design. The purpose of this research is to use a hybrid algorithm model to filter recommendations for web resources. In order to better design the hybrid recommendation model, the first step is for designing the recommendation function structure of the recommendation system. The first step of the structure design is to understand the characteristics of the web resources, then to consider the aspects of learners' ratings and learners' interests, and finally to select a suitable recommendation algorithm on this basis. This time, the model divides the knowledge points of music education online course resources into five parts, as shown in Figure 3.1.

As shown in Figure 3.1, the web resources knowledge points are divided into five modules, which are music theory basics, instrument learning, music history learning, different types of music appreciation and composition learning [14]. The five modules involve the cross-application of knowledge points, so it can be seen that learners need to learn multiple knowledge points meanwhile, and the quantity of recommended resource knowledge points in this study is designed with reference to this principle [15]. This method first requires labeling the above five learning resources knowledge points, and the multi-label web resources will have slightly different knowledge point biases when facing different people. Therefore, the study uses the TF-IDF algorithm to assign the weights of knowledge points to accommodate the different labeling needs of various populations. Based on this, the cosine similarity is used to classify the learning resource base, according to the similarity, to tailor the knowledge base for learners and recommend similar resources. The design of recommendation structure from the perspective of learners' ratings is a reasonable direction, but the evaluation scores do not exactly match the quality of learning resources, which can lead to interference with the algorithm. In view of this, this study will introduce another metric: the frequency of evaluation of resources. For this study, in terms of real-time recommendations, the last few ratings of learners are combined with the current resource ratings, and a list of recommendations is generated for similar resources in both. For newly registered users, the system also has a solution. When a new user registers, the system automatically recommends five interest modules for selection, and the user's selection becomes a feature tag, and the system then classifies the tags and sorts and recommends similar resources according to their scores from high to low. According to the previous recommendation system design idea, the flow chart of recommendation function structure is designed as shown in Figure 4.1.

The recommendation system is mainly divided into six modules, which are overall recommendation, interest recommendation, high-frequency resource recommendation, latest recommendation, resource display, and similar recommendation. After the user logs in, the overall recommendation module will show the user a list of resources generated based on historical ratings and the system's collaborative filtering algorithm. The interest recommendation module mainly creates a list of resources needed by learners based on current ratings and historical ratings. The High Score Recommendation module generates a list of learning resources based on the frequency of users' ratings, from highest to lowest. Latest recommendation is to generate a list of learning resources from near to far based on the publisher's release time. The Resource Display module mainly displays the resources so that learners can browse, rate and label the resources. The similarity recommendation module analyzes and organizes similar resources, and generates a list of resources with high similarity. In summary, when a user logs in, the system will recommend to the learner through a mixture of comprehensive recommendation, high frequency resource recommendation, interest recommendation, similar recommendation and latest recommendation.

The login and resource display generate data information, which is because during the login, users select the learning knowledge points according to their preferences, and in the comprehensive recommendation module, they rate the learning resources and other operations. The information from these two parts will be used as the basis for recommendations. The other modules are all about the use of information, based on the data information generated by the user, into the recommendation system for analysis and calculation, and then feedback to the user.

3.2. Hybrid algorithm music education recommendation model construction. The previous section is the structural design of the whole recommendation system for music education, and the later section will introduce in detail the hybrid approach of algorithms and multi-algorithm model construction in the recommendation resource system. Next, the similar recommendation module will be explained in detail. Word frequency -The term frequency-inverse document frequency (TF- IDF) is an algorithm for information retrieval and data mining The common weighting technique of [16]. The similarity module mainly uses a statistical algorithm TF-IDF, which operates in a special analytical mode for semantic contexts, where the TF part is expressed as shown in equation 3.1.

$$TF_{i,j} = \frac{n_{i,j}}{n_{*,j}} \tag{3.1}$$

As shown in equation 3.1, where is used to represent the sentence; $jn_{i,j}$ is used to represent the quantity of occurrences of in ij; $n_{*,j}$ denotes the total quantity of words in j; and the word frequency of i in j is represented by $TF_{i,j}$. After this operation, it is also necessary to express the weight of the words in terms of IDF, whose expression is shown in equation 3.2.

$$IDF_i = \log\left(\frac{N+1}{N_i+1}\right) \tag{3.2}$$

As shown in equation 3.2, N denotes the total quantity of sentences; N_i denotes the total quantity of sentences containing the word i. The overall formula of the unified calculation method is shown in equation 3.5.

$$TFIDF_{i,j} = TF_{i,j} \times IDF_i \tag{3.3}$$

As shown in equation 3.3, where TF localizes the word frequencies and then uses IDF to assign the weights of the words. The similarity recommendation module also uses the Collaborative Filtering recommendation (CFR) algorithm, which can be represented by $U = \{u_1, u_2, \ldots, u_i, \ldots, u_m\}$ for the set of learners m and $I = \{i_1, i_2, \ldots, i_i, \ldots, i_n\}$ for the set of resources [17]. The scoring matrix is shown in equation 3.4.

$$\begin{bmatrix} R_{11} & R_{12} & \cdots & R_{1n} \\ R_{21} & R_{22} & \cdots & R_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ R_{m1} & R_{m2} & \cdots & R_{mn} \end{bmatrix} \tag{3.4}$$

As shown in equation 3.4, R_{ij} represents the rating of by the user i_j . Equation 3.5 demonstrates the formula for calculating the cosine similarity.

$$\operatorname{sim}(u,v) = \cos(\overrightarrow{I_u}, \overrightarrow{I_v}) = \frac{\overrightarrow{I_u} \cdot \overrightarrow{I_v}}{\|\overrightarrow{I_u}\| \times \|\overrightarrow{I_v}\|} = \frac{\sum_{i=1}^{n} R_{ui} R_{vi}}{\sqrt{\sum_{i=1}^{n} R_{ui}^2} \times \sqrt{\sum_{i=1}^{n} R_{vi}^2}} \tag{3.5}$$

As shown in equation 3.5, u and v denote users and i denotes resources; R_ui serves as the rating of u to i and R_{vi} serves as the rating of v to i. The Pearson correlation coefficient is then introduced for calculating the similarity between u and v, and the expression is shown in equation 3.6.

$$\operatorname{sim}(u,v)=\frac{\sum_{i\in I_{uv}}(R_{ui}-\overline{R_u})(R_{vi}-\overline{R_v})}{\sqrt{\sum_{i\in I_{uv}}(R_{ui}-\overline{R_u})^2}\times\sqrt{\sum_{i\in I_{uv}}(R_{vi}-\overline{R_v})^2}} \tag{3.6}$$

As shown in equation 3.6, $\overline{R_u}$ denotes the average of all resource ratings by u and $\overline{R_v}$ denotes the average of all resource ratings by v ; the set of resources jointly evaluated by u and v is denoted by I_uv. Based on this, this study also introduced a correction factor α to improve the calculation of Pearson's correlation coefficient. The improved expression is shown in equation 3.7.

$$\operatorname{sim}'(u,v)=\frac{\min(|I_u\cap I_v|,\alpha)}{\alpha}\times\operatorname{sim}(u,v) \tag{3.7}$$

As shown in equation 3.7, a natural number α can be obtained by training; $I_u\cap I_v$ represents the intersection between the learners' ratings and $|I_u\cap I_v|$ represents the quantity of co-rated resources. Equation 3.7 can also be transformed to equation 3.8.

$$\operatorname{sim}'(u,v)=\begin{cases}\operatorname{sim}(u,v), & \text{if } |I_u\cap I_v|\geq\alpha\\ \frac{|I_u\cap I_v|}{\alpha}\times\operatorname{sim}(u,v), & \text{if } |I_u\cap I_v|<\alpha\end{cases} \tag{3.8}$$

As shown in equation 3.8, when $|I_u\cap I_v|$ is smaller than the correction factor α, the similarity needs to be corrected to prevent the transition prediction of the algorithm, and vice versa, no correction is needed. In addition to considering the similarity for ratings, the similarity between learners can also be considered by classifying the rating levels. Introduce the rank correlation formula as shown in equation 3.9 [18].

$$\operatorname{sim}(u,v)=\frac{\sum_{i\in I_{uv}}(Rank_{ui}-\overline{Rank_u})(Rank_{vi}-\overline{Rank_v})}{\sqrt{\sum_{i\in I_{uv}}(Rank_{ui}-\overline{Rank_u})^2}\times\sqrt{\sum_{i\in I_{uv}}(Rank_{vi}-\overline{Rank_v})^2}} \tag{3.9}$$

As shown in equation 3.9, where the set of resources jointly evaluated by u and v is represented by I_{uv} . Firstly, the learning resources are labeled with corresponding labels. Then the semantic frequencies and weights are calculated according to the TF-IDF algorithm. And finally, the similarity of the three resources is calculated by cosine similarity. The recommendation system proposed in this study lists the similar resources on this basis. In the overall recommendation section, the semantic modeling algorithm is applied, which is based on finding hidden features that are not easily detected in the learners and predicting how the learners rate the recommended resources. The semantic model algorithm uses the idea of regression and its expression is shown in equation 3.10.

$$\hat{R}_{m\times n}=P^T_{m\times k}\cdot Q_{k\times n}\approx R \tag{3.10}$$

As shown in equation 3.10, where $\hat{R}_{m\times n}$ is used to represent the prediction matrix, this matrix is calculated by learning the original matrix R using the algorithm of regression.

$$C=\sum_{(u,i)\in R_0}(R_{ui}-\hat{R}_{ui})^2+\text{Reg}=\sum_{(u,i)\in R_0}(R_{ui}-P_u^T\cdot Q_i)^2+\lambda\sum_u\|P_u\|^2+\lambda\sum_i\|Q_i\|^2 \tag{3.11}$$

As shown in equation 3.11, where Q and R denote two matrices, u denotes a user, and i denotes a different resource. R_{ui} and $\hat{R}_{ui}$ denote a point in the matrix. In order to avoid overfitting, the regularization operation [19] is introduced.

In the first, the learning matrix of users is established; in the second step, the real ratings of users in the database are input into the matrix. In the third step, for the unrated blank part, the system automatically uses the semantic model algorithm for predicting the user's rating; in the fourth step, the system recommends the user according to the rating level. The interest recommendation is obtained by combining the high frequency recommendation and the latest recommendation, in the latest recommendation u denotes the learner, K serves as the quantity of ratings, the set of resources is defined as IK, the set of similar resources is defined as JK , and the cosine similarity calculation formula is introduced again, as shown in equation 3.12.

$$\operatorname{sim}(m,n)=\frac{\sum_{i=0}^{k}(f_{mi}\times f_{ni})}{\sqrt{\sum_{i=0}^{k}f_{mi}^2}\times\sqrt{\sum_{i=0}^{k}f_{ni}^2}} \tag{3.12}$$

As shown in Exhibit 3.12, m and n represent any two resources in the set IK; f_{mi} represents the rating prediction for m and f_{ni} represents the rating prediction for n; i represents the ith learner. The high frequency recommendation is a ranked recommendation of the resources in the similar set, and this ranking before and after the computational expression is shown in equation 3.13.

$$P_{u,j}=\frac{\sum_{i\in IK}(\operatorname{sim}(j,i)\times S_i)}{\text{sim\_sum}}+lg\max(\text{highscore},1)-lg\max(\text{lowcore},1) \tag{3.13}$$

As shown in equation 3.13, a resource in the set of similar resources JK is represented by j. $P_{u,j}$ indicates the priority of the resource.

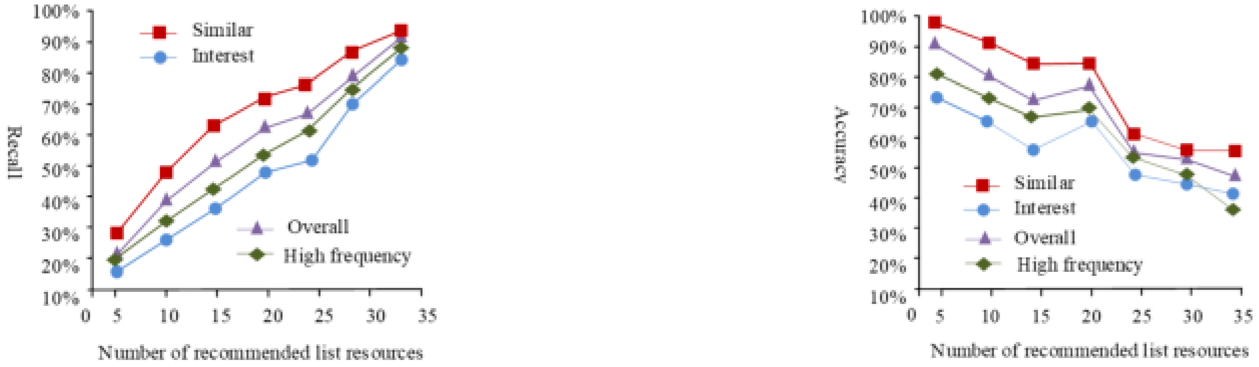
In the Linux environment module there are database with different types of back-end services, and also logs can be generated and processed in this module. Web uses a front-end and back-end separation method for services, and users can log in at the Windows remote end to access different pages[20]. The database is the foundation for the system's correct operation; the Web allows users to log in and provides services such as the presentation of learning resources, user ratings, and so on. The back-end service analyzes and processes system storage logs in order to provide a list of resource suggestions using a recommendation algorithm.

4. Evaluation analysis and comparison of hybrid algorithm music recommendation models.

4.1. Performance testing of hybrid algorithm music education recommendation model. In this study, 50 students in a flipped music education classroom were chose for taking part in in an experiment using the recommendation system. To objectively demonstrate the function of the recommendation system, the students in the selected classroom included three levels of learning: college students, graduate students, and doctoral students. Introducing accuracy and recall as evaluation indicators, we analyzed the recommendations of four modules: similar recommendation, interest recommendation, overall recommendation, and high-frequency recommendation. The changes in accuracy and recall with the number of recommendation list resources are shown in Figure 4.1.

From Fig. 4.1a, it can be seen that the recall rate of each module recommendation algorithm increases steadily with the increase of the quantity of recommendation list resources, among which the recall rate is similar recommendation module, overall recommendation module, high frequency recommendation module, and interest recommendation module in order from high to low; the recall curve of the recall rate of interest module recommendation algorithm fluctuates relatively more, while the rest are relatively stable. Figure 4.1b illustrates that the accuracy rate of the four module algorithms increases when the number of recommended resources is 15 to 20; with the increase of the quantity of recommended list resources, the accuracy rate of each module recommendation algorithm shows an overall decreasing trend, in which the accuracy rate is similar recommendation module, overall recommendation module, high frequency recommendation module, and interest recommendation module in order from high to low. This experiment continues to introduce the F1 value, which is a comprehensive index of recall and accuracy, to judge the number of resources in the best recommendation list, as shown in Figure 4.2.

As can be seen from Figure 4.2, when the number of recommendation list resources is from 5 to 20, the F1 values of all the four module recommendation algorithms increase rapidly; when the number of recommendation list resources is from 20 to 25, the F1 values of the four module recommendation algorithms rise gently; when the number of recommendation list resources is from 25 to 35, the F1 values of the similar recommendation module and the overall recommendation module begin to decline, and the F1 curves of the interest module and



(a) Change in accuracy

(b) Change in recall rate

Fig. 4.1: The variation of recommendation accuracy and recall rate in different sections with the number of recommendation list resources

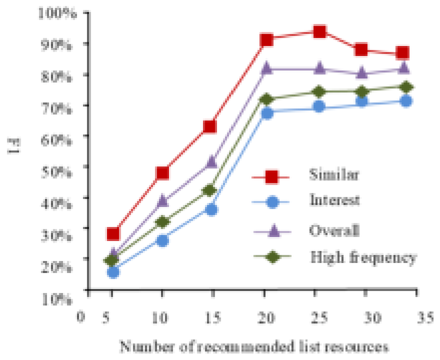


Fig. 4.2: The variation of recommended F1 values for different sections with the number of recommended list resources

the HF module continue to rise gently When the quantity of resources in the recommendation list is between 25 and 35, the F1 values of the similar recommendation module and the overall recommendation module start to decline, while the F1 values of the interest module and the high frequency module continue to rise gently, but the rise is smaller and tends to be horizontal. Overall, when the quantity of resources in the recommendation list is 25, the F1 values of the recommendation algorithms of the four recommendation modules are higher, which can better recommend for learners and bring into play the advantages of the model proposed in this study, in which the F1 values are similar recommendation module, overall recommendation module, high frequency recommendation module, and interest recommendation module in descending order, with F1 values of 95%, 82%, 73%, and 65%, respectively, indicating that the The similar recommendation module runs the best and has the highest accuracy rate, which brings a better experience to users. For further testing the superiority of the model constructed in this experiment and to conduct a comparative analysis of its performance, 150 sets of recommendation algorithm sample data were selected, the same population size and number of iterations were set, and after several iterations, the common recommendation based on Back Propagation (BP) neural network algorithm, k-Nearest Neighbor (KNN) classification algorithm, the traditional CF recommendation model, and the improved CF algorithm used in the similar recommendation module designed in this study were compared

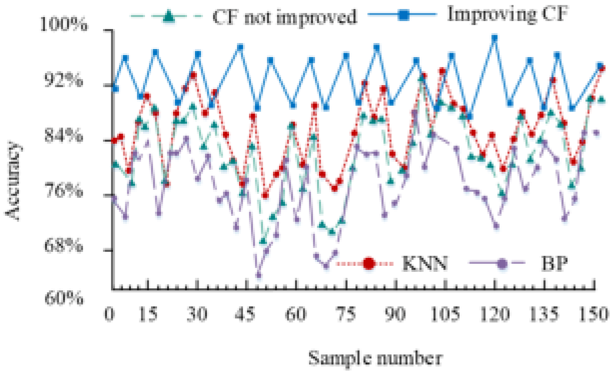


Fig. 4.3: Comparison of accuracy of different model algorithms in different samples

Table 4.1: Page aesthetics evaluation

Page	Extremely beautiful and easy to browse	Beautiful and easy to browse	Beautiful but not easy to browse	Easy to browse but not aesthetically pleasing	Neither aesthetically pleasing nor easy to browse
Navigation	302	169	26	3	0
Registration	185	309	4	1	1
Login	23	400	71	5	1
Interest	34	358	99	7	2
High frequency	187	201	92	20	0
Latest Release	209	189	25	72	5
Resource display	237	205	27	26	5
Overall	398	78	23	1	0

and analyzed, and the accuracy of different model algorithms were compared as shown in Figure 4.3.

Figure 4.3 indicates the accuracy of 100 sets of samples ranged from 60% to 100%. The accuracy of the model based on the improved CF algorithm ranged from 85% to 99%; the accuracy of the model based on the unimproved CF algorithm ranged from 68% to 93%; the accuracy of the model based on the KNN algorithm ranged from 77% to 95%; and the accuracy of the model based on the BP algorithm ranged from 63% to 85%. The accuracy rate based on the improved CF algorithm model was significantly higher than the other algorithms, with an average accuracy rate of 95% for 100 groups; the average accuracy rate for 100 groups based on the unimproved CF algorithm model was 83%; the average accuracy rate for 100 groups based on the KNN algorithm model was 89%; and the average accuracy rate for 100 groups based on the BP algorithm model was 78%. Overall, the improved CF algorithm possesses a higher accuracy rate, not only on the unimproved CF algorithm, but also higher than other algorithms generally applied to recommendation models. Thus, it demonstrates that the improved CF algorithm can be well applied in the music education recommendation model.

4.2. Comparative analysis of hybrid algorithm music education recommendation models. For comprehensively evaluating the superiority of the music education recommendation model proposed in this study, a questionnaire survey will be conducted on 500 users from different aspects to evaluate the hybrid algorithm music education recommendation model proposed in this study, and then the results of the questionnaire survey will be organized and analyzed, as Table 4.1 indicates the evaluation of different users on the aesthetics of the eight pages designed in the method section.

As can be seen from Table 4.1, the overall recommendation page and the login page were more popular

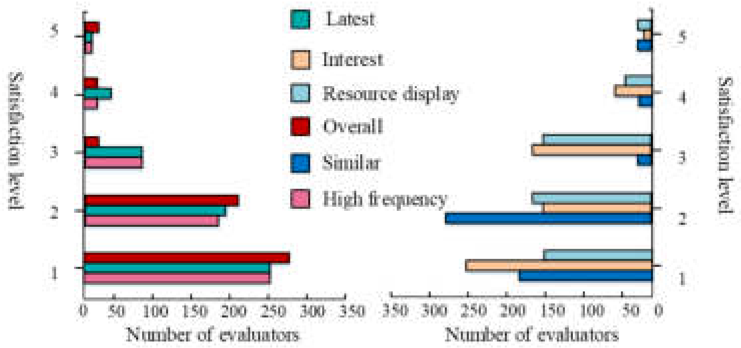


Fig. 4.4: Survey on satisfaction level of different modules in recommendation systems

among users. 398 people rated the overall recommendation page as beautiful and easy to navigate, 78 people rated it as beautiful and easy to navigate, 23 people rated it as beautiful but not easy to navigate, 1 person rated it as easy to navigate but not beautiful, and 0 people gave poor ratings as neither beautiful nor easy to navigate. The login page was rated as beautiful and easy to navigate by 185 people, beautiful and easy to navigate by 309 people, beautiful but not easy to navigate by 4 people, easy to navigate but not beautiful by 1 person, and not beautiful nor easy to navigate by 1 person. Overall, 87.1% of the users rated the pages as beautiful and easy to navigate, beautiful and easy to navigate, which shows that the system pages designed in this study are liked by most users and proves the reasonableness of the pages designed in this study. Next, the six templates designed in this study will be evaluated, as shown in Figure 4.4, which shows the ratings of 500 users on the satisfaction level of each template.

Figure 4.4 illustrates the vertical coordinates of satisfaction level 1 to 5 represent very satisfied, more satisfied, average satisfaction, dissatisfied, and very dissatisfied respectively. It can be seen that users are more satisfied with the interest recommendation module and the overall recommendation module, and the total number of people who are very satisfied and more satisfied with the interest recommendation module is 478; the total number of people who are very satisfied and more satisfied with the overall recommendation module is 472. Overall, users were satisfied with the design of the module, and only 3% of them were dissatisfied or very dissatisfied. For further testing the superiority of the recommendation model for music education constructed in this study, the association rule-based recommendation model and the content-based recommendation model were introduced and compared with the one proposed in this study for analysis, and 20 experts were invited to rate the performance of these three models respectively, as shown in Figure 4.5.

As can be seen from Figure 4.5, the average score of the model in view of association rules is 72, and the score curve fluctuates widely, representing a large difference in experts' evaluation; the average score of the recommendation model based on content is 89, and the score curve fluctuates widely still, with a large difference in experts' evaluation; the model constructed in this study has received a higher evaluation from experts, with an average score of 98, and more than 97% of people give The average score is 98, and 97% of them give a high score of 95 or more, and the fluctuation of the score curve is smaller, and the experts' evaluation tends to be consistent. It proves that the model proposed in this study has superior performance and can give resource recommendations that are more suitable for learners' situations, and has certain applicability in online music education.

5. Conclusion. With the online development of music education, more and more scholars join the research of resource recommendation, and in order to make better recommendations for learners, this research mixes various algorithms to construct a recommendation model for music education resources. In the performance test experiments, the F1 values of the recommendation algorithms of the four recommendation modules are higher when the number of resources in the recommendation list is 25, which are 95%, 82%, 73%, and 65%, respectively, indicating that the table can better recommend for learners when the number of resources is 25,

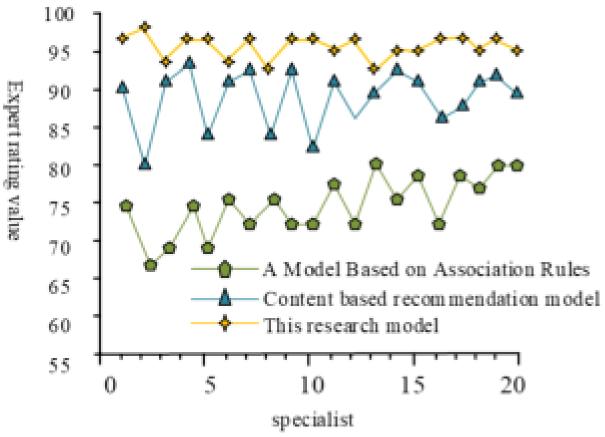


Fig. 4.5: Rating of three models by 20 experts

and take advantage of the model proposed in this study. For the back propagation BP neural network algorithm based on K nearest neighbor Classification algorithm, traditional CF recommendation model, and the improved CF algorithm used in the similar recommendation module designed in this study, the outcomes demonstrate that the improved CF algorithm has a higher accuracy rate of 95%, which shows that the improved CF algorithm can be well applied in the music education recommendation model. In the user questionnaire survey, 87.1% of the users rated the page as beautiful and easy to navigate, more beautiful and easier to navigate; only 3% of the users rated the module as unsatisfactory or very unsatisfactory in terms of satisfaction. In terms of expert ratings, the recommendation model based on association rules and the recommendation model based on content were introduced and compared with the model proposed in this study, and the outcomes indicated that the model constructed in this research received higher ratings from experts, with an average score of 98, and 97% of people gave high ratings of 95 or more, and the fluctuation of the score curve was small, and the expert ratings tended to be consistent. It proves that the model proposed in this study has superior performance and can give resource recommendations that are more suitable for learners' situations, and has some applicability in online music education. However, there are still shortcomings in this study, the update speed of learning resources is slow, and the update speed of resources will be improved on the basis of this study in the future.

REFERENCES

[1] Shaw, R. & And, M. and distance learning during COVID-19: A survey. *Arts Education Policy Review.* **123**, 143-152 (2022)

[2] Ng, D. Ng E H L. *Chu S K W. Engaging Students In Creative Music Making With Musical Instrument Application In An Online Flipped Classroom.* **27**, 45-64 (2022)

[3] Technologies, M. & Impact, T. in Music Education/Aportul tehnologiilor digitale ̂in educația muzical . *Tehnologii Informatice Şi De Comunicaţii În Domeniul Muzical.* **12**, 13-19 (2021)

[4] Wu, C., Liu, S. & Zeng, Z. Knowledge graph-based multi-context-aware recommendation algorithm. *Information Sciences.* **595**, 179-194 (2022)

[5] Bath, N., Daubney, A., Mackrill, D. & Spruce, G. The declining place of music education in schools in England. *Children & Society.* **34**, 443-457 (2020)

[6] Camlin, D. & Lisboa, T. The digital 'turn' in music education. *Music Education Research.* **23**, 129-138 (2021)

[7] Kruse, A. & Hill, S. Exploring hip hop music education through online instructional beat production videos. journal of Music. *Technology & Education.* **12**, 247-260 (2019)

[8] Daubney, A. & Research, F. music education in a time of pandemic [J]. *British Journal Of Music Education.* **37**, 107-114 (2020)

[9] Cheng, L. & Lam, C. The worst is yet to come: the psychological impact of COVID-19 on Hong Kong music teachers. *Music Education Research.* **23**, 211-224 (2021)

[10] Zhao, P., Luo, A., Liu, Y., Xu, J., Li, Z., Zhuang, F. & Zhou, X. Where to go next: a spatio-temporal gated network for next poi recommendation. *IEEE Transactions On Knowledge And Data Engineering.* **34**, 2512-2524 (2020)

[11] Li, H., Zhong, Z., Shi, J., Li, H. & Zhang, Y. Multi-Objective Optimization-Based Recommendation for Massive Online

Learning Resources. *IEEE Sensors Journal.* **21**, 25274-25281 (2021)

[12] Mou, X., Xu, F. & Du, J. Examining the factors influencing college students' continuance intention to use short-form video APP. aslib Journal of Information Management. (2021)

[13] Liang, X. & Yin, J. Recommendation Algorithm for Equilibrium of Teaching Resources in Physical Education Network Based on Trust Relationship. journal Journal of Internet Technology. (2022)

[14] Ornoy, E. & Cohen, S. The effect of mindfulness meditation on the vocal proficiencies of music education students. *Psychology Of Music.* **50**, 1676-1695 (2022)

[15] Piazza, E. & Talbot, B. Creative musical activities in undergraduate music education curricula. *Journal Of Music Teacher Education.* **30**, 37-50 (2021)

[16] Nsugbe, E. Toward a Self-Supervised Architecture for Semen Quality Prediction Using Environmental and Lifestyle Factors[C]//Artificial Intelligence and Applications. 2023. (0)

[17] Sirbiladze, G., Midodashvili, B. & Midodashvili, L. About One Representation-Interpreter of a Monotone Measure. *Journal Of Computational And Cognitive Engineering.* **1**, 51-55 (2022)

[18] Guo, Y., Mustafaoglu, Z. & Koundal, D. Spam Detection Using Bidirectional Transformers and Machine Learning Classifier Algorithms. journal of Computational and Cognitive Engineering. (2023)

[19] Choudhuri, S., Adeniye, S. & Sen, A. Distribution Alignment Using Complement Entropy Objective and Adaptive Consensus-Based Label Refinement For Partial Domain Adaptation[C]//Artificial Intelligence and Applications. 2023. (0)

[20] Wu, S., Tang, Y., Zhu, Y., Wang, L., Xie, X. & Tan, T. Session-based recommendation with graph neural networks [C]//Proceedings of the AAAI conference on artificial intelligence. 2019. (0)

Edited by: Kumar Abhishek
Special issue on: Machine Learning and Block-chain based Solution for Privacy and Access Control in IoT
Received: Oct 7, 2023
Accepted: Feb 11, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1466–1477, DOI 10.12694/scpe.v25i3.2674

ENHANCING IOT SECURITY IN RUSSIAN LANGUAGE TEACHING: A IMPROVED BPNN AND BLOCKCHAIN-BASED APPROACH FOR PRIVACY AND ACCESS CONTROL

QI JIA[*]

Abstract. Russian language instruction emerges as a pivotal course in tertiary education, necessitating novel approaches to maintain instructional quality and efficacy. This study introduces a novel approach to Russian language teaching that combines the robustness of Machine Learning with the security framework of Blockchain technology and is tailored to the unique needs of the Internet of Things (IoT) environment. At its core, the study creates an advanced back-propagation deep neural network enriched with a deep noise-reducing auto-encoder and a support vector machine to improve privacy and access control in IoT-based educational platforms. The proposed model employs a polynomial kernel function and a one-error penalty factor in a single hidden layer, resulting in a system that is not only efficient in handling small-scale data samples but also adept at processing larger data volumes, a common scenario in IoT settings. This design effectively overcomes the problems of overfitting and slow convergence that are common in traditional models. Furthermore, the incorporation of blockchain technology ensures a decentralized and secure data handling framework, reinforcing the privacy and access control aspects that are critical in the digital education domain. The combination of these technologies yields a more rational, scientifically based evaluation system, propelling the standardization and enhancement of Russian language instruction forward. This method not only improves language teaching quality, but it also paves the way for more secure, scalable, and efficient IoT applications in educational settings.

Key words: Teaching quality evaluation; Back propagation; Neural networks; Noise reduction; Support vector machines

1. Introduction.

A major attempt to raise the calibre of instruction and teaching is teaching assessment. The assessment results provide feedback on the quality of the teachers' instruction and serve as a foundation for developing more effective teaching strategies. Learning outcomes are also reflected in teaching assessment, which can be used by students to modify their learning strategies and progress. It is an effective technique to support the management of education and teaching in a scientific and logical fashion, as well as to create a teaching force that is more targeted and concentrated. The variety of contemporary indicators for assessing teaching quality and the complexity of evaluation index aspects make it difficult to quantify a particular indicator in the teaching evaluation process during the teaching phase. The teaching process is characterised by a multi-factor loop, and the interdependence of teachers, teachers, and students creates a straightforward non-linear challenge for evaluating the quality of the instruction. Neurons are arranged in layers in non-linear systems called neural networks. Deep learning's robust information processing capabilities give teaching quality evaluation a contemporary instrument, significantly lowering the subjectivity of conventional teaching evaluation and enhancing its rationality. This shows the value and importance of using neural networks to create a model for assessing the quality of training with the goal to progress scientific teaching objectives and enhance standards for education and instruction.

The Internet of Things (IoT) has emerged as a critical component in the rapidly evolving landscape of digital education, revolutionizing how educational content, including language instruction, is delivered and managed. While this transformation provides unprecedented opportunities for interactive and personalized learning experiences, it also poses significant challenges in terms of data privacy, security, and access control. Russian language instruction in tertiary education, which is becoming increasingly important as Russia's global influence grows, is at the forefront of this digital shift. In this IoT-driven environment, the need to safeguard sensitive educational data and ensure the integrity of the teaching process is more pressing than ever. Our research focuses on developing an improved back-propagation deep neural network model with elements such

[*] School of International Education (Department of Foreign Language Teaching), Yellow River Conservancy Technical Institute, Kaifeng 475004, China (jiaqimark0107@outlook.com)

as a deep noise-reducing auto-encoder and a support vector machine that is specifically tailored for the context of Russian language instruction in IoT environments. This model aims to improve not only the effectiveness of language teaching but also the inherent security concerns associated with IoT-based educational systems. Within this framework, the integration of ML and Blockchain promises to deliver a more secure, efficient, and personalized educational experience.

Use of neural networks for teaching assessment relies on the development of a solid scientific model to evaluate educational quality [1, 2]. However, the low computing efficiency, sluggish convergence, and insufficient accuracy of current assessment models make further investigation and development of evaluation models necessary. To address the issues of overfitting and poor accuracy of existing models, this research suggests adaptive backpropagation neural networks and includes deep noise reduction autoencoders and support vector institutions to develop deep backpropagation neural networks on this basis. The research's goal is to create models for evaluating teaching quality that can handle samples from massive data sets.

The application of technology in an IoT environment for educational purposes. This novel approach addresses critical issues in digital education, specifically Russian language teaching. The use of these technologies in the classroom is a significant step forward in terms of improving both the quality of instruction and the security of the digital learning environment. The creation of an improved back-propagation deep neural network model that incorporates a deep noise-reducing auto-encoder and a support vector machine represents a significant step forward in the evaluation and improvement of language instruction quality. This model was created specifically to process and analyze the complexities of language teaching data, making it a useful tool for educational institutions.

2. Related Works. Numerous experts and academics have conducted a number of studies on the conventional teaching quality assessment system in an effort to enhance the teaching quality assurance system, fairly evaluate teaching quality in order to improve teaching standards, and advance education teaching towards scientific standardisation. In order to selectively label sample features, A system of active learning developed by Huang W combines Gaussian process and sparse Bayesian learning. The algorithm's improved performance was later confirmed [3]. Yuan Z analysed and evaluated feature selection techniques based on current automatic scoring systems, employed multiple regression techniques for score evaluation, and confirmed the effectiveness of the algorithm model through carefully controlled tests with the goal to expand the English translation scoring system [4]. Xiaolong developed a model employing evaluation indices from diverse viewpoints for assessing the effectiveness of online education programmes for colleges and universities. The experimental findings revealed that the algorithm model's training error was relatively small [5]. On the basis of the empirical modal decomposition approach and the adaptive complementary method, Sun Q developed the classroom theory teaching quality evaluation model and improved the correlation vector machine. After employing the baseline weights of the genetic algorithm algorithm network, the model can effectively assess the quality of English interpretation training using a process based on genetic algorithms [7].

To enhance the scientific rigour and applicability of teacher assessment, Lin L applied data mining techniques and machine learning methods for data analysis and joint model creation [8]. This was done to prevent subjectivity from influencing teaching evaluation and to advance the thoughtful growth of teaching evaluation. In order to incorporate artificial intelligence methods into classroom evaluation activities, Guo J suggested an integrated model including statistical modelling and integrated learning based on computer vision and intelligent voice recognition. The experimental findings demonstrated the model's superior functionality, with model accuracy as high as 0.905 [9]. To improve the efficiency of online teaching, Ding X et al. used association rule mining techniques for segmentation fusion and autocorrelation matching detection of teaching timeliness and developed an online teaching timeliness evaluation model based on intelligent learning. The simulation results show that the approach has a high level of confidence for assessing how timely online education is [10]. In order to assess the effectiveness of evaluating ideological and political education, Wang Y et al. employed machine learning and artificial intelligence to develop a fuzzy hierarchical analytic model of the quality of ideological and political teaching. The model uses a three-layer structure to establish a model network structure for data administration, modification, and management of the model assessment. A database for real-time updating was also built. The outcomes of the simulation experiment show that the research model meets the criteria for assessing the efficacy of ideological and political training in universities and other institutions [11]. To address

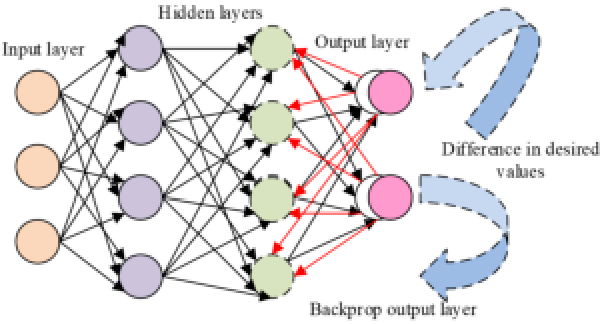


Fig. 3.1: Functional diagram of the back propagation method

the flaws in the taekwondo teaching model used in colleges and universities and to enhance the teaching effect of taekwondo, Liang H developed a taekwondo teaching effect evaluation model based on the intelligent algorithm of human feature recognition using support vector institutions. The performance of the model was confirmed using controlled trials and quantitative statistical techniques, and the concept has some practical applications in classroom education [12].

The aforementioned research on teaching quality evaluation models demonstrates that there are still some gaps in the current findings, and it is relatively uncommon to see teaching quality evaluation models built by combining deep neural networks to solve the issues of fuzzy model index weights, excessive randomness, and easy over-fitting of models, which have significant ramifications for handling large-scale data set samples.

3. Deep Neural Network Evaluation Model Construction Based on Improved BP.

3.1. Construction of a BPNN Evaluation Model Incorporating Adaptive Learning Rate and Momentum Terms. This study addresses the shortcomings of existing models and methods for processing evaluation datasets, with such improvements enhancing the gradient descent method of back propagation neural networks (BPNN) and speeding up the convergence of the model. This study also adds support for vector institutions and deep noise reduction autoencoders to the adaptive BP neural network, a change that can help to handle large evaluation sample data.

The basic processing units for algorithm learning are the neurons, which are the building blocks of artificial neural networks. The function of BPNN's back propagation approach is depicted in Figure 3.1 [13, 14]. Back propagation neural networks are more fundamental neural networks that use forward propagation for output results and back propagation for error propagation. Figure 3.1 illustrates the process of input, processing, computation and output of data in forward propagation. When the error is back-propagated, the error layer determines the discrepancy between the target's desired value and the output's actual value. The error value will adjust the neuron weights and thresholds of each layer until the error reaches the required end of the algorithm. The error is transported forward from the output layer(OL) through the HL in an inverse forward propagation way.

The study builds a three-layer BPNN using forward propagation learning to initialise the network with the number of god will elements as n, p, and q in each layers, respectively. Equation 3.1 shows the input value.

$$h_{ij}(k) = \sum_{i=1}^{n} w_{ij} x_i(k) - b_j \tag{3.1}$$

In equation 3.1, k is a sample chosen at random, $x(k)$ is the input vector, W_{ij} is the network's connection weight, and b_j is the threshold value chosen at random from the range $(-0.5, 0.5)$. Equation 3.2 displays each neuron's output value from the HL, h_o..

$$h_o(k) = f(h_{ij}(k)) \tag{3.2}$$

In accordance with 3.3, the method for determining each neuron's input value y_i in the OL based on the HL's output value, the connection weights, and the OL's threshold value is shown.

$$y_{i_t}(k) = \sum_{j=1}^{p} w_{jt} h_{o,j}(k) - b_t \tag{3.3}$$

W_{jt} is the connection weight and b_t is the threshold value in equation (3). Similar to how the input value y_i of the neuron is used to determine the output value y_o , as shown in equation 3.4.

$$y_{o,t}(k) = f(y_{i,t}(k)) \tag{3.4}$$

The target accuracy of the network is set to ϵ during the backpropagation phase of the BPNN, commonly known as error backpropagation. When the accuracy is less than the set accuracy, the bias derivatives of the neurons in the OL are calculated in equation 3.5.

$$\delta_t(k) = y_o(k)(1 - y_o(k))(t(k) - y_o(k)) \tag{3.5}$$

$\delta_t(k)$ stands for the partial derivative in equation 3.5. Equation 3.6 illustrates how the connection weights AW_{jt} and the threshold b_j AA between the HL and the OL are corrected using the derived partial derivatives and the neuron outputs of the HL. In equation 3.6, N and $N+1$ represent before and after correction respectively, and μ represents the learning step.

$$\begin{cases} w_{jt}^{(N+1)}(k) = w_{jt}^{N}(k) + \mu\delta_t(k)h_{o,j}(k) \\ b_t^{(N+1)}(k) = b_t^{N}(k) + \mu\delta_t(k) \end{cases} \tag{3.6}$$

In a similar manner, the HL neuron's partial derivative $\delta_h(k)$ is computed, as shown in equation 3.7.

$$\delta_h(k) = \left[\sum_{t=1}^{q} \delta_t(k) w_{jt}\right] h_{o,j}(k)(1 - h_{o,j}(k)) \tag{3.7}$$

The connection weights W_{ij}, b_j between the input and HLs are corrected, and the procedure is shown in equation 3.8.

$$\begin{cases} w_{ij}^{(N+1)}(k) & = w_{ij}^{N}(k) + \mu\delta_h(k)x_i(k) \\ b_j^{(N+1)}(k) & = b_j^{N}(k) + \mu\delta_j(k) \end{cases} \tag{3.8}$$

Finally, determine whether the global error meets the required precision, if so, the algorithm learning ends; otherwise, samples are chosen to recalculate the input and output values of the HL neurons until the error meets the requirements or the algorithm iteration ends. The global error calculation is illustrated in equation 3.9.

$$E = \frac{1}{2n} \sum_{k=1}^{n} \sum_{t=1}^{q} (t_t(k) - y_t(k))^2 \tag{3.9}$$

The BPNN takes a long time to train or even fails to converge well, may fall into local minima during learning, uses gradient descent to make the error converge very slowly, and the training results are unstable. To address these problems, the model enhances the BPNN by introducing adjustable learning rate and momentum components. The number of neurons n and q in the input and OLs are determined according to the input sample dimension and the output result dimension, and the number of neurons p in the HL is determined according to the empirical equation 3.10. in equation 3.10 is a constant between $[1, 10]$.

$$p = \sqrt{n + m} + a \tag{3.10}$$

The Adaptive Gradient (AdaGrad) method's learning rate dynamically adapts in response to network fault[15-16]. Equation 3.11 illustrates the process of adaptive learning rate change. In equation (11), $\mu(0)$ represents the starting learning rate, and in this investigation, β and γ have the values 1.05 and 0.7, respectively.

$$\mu(n)=\begin{cases}\beta\mu(n-1) & \text{if } E(n)<E(n-1) \quad \text{and} \quad 1<\beta<1.5\\ \gamma\mu(n-1) & \text{if } E(n)>E(n-1) \quad \text{and} \quad 0.5<\gamma<1\\ \mu(n-1) & \text{otherwise}\end{cases} \tag{3.11}$$

Equation 3.12 illustrates the inclusion of the momentum factor in the adaptive learning rate approach, which serves as a dampener in the process of the error back propagation correction weight. In equation 3.12, α stands for the momentum term, w for weight, and w for moment.

$$\Delta w(n)=-\mu\sum_{t=0}^{n}\alpha^{(n-t)}\frac{\partial E(n)}{\partial w(n)} \tag{3.12}$$

The weighting adjustment equation is shown in equation 3.13, where the learning rate is represented by μ and the error is represented by $E(n)$.

$$w(n+1)=w(n)-\mu(n)\sum_{t=0}^{n}\alpha^{(n-t)}\frac{\partial E(n)}{\partial w(n)} \tag{3.13}$$

3.2. Deep Neural Network Evaluation Model Construction Based on Improved BPNN. The momentum terms as well as the adaptive learning rate in the BPNN evaluation model have certain advantages when handling small-scale datasets, but their capacity to handle complicated and high-dimensional large dataset samples is constrained. The paper builds a deep network model to solve this issue by layering deep noise reduction autoencoders over BP neural networks and adding Support Vector Regression (SVR) to the OL.

Artificial neural networks are deepened by deep neural networks, which have many HL. Deep noise reduction autoencoders have a stronger ability to extract essential features than the original autoencoders because they consist of many autoencoders that add noise to the data set to prevent overfitting during training [17, 18].

As seen in Figure 3.2, the feature of zeroing is mostly used for noise reduction processing of the noise contained in the input original data. First, set a particular probability to set part of the data in the original matrix x to 0 to get the residual input matrix $\widetilde{x}$ with lost data. The compressed matrix y is obtained by layer-by-layer coding, followed by layer-by-layer pass to obtain x' , error between x and x' for network parameter learning, and iteration to obtain the compressed coded y . The entire training process improved in robustness and generalizability.

To minimise the error and complete reconstructing the original input dataset, the error between the reconstructed dataset and the original dataset is then calculated using an error function, and the BP algorithm is used to propagate the error to the entire depth noise reduction autoencoder and modify the weights and thresholds. The cost function is the mean squared error function, whose expression is given in equation; the weights and thresholds are updated using the gradient descent method 3.14.

$$L(x,y)=\frac{1}{n}\sum_{i=1}^{n}(y_i-x_i)^2 \tag{3.14}$$

The Adam algorithm combines the two well-known techniques "Adagrad" (for sparse gradients) and "RMSPro" to solve optimisation issues involving vast amounts of data and high feature latitude. (for non-stationary data). Figure 3.3 illustrates how the Adam approach, which is computationally efficient and ideal for very noisy and sparse gradient issues, can replace the conventional random gradient descent method to update the network weights more effectively. It also functions as a deep noise reduction autoencoder. The deep noise reduction autoencoder's OL neurons are shown as dashed circles in Figure 3.3, while the true output is a classifier or predictor.

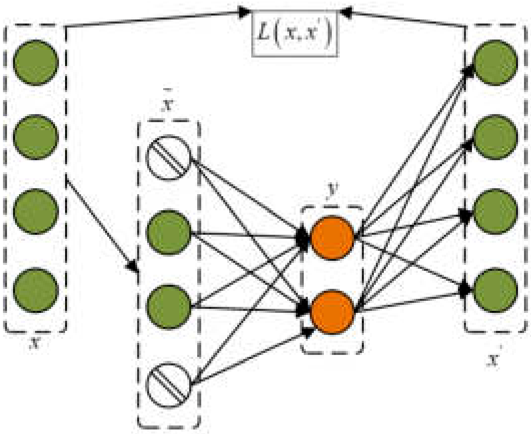


Fig. 3.2: Structure of Noise Reduction Automatic Encoder

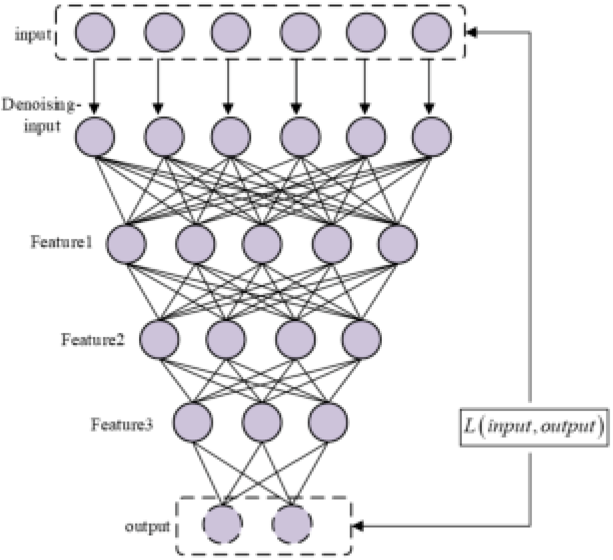


Fig. 3.3: Structure of deep automatic noise reduction encoder

Support Vector Machines (SVM) are a subclass of generalised linear classifiers that conduct supervised learning (supervised learning) binary categorization of data for tasks including regression, classification, and recognition. A linear classifier is a line that divides two groups of data in a two-dimensional plane. A plane serves as a linear classifier in three dimensions. In higher dimensions, a hyperplane is created from a linear classifier. Linear and non-linear regression are two categories for support vector analysis. In the former, complex nonlinear relationships are mapped into a high-dimensional space, where they are then realised and behave like linearized relationships in the latter [19, 20]. In response to the various indications for evaluating the quality of Russian language education and the complex non-linear connection between indicators and evaluation conclusions, the study uses support vector non-linear regression.Support vector regression serves as the predictor for the final OL in the deep noise reduction auto-coding unsupervised training layer of the final deep neural network evaluation model built using the improved BPNN, and the error between the original input dataset and the unsupervised

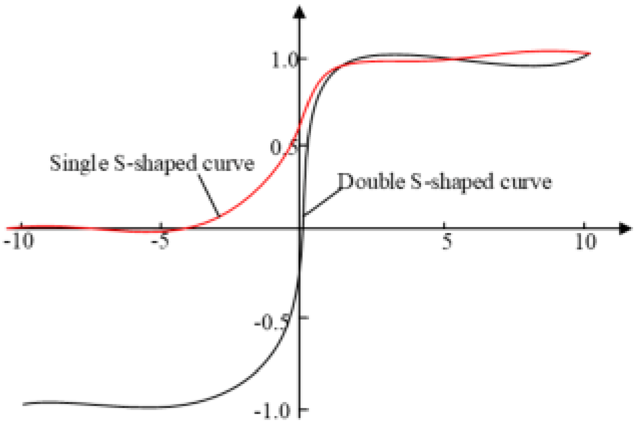


Fig. 3.4: Sigmoid function curve

training output data is minimised to obtain the feature vectors of the original input dataset. The support vector regression model's structure is depicted in

The data is often pre-processed before being fed into the model, and the study normalises the sample data to transform the data between intervals . One way to lessen the difficulty of weight adjustment is to scale back the amount of the input value change. On the other hand, Figure 4.1 depicts the activation function of the BPNN as a double S-shaped curve.

The transformation between [-1,1] and the derivative of the activation function is identical, Which is shown in Equation 3.15.

$$y = f(x) = \frac{1 - e^{x}}{1 + e^{x}} \tag{3.15}$$

The normalisation operation can speed up the convergence of the network and improve the computational efficiency. The operation process is shown in equation . In equation , x_{max} and x_{min} denote the maximum and minimum values in the data, x_i and z_i denote the data output before and after processing, and x_{mid} denotes the intermediate values of data changes.

$$\begin{cases} z_i = \frac{x_i - x_{min}}{\frac{1}{2}(x_{max} - x_{min})} \\ x_{mid} = \frac{x_{max} + x_{min}}{2} \end{cases} \tag{3.16}$$

4. Performance Testing of Improved BP Deep Neural Network Evaluation Models.

4.1. Test Experimental Protocol Design and Model Parameter Analysis. A test experiment was created to confirm the effectiveness of the built model. The experiment identifies metrics for assessing the effectiveness of Russian language instruction from two perspectives—student evaluation and teaching supervision groups—as well as from two dimensions—preparation before instruction and during instruction. These metrics include teaching attitude, teaching content, teaching methods, and answering questions after class. They also include professional quality, teaching ability, preparation before instruction, and the energy of the classroom environment. The dataset originates from a university academic system's dataset on the evaluation of Russian language courses and contains 3684 examples of data. Student evaluations are used as model input values and the evaluations of the teaching supervisory team are used as the target expectation values of the algorithm model. Finally, all data are normalised to increase the algorithm's calculation efficiency.

The amount of neurons in the input layer was set to 30, and the amount of neurons in the output layer to 1, the growth ratio of the adaptive learning rate to 1, the decline ratio to 0, and the momentum term to 0.65

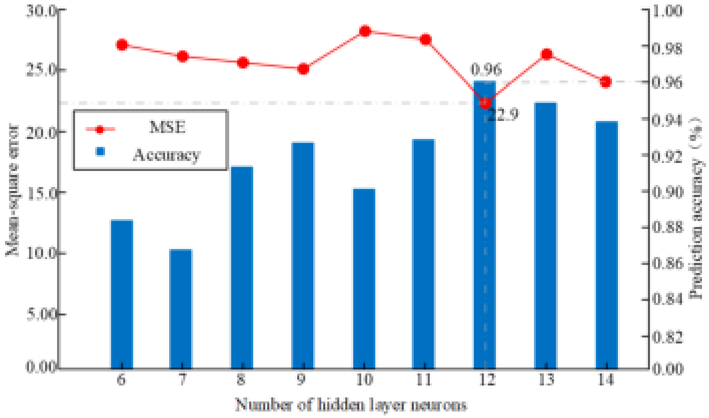


Fig. 4.1: Effect of the number of neurons on mean square error and accuracy

in order to compute the number of neurons in the HL. The number of neurons in the HL was determined by applying Equation 3.10 and utilising the mean-square error (MSE) and prediction accuracy as the assessment indices. The optimum amount of neurons for the HL was determined using Mean Square Error (MSE) and prediction accuracy. The training outcomes are shown in Figure 4.1. Figure 4.1 demonstrates that the mean squared error is at a minimum of 22.9 and the prediction accuracy is at a maximum of 0.96 when there are 12 neurons in the HL. The minimal prediction accuracy is only approximately 86%, and the mean squared error does not change significantly when the number of neurons changes, but the accuracy value fluctuates more. The evaluation model's accuracy is taken into account for determining the HL's number of neurons, which is set at 12.

2, 3, 4, and 5 HLs, together with 12 HL neurons, were chosen as the parameters. The unsupervised training was done using Adam's technique, and the evaluation index was the difference in error between the reconstructed data feature vector and the original data set. Figure 7 displays the training outcomes after 5000 iterations. Figure 4.2 illustrates how the error value curves all exhibit a declining trend as the number of iterations rises. The model with two HLs exhibits the highest decline in error value, with a 68.% drop from the start of the iteration. When there are just two HLs, the algorithm model's training result is perfect; nevertheless, with the same number of repetitions, the error value climbs steadily as the number of HLs rises.

The penalization coefficient and the kind of kernel function are the two primary factors influencing support vector regression in the supervised prediction output process. Model complexity and empirical riskiness are both impacted by the penalty coefficient, and modifying these two factors enhances the algorithm's overall performance. To calculate the Mean Absolute Percentage Error (MAPE) between the predicted evaluation result value and the actual evaluation value, the penalty coefficients are taken to be in the range of 1 to 9, and the kernel functions are taken into consideration to be Liner, Poly, radial basis function, and Sigmoid function. Figure 4.3 displays the model training outcomes and the evaluation index, the MAPE. In accordance to Figure 8, despite the lesser degree of error fluctuation, the excessive penalty still causes the MAPE to be excessively large. The MAPE of the support vector machine consisting of all kernel functions roughly tends to increase as the error penalty factor increases. In comparison to the other three types of functions, the polynomial function has the significantly lowest MAPE value, with a MAPE value of only 0.0506 when the penalty coefficient is 1. Support vector regression uses the polynomial function as its kernel function.

4.2. Quality Analysis of Model Training Results. With two HLs, 12 neurons each, and an error penalty value of 1, the adaptive learning rate was set to 1.1 for growth ratio, 0.8 for decline ratio, and 0.65 for momentum term. The kernel function for the training of the model was decided to be a polynomial function. To compare the MAPE, MSE, Root Mean Square Error (RMSE), and Symmetric Mean Absolute Percentage Error (SMAPE) of the various algorithms, the BP deep neural network developed in this study was first compared

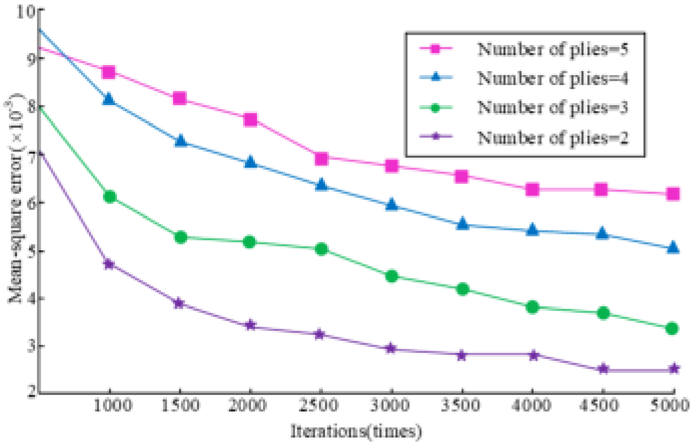


Fig. 4.2: Effect of the number of HLs on mean square error

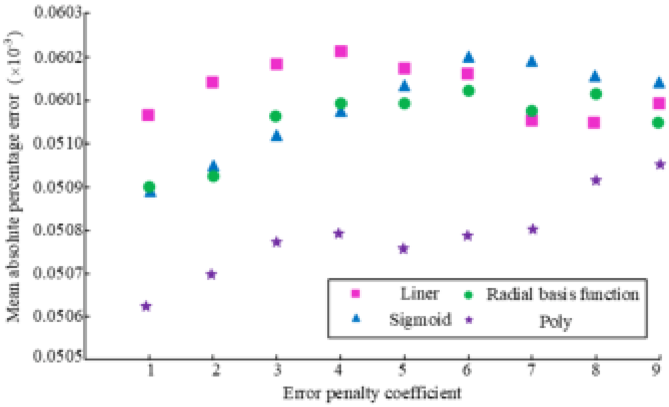
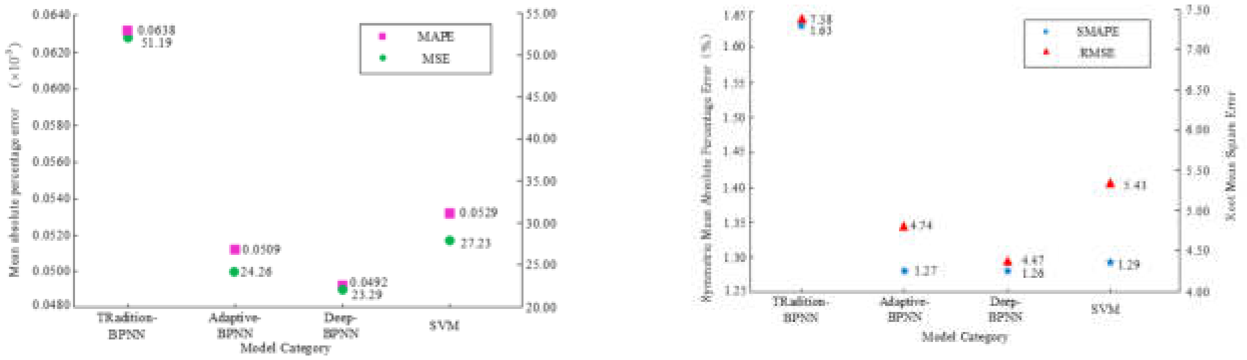


Fig. 4.3: Effect of error penalty coefficient on MAPE of different kernel functions

with the adaptive BPNN, traditional BPNN, and support vector machine algorithms. Figure 8 displays the training outcomes and the RMSE. Figure 4.4 shows that the BP deep neural network developed in this study had the lowest values for all four metrics, with MAPE of 0.0492, MSE of 23.29, SMAPE of 1.26, and RMSE of 4.47. The typical BP neural network had higher error values for all four metrics. In particular, the MAPE and MSE measures were 25.34 and 111 percentage points higher than the BP deep neural network's, making them inappropriate for use as direct assessment models in comparison. The SMAPE values of the adaptive BPNN were the ones that were closest to those of the BP deep neural network. Although they performed slightly worse in terms of the magnitude of the other three errors, overall performance was not significantly different, proving that the construction of adaptable BP neural networks was correct and highlighting the need for further advancements in adaptable BP neural networks. When the algorithm learning was finished, the comparison of training time and accuracy is continued, and the results are displayed in Figure 4.5. Figure 4.5 demonstrates that the enhanced adaptive BP neural network and the deep neural network have much higher accuracy values, up to 5 percentage points higher than the traditional BP neural network. However, the accuracy rates of the four networks are not significantly different. The adaptive BP neural network, however, took the shortest amount of time to train—only 1.07s—a difference of 10.63 seconds from the traditional BPNN and 2.9 seconds from the deep BP neural network. This shows that the adaptive BP neural network handles the concerns with



(a) MAPE, MSE

(b) SMAPE, RMSE

Fig. 4.4: Comparison of network performance of different algorithms

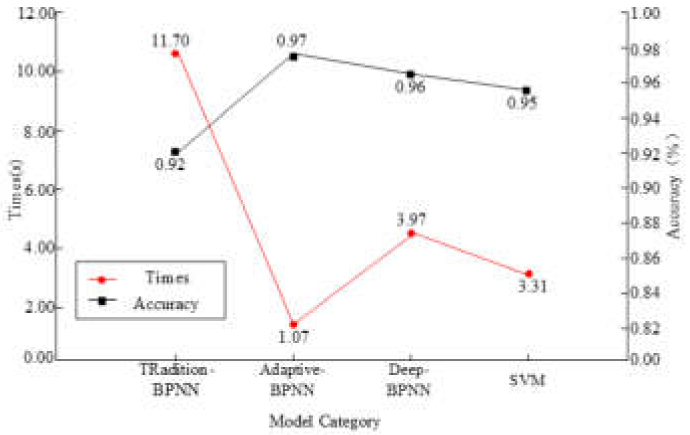


Fig. 4.5: Network training time and accuracy of different algorithms

delayed convergence and slipping into local minima that the classic neural network experiences and is more suited to handling tiny data sample sets. The deep BPNN exhibits some improvements in error values, but these advantages are not significant because the deep BPNN's structure is complex and the number of HLs grows, which lengthens processing time for small sample sets. The performance of the BP deep neural network was then compared when it was used with various optimisation techniques, such as BPNN-Gradient Descent, BPNN-Momentum, and BPNN-RMSProp. The training results are displayed in Figure 4.6s. The BP deep neural network and the BPNN-RMSProp algorithms' mean squared error values reduced the quickest and had the sharpest curve trend below 1000 iterations, as can be seen in Figure 4.6. As the number of iterations increases, the error curve flattens out and the error values do not decrease significantly, even though the BPNN created using Adam's optimisation algorithm in this study had the best results in terms of reconstructing the input data at the end of unsupervised learning training and had the lowest error values at the end of the iterations. The error values of the Gradient Descent algorithm and Momentum algorithm also showed a decreasing trend, but the error values were larger and the algorithm's overall training performance was not better. However, the error curves of the Gradient Descent and Momentum algorithms also show a decreasing trend. Finally, the sample data were normalised on the large-scale dataset to highlight the benefits of BP deep

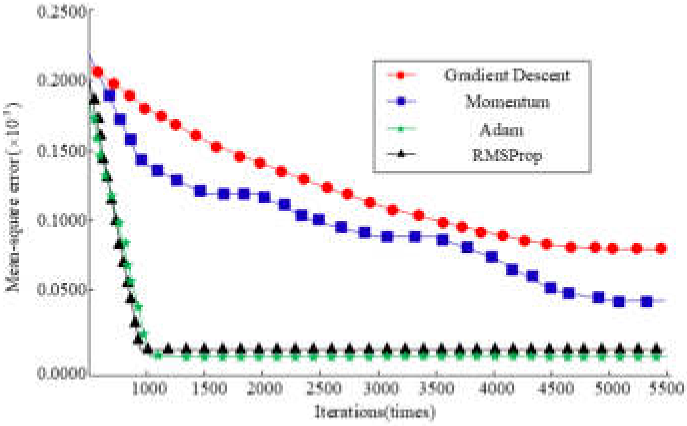


Fig. 4.6: Performance Comparison of Different Optimization Algorithms BPNN

Table 4.1: Performance Comparison of Large Datasets

Model	MAPE	MSE	SMAPE	RMSE	Times
Adaptive-BPNN	0.2453	68.9	6.890	9.087	45.98
Deep-BPNN	0.0876	28.6	3.002	7.930	79.24
TRadition-BPNN	1.6274	106.6	8.236	11.231	123.69
SVM	0.8952	89.7	7.263	8.563	86.66

neural networks on large-scale datasets. The final training results are displayed in Table 1 after comparing the MAPE, MSE, SMAPE, and RMSE values and training duration of the four networks. Table 4.1 demonstrates how Deep-BPNN outperforms Adaptive-BPNN in terms of error performance measures when processing massive datasets. These metrics are all significantly lower for Deep-BPNN. The training duration was 79.24 seconds, but even though there were more HLs and total HL neurons, the training time was still within a reasonable range.

5. Conclusion. This study enhanced the conventional BP neural network by including a support and a deep noise reduction autoencoder vector mechanism to the adaptive learning BPNN to create a deep neural network to meet the challenging nonlinear problem of evaluating the quality of teaching Russian. The results of the model performance test indicate that 12 neurons, with a mean squared error of 22.9 and a prediction accuracy of 0.96, are the ideal number for a single HL. When there are two HLs, the error curve of the built-in deep neural network model shrinks the quickest, reaching a maximum reduction of 68.3%. The error penalty factor was adjusted to 1 using the polynomial function, which is best for enhancing the algorithm's overall performance. The MAPE value at the end of the model training was only 0.0506. With a MAPE of 0.0492, an MSE of 23.29, a SMAPE of 1.26, and an RMSE of 4.47, the BP deep neural network outperformed the adaptive BPNN, regular BPNN, and support vector machine algorithms in terms of error values and accuracy magnitudes. The adaptive BP neural network is better suited for processing small-scale data sample sets because its error value is marginally larger than that of the BP deep neural network, but its training time is shorter—only 1.07s, 10.63s less than the traditional BPNN and 2.9s less than the deep BP neural network. With the lowest error value at the end of the iterations and the highest performance at reconstructing the input data throughout unsupervised learning training, the deep BPNN built utilising Adam's optimisation technique clearly has an edge when working with large-scale data sets. Further study is still required to determine the effectiveness and duration of training for the built-in deep neural network model.

REFERENCES

[1] Rong, Z. & Gang, Z. An artificial intelligence data mining technology based evaluation model of education on political and ideological strategy of students. *Journal Of Intelligent & Fuzzy Systems: Applications In Engineering And Technology.* **40**, 3669-3680 (2021)

[2] Liu, G. & Zhuang, H. Evaluation model of multimedia-aided teaching effect of physical education course based on random forest algorithm. *Journal Of Intelligent Systems.* **31**, 555-567 (2022)

[3] Huang, W. Simulation of English teaching quality evaluation model based on gaussian process machine learning. *Journal Of Intelligent And Fuzzy Systems.* **40**, 2373-2383 (2021)

[4] Yuan, Z. Interactive intelligent teaching and automatic composition scoring system based on linear regression machine learning algorithm. *Journal Of Intelligent & Fuzzy Systems: Applications In Engineering And Technology.* **40**, 2069-2081 (2021)

[5] Wen, X. Research on the teaching quality evaluation model of distance education in colleges based on analytic hierarchy process. *International Journal Of Continuing Engineering Education And Life-Long Learning.* **32**, 796-810 (2022)

[6] Sun, Q. Evaluation model of classroom teaching quality based on improved RVM algorithm and knowledge recommendation. *Journal Of Intelligent And Fuzzy Systems.* **40**, 2457-2467 (2021)

[7] Lu, C., He, B. & Zhang, R. Evaluation of English interpretation teaching quality based on GA optimized RBF neural network. *Journal Of Intelligent And Fuzzy Systems.* **40**, 3185-3192 (2021)

[8] Lin, L. Smart teaching evaluation model using weighted naive bayes algorithm. *Journal Of Intelligent & Fuzzy Systems: Applications In Engineering And Technology.* **40**, 2791-2801 (2021)

[9] Guo, J., Bai, L., Yu, Z., Zhao, Z. & Wan, B. An AI-Application-Oriented In-Class Teaching Evaluation Model by Using Statistical Modeling and Ensemble Learning. *Sensors.* **21**, 241-251 (2021)

[10] Ding, X., Salam, Z. & Lv, W. Research on Timeliness Evaluation Model of Online Teaching Based on Intelligent Learning. *International Journal Of Continuing Engineering Education And Life-Long Learning.* **31**, 263-275 (2021)

[11] Wang Y. Ideological and political teaching model using fuzzy analytic hierarchy process based on machine learning and artificial intelligence. *Journal Of Intelligent & Fuzzy Systems: Applications In Engineering And Technology.* **40**, 3571-3583 (2021)

[12] Liang, H. Role of artificial intelligence algorithm for taekwondo teaching effect evaluation model[J]. *Journal Of Intelligent And Fuzzy Systems.* **40**, 3239-3250 (2021)

[13] Zeng, Q., Jiang, H., Liu, Q., Li, G. & Ning, Z. Design of a high-temperature grease by BP neural network and its preparation and high-temperature performance studies. *Industrial Lubrication And Tribology.* **74**, 564-571 (2022)

[14] Zhu, D., Cheng, C., Zhai, W., Li, Y. & Chen, B. Multiscale Spatial Polygonal Object Granularity Factor Matching Method Based on BPNN. *International Journal Of Geo-Information.* **10**, 75-91 (2021)

[15] Xue, Y., Wang, Y. & Liang, J. self-adaptive gradient descent search algorithm for fully-connected neural networks. *Neurocomputing.* **478**, 70-80 (2022)

[16] Bharathi, L. & Chandrabose, S. Machine Learning-Based Malware Software Detection Based on Adaptive Gradient Support Vector Regression. *International Journal Of Safety And Security Engineering: An Interdisciplinary Journal For Research And Applications.* **12**, 39-45 (2022)

[17] Chen, F., Liu, L., Tang, B. & Zhang, F. novel fusion approach of deep convolution neural network with auto-encoder and its application in planetary gearbox fault diagnosis[J]. *Journal Of Risk And Reliability.* **235**, 3-16 (2021)

[18] Wu, H., Liu, S., Cheng, C., Cao, S., Cui, Y. & Zhang, D. Multiscale variational autoencoder aided convolutional neural network for pose estimation of tunneling machine using a single monocular image. *IEEE Transactions On Industrial Informatics.* **18**, 5161-5170 (2021)

[19] Sweeti Attentional load classification in multiple object tracking task using optimized support vector machine classifier: a step towards cognitive brain–computer interface. *Journal Of Medical Engineering & Technology.* **46**, 69-77 (2022)

[20] Sun, F. & Shi, G. Study on the application of big data techniques for the third-party logistics using novel support vector machine algorithm. *Journal Of Enterprise Information Management.* **35**, 1168-1184 (2022)

Edited by: Kumar Abhishek
Special issue on: Machine Learning and Block-chain based Solution for Privacy and Access Control in IoT
Received: Oct 7, 2023
Accepted: Dec 19, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1478–1491, DOI 10.12694/scpe.v25i3.2675

BLOCKCHAIN ENHANCED STUDENT PHYSICAL PERFORMANCE ANALYSIS USING MACHINE LEARNING-IOT AND APRIORI ALGORITHM IN PHYSICAL EDUCATION NETWORK TEACHING

JIANING LI[*], ZHEPING QUAN[†], AND WEIJIA SONG[‡]

Abstract. In the digital era, particularly with the rise of online teaching, traditional approaches to college physical education face challenges in adequately monitoring and enhancing students' physical fitness. This study introduces a novel approach that integrates blockchain technology with a Machine Learning-IoT framework to evaluate and improve students' physical performance. Utilizing the Apriori algorithm, enhanced with particle swarm optimization and an improved K-means methodology, this system offers a robust tool for correlating student behavior with sports performance in a secure and decentralized manner. The proposed system uses blockchain for safe data management and IoT for real-time data collection, ensuring privacy as well as efficiency. The algorithm's accuracy, recall, and F1 values on the Iris dataset are 0.947, 0.931, and 0.928, respectively, with a considerable Calinski Harabasz score of more than 240. When applied to university student behavior data, the blockchain-enhanced system successfully mined association rules with a maximum confidence level of 0.923.

Key words: Apriori; Relevance; Clustering Algorithm; Network Teaching; Behavior Analysis; College Student

1. Introduction. The integration of sophisticated technologies such as the Internet of Things (IoT), Machine Learning (ML), and Blockchain has changed data analysis and decision-making processes in today's educational landscape. This is especially true in physical education, where the requirement to adequately evaluate and improve students' physical performance has become increasingly important. The introduction of online education practices has heightened the need for innovative ways that go beyond traditional limits.

With its interconnected network of physical devices, the Internet of Things provides an invaluable platform for real-time data collecting in educational contexts. This Internet of Things-based data gathering is critical for monitoring student actions and behaviors, especially in physical education, where performance indicators are dynamic and multidimensional. Machine Learning, on the other hand, provides powerful analytical capabilities for identifying patterns and insights in the massive amounts of data created by IoT devices. ML algorithms can help educators acquire a better understanding of student performance and develop more successful teaching tactics. By ensuring data integrity, security, and privacy, blockchain technology adds a key component to this ecosystem. Blockchain provides a decentralized and tamper-proof platform in educational situations where sensitive student data is involved, ensuring that the data utilized for analysis is reliable and safeguarded against unwanted access and manipulation.

In the context of educational reform in the new era, current college education emphasizes comprehensive development, in which physical education is an indispensable link. With the gradual improvement of information technology construction, the teaching of various subjects in universities is gradually networked [1]. Against this background, the current problem of less attention paid to physical education and the cultivation of students' physical fitness in college education is even more obvious [2]. Therefore, it is necessary to provide a practical evaluation tool for online teaching of physical education in universities. Apriori is the most classical and widely used association rule mining algorithm [3]. The algorithm uses a layer by layer iterative search method to mine hidden Boolean association rules between data [4]. K-means is one of the main tools for data analysis, which performs tasks such as data classification through clustering analysis [5]. This study introduces K-means into Apriori algorithm to construct a correlation algorithm between student behavior and sports performance to

[*] College of Physical Education, Taiyuan Normal University, Jinzhong, 030619, China (facejob2023@163.com)

[†] College of Physical Education, Taiyuan Normal University, Jinzhong, 030619, China (Corresponding author, zhepingquan@outlook.com)

[‡] Wesleyan College, Cavite State University, Manila, 0900, The Philippines (songweijia199472@163.com)

optimize the current situation of online sports teaching and help improve students' physical fitness. During the construction process, the K-means clustering results and the efficiency shortcomings of Apriori are also optimized. The purpose of this study is to bring practical results to the field of online physical education in universities. This research presents a comprehensive solution that synergizes IoT, ML, and Blockchain technologies to analyze and enhance student physical performance in an educational context. The main contributions of this research are:

1. We present a novel architecture that blends IoT's real-time data collection capabilities with ML's analytical prowess, all while being supported by blockchain technology's security and integrity. This comprehensive technique ensures a comprehensive examination of student physical performance while protecting data privacy and security.
2. We have adapted the Apriori algorithm, which has historically been employed in market basket analysis, for use in physical education. This modification, enhanced with particle swarm optimization and better K-means techniques, provides an effective tool for linking student behavior with sports success, giving instructors with actionable data.
3. Our research addresses the crucial concerns of data integrity and privacy in educational data analysis by utilizing blockchain technology. This ensures that the findings of the study are founded on secure and accurate data, which is critical in educational settings.

2. Related works. Apriori, as a major algorithm in the field of association mining, has been widely studied and applied so far. Hazelton J proposed a cost estimation model for engineering calculation based on Apriori association mining. Experimental results showed that the model could accurately estimate the cost of engineering construction [6]. Karthik S and Velu CM proposed a peak season sales forecasting system based on user data sets using Apriori algorithm instead of traditional algorithms [7].

According to the comparison results between this system and traditional algorithms, the accuracy, sensitivity, and specificity of this system reached 73%, 78%, and 80%, respectively, which were much higher than traditional algorithms. Moreover, the running time of this system was also shorter, so it had higher applicability. Sornalakshmi M et al. provided the healthcare industry with a new improved Apriori algorithm that can separate frequent itemsets and delete abnormal data at the beginning of the algorithm, thereby reducing the resource requirements for algorithm operation [8]. According to the test results, the results generated by this algorithm in the local minimum support degree of healthcare databases were relatively reliable, and can effectively reduce medical costs. K-means, as a widely used clustering algorithm, also had a lot of related research. Sun Z et al. proposed a short-term traffic flow prediction model based on K-means clustering and gated recursive units. This model can predict the traffic flow in the future based on the characteristics of historical traffic flow data [9] Compared with methods such as random forest and support vector machine regression, this method fully considered the pattern diversity of traffic flow, and had higher prediction accuracy. Kumar R U and Jeeva J B proposed a color cutting method using the clustering analysis function of K-means, which was used to evaluate and predict the healing of skin surface wounds [10]. This method was non-invasive and can effectively reduce the pain of patients, thereby reducing the tension and anxiety of patients facing medical facilities. Nyanjara S et al. with the help of experts to develop an assessment model for maternal and neonatal health quality in developing countries, which is based on K-means [11]. The model can classify and summarize the health data of the target population, and assign different data points to the most appropriate clustering. The test results showed that the classification accuracy of this method exceeded 73In the field of education, the teaching of physical education courses has always been one of the focuses of attention of educators. Huang Y Y studied the integration of online courses and physical education in current higher vocational education, and provided guiding suggestions for online physical education [12]. Guo J and Sun C proposed a real-time monitoring system for physical education classes based on the Internet of Things and cloud computing, which can automatically locate students' positions in sports scenes and evaluate the quality of their courses [13]. The system to some extent achieved the monitoring of students' performance in physical education courses, but failed to comprehensively evaluate students' physical exercise both inside and outside of class. Wang GR analyzed the current situation of physical education teaching in schools from a medical perspective [14]. According to the analysis conclusion, the current physical education teaching industry lacked teachers with professional medical knowledge, and the physical education teaching in various schools also lacked medical characteristics.

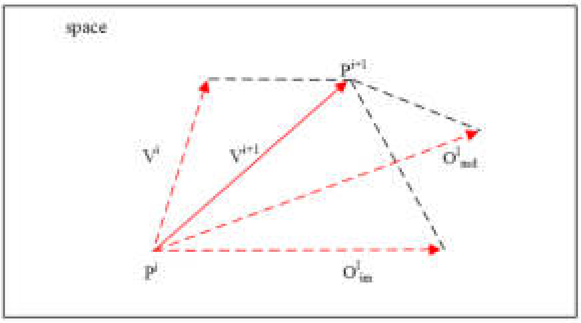


Fig. 3.1: Theory of Particle Swarm Optimization

The study believed that physical education combined with medical theory can reduce accidental injuries among students and further increase their exercise effectiveness.

Through sorting out relevant research results, it is found that most of the research on physical education teaching in universities focused on improving the quality of courses and the remoteness of the network. However, students' physical performance exists not only in physical education classes, but also in various aspects outside of class. Therefore, a comprehensive evaluation tool is necessary. Although Apriori and K-means are widely used, there is still a gap in intelligent sports teaching. Therefore, this study proposes a correlation algorithm between student behavior and sports performance based on Apriori.

3. Apriori Algorithm Construction for Correlativity between Student Behavior and Sports Performance.

3.1. Adaptive Weighted K-means Algorithm for Apriori Analysis. The behavior data of students is mainly continuous data, such as consumption data, network access data, etc. Due to the small dimensions and values of such data, K-means algorithm is suitable for clustering analysis. This is because K-means is a machine learning which has high efficiency and scalability, and low implementation costs [15]. However, the output effect of K-means is highly dependent on the selection of cluster centers. The selection of cluster centers is random, and the number of cluster centers is set by the operator through experience and test results, with a low degree of automation. In addition, K-means also has problems with local optimal solutions and unstable clustering results [16]. To address these shortcomings and make K-means more efficient in serving the processing of student behavior data, an optimized K-means based on Particle Swarm Optimization (PSO) is proposed which resembles like deep learning model. The principle of PSO is shown in Figure 3.1. In multi-dimensional space, each particle is searching for other individuals, and they constantly obtain the location of the next search, and continue to iterate until they find the optimal location. Introducing PSO into K-means can effectively optimize the randomness of cluster center selection, ensuring that K-means can find relatively excellent initial cluster centers. In addition, an adaptive weighting strategy is proposed to solve the problem of unstable clustering results.

Firstly, the PSO optimization part is constructed. In a search space with dimension M , the number of particles is N . The position of particle i can be expressed by equation 3.1.

$$p_{1m}, p_{2m}, \ldots, p_{im} \tag{3.1}$$

The P_{im} in equation 3.1 represents the position of particle i . The velocity of the particle is set to V, and the velocity vector of the particle is shown in equation 3.2.

$$V_{im} = (v_{i1}, v_{i2}, \ldots, v_{im}) \tag{3.2}$$

The position and velocity of particle is i . The optimal solution of its individual is V , as shown in equation 3.3.

$$\begin{cases} O_{im} & = (o_{i1}, o_{i2}, \ldots, o_{im}) \\ O_d & = (o_{1,d}, o_{2,d}, \ldots, o_{M,d}) \end{cases} \tag{3.3}$$

In equation 3.3, O_d represents the group optimal solution. After obtaining individual and group optimal solutions, the particle swarm can update its iteration speed and position. The mathematical expression of the velocity vector V_{im}^k of the updated particle in the mth dimension is shown in equation 3.4.

$$V_{im}^I = \omega v_{im}^{I-1} + (o_{im}^I - p_{im}^I)c_1 r_1 + (\mathbf{o}_{md}^I - p_{im}^I)c_2 r_2 \tag{3.4}$$

In equation 3.4, I represents the current number of iterations. ω is the inertia weight, and the weight value is proportional to the global optimization ability of the algorithm, while the local optimization ability is inversely proportional. c_1 and c_1 represent learning factors, the former being individual learning factors, and the latter being group learning factors. r_1 and r_2 are the random number used to increase the randomness of the search, and the values of both are [0,1]. o_{im}^I and o_{md}^I respectively identify the historical optimal positions of individuals and groups. The position vector P_{im}^I of the particle in mth dimension after updating is shown in equation 3.5.

$$p_{im}^I = p_{im}^{(I-1)} + v_{im}^{(I-1)} \tag{3.5}$$

After completing the construction of the particle swarm, you can introduce it into K-means. It is necessary to first calculate the particle fitness to obtain the global optimal solution. After inputting the optimal solution information and the data to be predicted, the reciprocal of the total distance is calculated from all sample points to the corresponding cluster center, as shown in equation 3.6.

$$\mathrm{FIT} = \sum_{k=1}^{K} \sum_{p_i \in P_K} d(p_i, C_k) \tag{3.6}$$

In equation 3.6, FIT represents the fitness function value, and $d(p_i, C_k)$ is the distance from the sample point to its corresponding cluster center. K refers to the k-value of the algorithm. The role of particle swarm optimization in this algorithm is to optimize the selection of clustering centers. In addition, adaptive weights should be used to optimize the performance instability of K-means. Assuming that there are n data to be processed with a dimension of m , the data can be converted into a matrix as shown in equation 3.7.

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \tag{3.7}$$

If the number of clusters is the same as the K value of the algorithm, the total intra class distance of all clusters in the algorithm on the uth dimension is shown in equation 3.8.

$$D_{in} = \sum_{k=1}^{K} \sum_{j=1}^{n} (x_{ju} - m_{ku})^2 \tag{3.8}$$

In equation 3.8, m_{ku} is the mean value of the k-th cluster on the attributes of the th dimension. After obtaining this value, you can further obtain the sum of the inter class distances of all clusters on this dimension attribute, as shown in equation 3.9.

$$D_{out} = \sum_{k=1}^{K} (m_{ku} - m_u)^2 \tag{3.9}$$

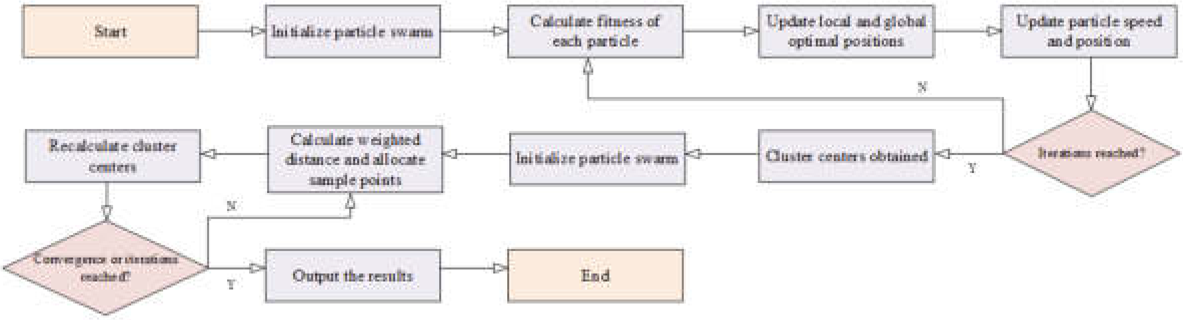

Start
Initialize particle swarm
Calculate fitness of each particle
Update local and global optimal positions
Update particle speed and position
Iterations reached?
N
Y
Cluster centers obtained
Initialize particle swarm
Calculate weighted distance and allocate sample points
Recalculate cluster centers
Convergence or iterations reached?
N
Y
Output the results
End

Fig. 3.2: The Improved K-means Algorithm for Student Behavior Data Processing

From this, it is possible to calculate the impact of the attributes of the uth dimension on the clustering results, as shown in equation 3.10.

$$\text{impact}_u = \frac{D_{in}}{D_{out}} \tag{3.10}$$

When a sample in a space has compact intra cluster distances and large inter cluster distances, it indicates that the clustering results are good, further indicating that the attribute dimension has a significant impact on the clustering results. Assigning weights to attributes with different dimensions and giving greater weights to more important attributes can optimize the clustering results of the entire algorithm. Set the attribute weight of dimension u to W , as shown in equation 3.11.

$$W_u = \frac{\text{impact}_u}{\sum_{u=1}^{M} \text{impact}_u} \tag{3.11}$$

In equation 3.11, the value of each W is between 0 and 1, and the sum of W for each dimension is 1. Adding the calculated attribute weights to the traditional K-means Euclidean distance equation can obtain a new distance equation, as shown in equation 3.12.

$$D(a,b) = \sqrt{\sum_{u=1}^{M} W_u (x_{au} - x_{bu})} \tag{3.12}$$

The flow of the improved K-means algorithm for student behavior data processing is shown in Figure 3.2. Before performing the K-means operation, the number of cluster centers is first obtained through PSO, and then the weights are initialized. In this algorithm, sample points are calculated based on weighted Euclidean distances, and then K-means iteration is performed. Compared with traditional K-means, this algorithm solves the problems of cluster center selection and sample calculation stability through PSO and weighted distance.

Apriori is one of the most important algorithms in the field of data mining, and its specific process is shown in Figure 3.3. Based on the preset support and confidence levels, the algorithm will determine the threshold value of the candidate item set and decide to prune or access the next step [17]. When the processed data is greater than or equal to the minimum confidence and support, it is considered that there is correlation between the data, and the algorithm can derive its association rules [18]. However, Apriori running according to this process has two main drawbacks. Firstly, the algorithm will generate a large number of candidate itemsets, a large portion of which are unnecessary for computation, which has a negative impact on the efficiency of the algorithm and increases the computational cost [19]. Secondly, Apriori scans the transaction database too much during the calculation process, which leads to a significant increase in the algorithm's calculation time, seriously affecting the efficiency of the algorithm [20].

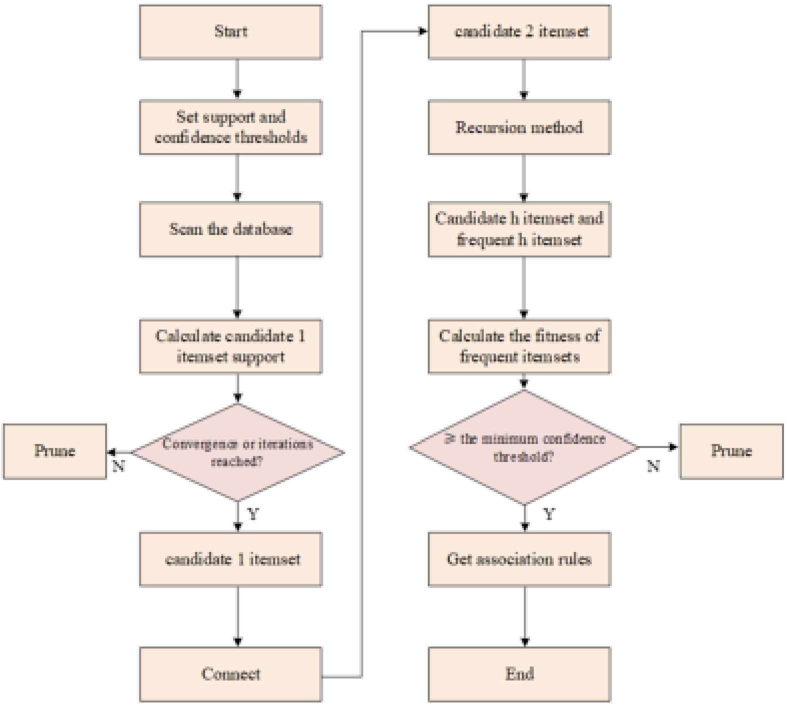

Start
Set support and confidence thresholds
Scan the database
Calculate candidate 1 itemset support
Convergence or iterations reached?
N
Prune
Y
candidate 1 itemset
Connect
candidate 2 itemset
Recursion method
Candidate h itemset and frequent h itemset
Calculate the fitness of frequent itemsets
≥ the minimum confidence threshold?
N
Prune
Y
Get association rules
End

Fig. 3.3: Structure of traditional Apriori algorithm

The analysis of student behavior data usually involves the entire school's student data, and the amount of data is relatively large, so there are high requirements for algorithm efficiency. To improve the efficiency of Apriori, an optimization method based on pre pruning is proposed to address its main shortcomings. Pruning is an operation of Apriori to filter itemsets that do not meet the support and confidence levels, and must be run at least twice in a round of algorithm runs. Therefore, performing pre pruning during the operation of the algorithm can reduce the number of candidate itemsets and the number of times the algorithm scans the transaction database, effectively improving the two main defects of Apriori. There is a set of frequent items, and data appears times in its subset, equation 3.13 can be obtained.

$$|L_{h-1}(z)| \geq h-1 \tag{3.13}$$

In equation 3.13, $L_{h-1}(z)$ refers to the number of occurrences of in frequent itemset L_{h-1}. According to this equation, the conditions for judging the h term set Y as an infrequent term set are shown in equation 3.14.

$$\begin{cases} |L_{h-1}(z)| < h-1 \\ z \in Y \end{cases} \tag{3.14}$$

When Y is a non frequent itemset, determine whether the $h$itemset connected to the Y_j existing in itemset Y is a non frequent itemset, and the judgment conditions are shown in equation 3.15.

$$\begin{cases} |L_{h-1}(z)| < h-1 \\ \exists z \in Y_j \end{cases} \tag{3.15}$$

When equation 3.15 is satisfied, Y_j does not need to add connections, and it is pruned during the operation of the algorithm, thereby increasing the efficiency of the algorithm. The improved Apriori algorithm flow is

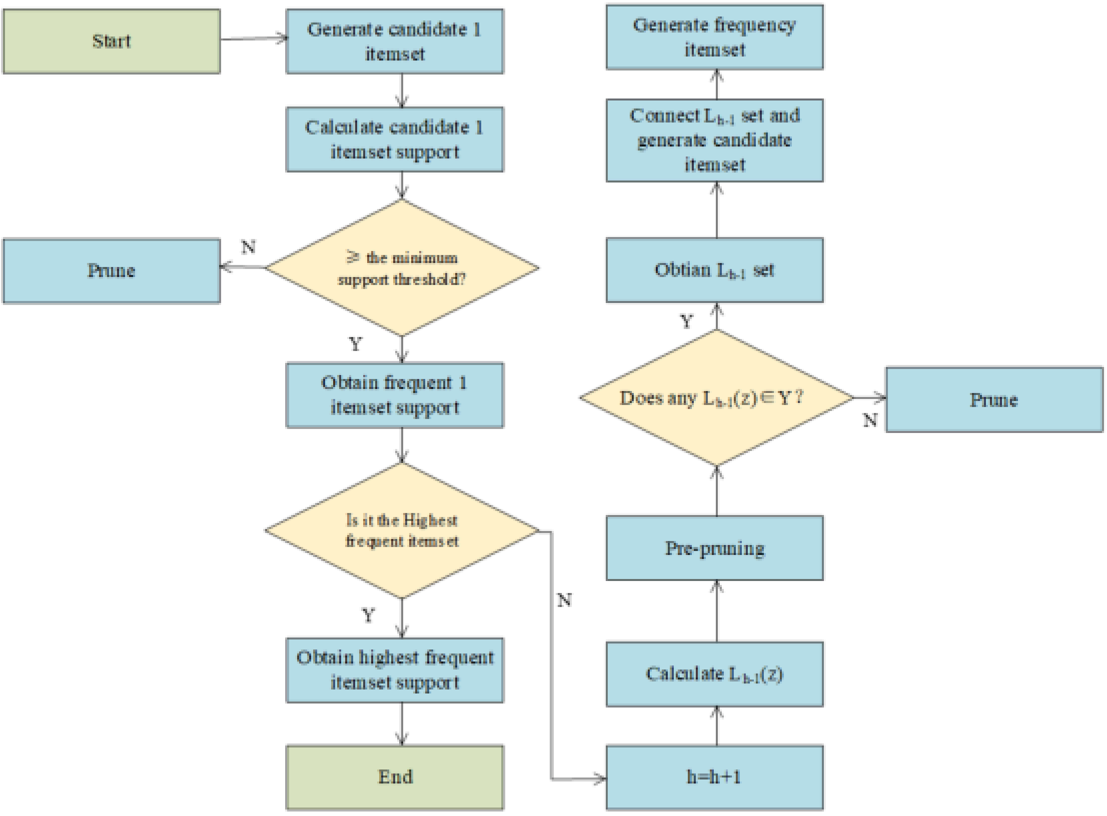

Start
Generate candidate 1 itemset
Calculate candidate 1 itemset support
≥ the minimum support threshold?
N
Prune
Y
Obtain frequent 1 itemset support
Is it the Highest frequent itemset
Y
N
Obtain highest frequent itemset support
End
h=h+1
Calculate L_{h-1}(z)
Pre-pruning
Does any L_{h-1}(z)∈Y?
N
Prune
Y
Obtian L_{h-1} set
Connect L_{h-1} set and generate candidate itemset
Generate frequency itemset

Fig. 3.4: Improved Apriori Operation Flow

shown in Figure 3.4. The data processed by improved K-means clustering is input into the algorithm and a candidate item set is generated. During the operation of the algorithm, some steps, including calculating confidence levels, are replaced with pre pruning steps, thereby reducing the amount of itemset generation and scanning times, and increasing operational efficiency.

After completing the construction of the correlation algorithm, specific settings need to be made for student data processing and correlation analysis. First, the characteristics and corresponding labels of different types of students are designed based on the type of student behavior data, as shown in Table 3.1. Student behavior is divided into three categories, including consumer behavior, life behavior, and sports behavior. The information included in consumption behavior mainly includes total sales volume, single consumption fluctuation, consumption volume for three meals, consumption frequency, and additional sales volume not necessary for daily life. Consumer behavior can reflect students' living habits and daily activities to a certain extent, and is an important evaluation dimension. The information included in life behavior mainly includes the time of students' meals, time of getting up and sleeping, length of sleeping, and time of surfing the internet. Life information reflects whether students' lives are regular, healthy, and have time for exercise. Healthy living habits are the foundation of excellent sports performance and physical fitness, so this is a major evaluation dimension. The classification of students' living habits is mainly based on whether their life data are regular. The last type of behavior is sports behavior, which mainly includes information about the number and time of entering the stadium and gymnasium, as well as the length of each exercise, and the use of sports equipment. This dimension of information can most directly reflect the students' sports situation.

After defining the behavioral characteristics and corresponding labels of students, it is also necessary to define their physical performance. The definition of physical education performance is based on the students'

Table 3.1: Indicator Labels for Association Rule

Categories	Features	Tags
Consumer behavior	Low living expenses, stable and regular consumption, basically no extra expenses	Low expenditure type
	The cost is too high, the single consumption fluctuates greatly, the takeout consumption frequency is high, and the extra cost is high	High expenditure type
	Low living expenses, stable and regular consumption, low extra expenses	Frugal type
	Moderate spending, moderate consumption frequency and regular consumption	Moderate type
	Low consumption frequency, low cost of meals, and high extra cost	In campus low expenditure type
Daily life behavior	Irregular meals, dependence on takeout, basically not getting up early, long internet time	Irregular type
	Regular meals, occasional takeout, frequent early getting up, short internet time	very regular type
	Meals are irregular, the number of takeouts is small, the early getting up is not frequent, and the internet time is moderate	regular type
	Irregular meals, seldom get up early, often order takeout, and spend a long time online	less regular type
Sports bahavior	The equipment is highly used, the number of times of admission is many, and the exercise time is long	Very effort type
	Average use of equipment, many times of admission and long exercise time	effort type
	The use of equipment is small, the number of admission is average, and the exercise time is short	less effort type
	Less equipment use, less admission times and less exercise time	effortless type

annual physical examination performance and the results of each semester's physical education courses. Sports performance is classified using a comprehensive scoring system. The score is calculated based on the weighted average score on which the definition of physical performance is based. The scoring range is 1 to 5 points, with 4-5 points representing outstanding physical performance, 3-4 points representing less outstanding physical performance, 2-3 points representing average physical performance, and 1-2 points representing Bad physical performance. When a student's physical performance is 1-2 points, their physical performance is extremely poor and they fail in terms of physical fitness.

4. Apriori Algorithm Experiment and Application for Correlating Student Behavior and Sports Performance.

4.1. Correlation Algorithm Performance Test. To evaluate the value of the proposed algorithm, it is necessary to test the algorithm. The test consisted of two parts, namely, algorithm performance test and practical application test. Performance testing tested the computational capabilities and characteristics of the improved K-means and Apriori algorithms in the algorithm. Due to the significant impact of hardware and software environments on algorithm performance, this experiment was conducted in a fixed environment configuration. The detailed configuration is shown in Table 4.1. The experimental operating system was Windows 10, and the programming environment was Python. The data used for the test was from the UCI database, where three datasets, Iris, Glass, and Wine, were used to test the performance of the algorithm.

Firstly, the clustering accuracy of the proposed algorithm was tested. Because the clustering of the algo-

Table 4.1: Configuration Related to Performance Test

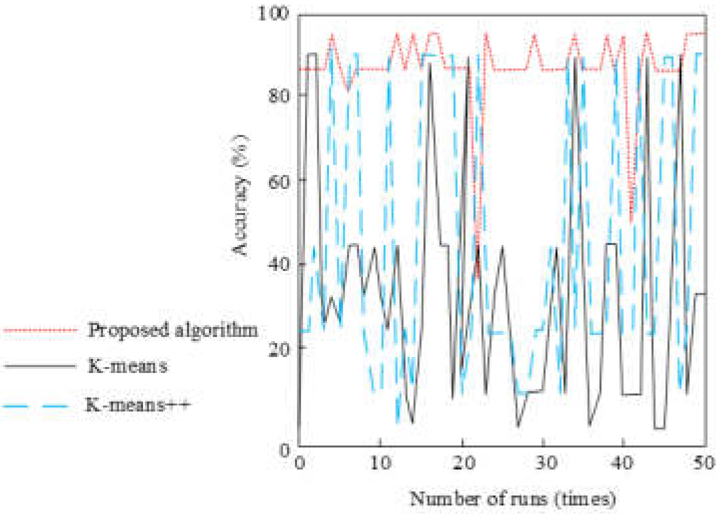
CPU	Intel(R) i7-8700k	
RAM	16GB	
Operating system	Windows 10	
Hard disk	1 TB HDD	
Programming environment	Python	
Experimental database	UCI	

Table 4.2: Assessment Data of the Proposed Algorithm and Comparative Algorithms

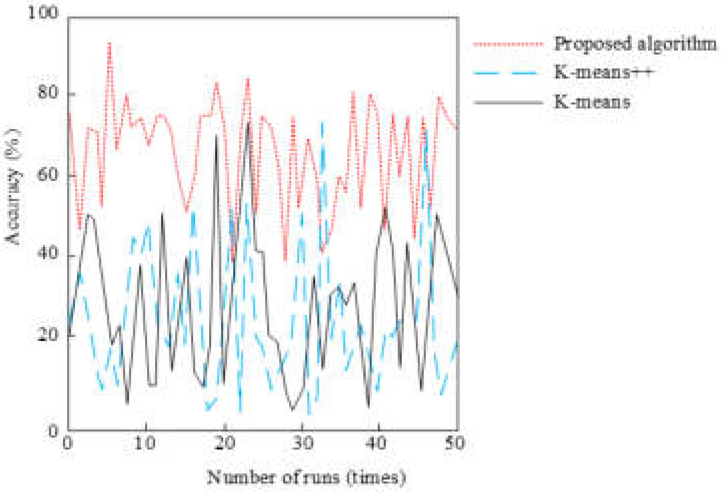
Data sets	Algorithm	Precision	Recall	F1
Iris	K-means	0.907	0.874	0.892
	K-means++	0.907	0.893	0.895
	Proposed algorithm	0.947	0.931	0.928
Wine	K-means	0.531	0.455	0.470
	K-means++	0.548	0.443	0.481
	Proposed algorithm	0.594	0.497	0.509
Glass	K-means	0.135	0.197	0.127
	K-means++	0.138	0.158	0.159
	Proposed algorithm	0.342	0.313	0.233

rithm was performed by an improved K-means algorithm, the traditional K-means algorithm and the improved K-means++algorithm based on K-means were used here as a comparison algorithm. The results of the accuracy test are shown in Figure 4.1. Iris and Wine datasets were used to test the accuracy of centralized algorithms under different data sets. The three algorithms were run 50 times on each dataset. Figure 4.1a shows the performance of the three algorithms in the Iris dataset, and Figure 4.1b shows the performance of the three algorithms in the Wine dataset. Under the two datasets, the accuracy rates of the three algorithms showed significant fluctuations, which was caused by the characteristics of the K-means algorithm itself. Observing the graph, the proposed algorithm exhibited minimal fluctuations in both data sets, and its accuracy curve was mostly higher than K-means algorithm and K-means++algorithm. In the Iris dataset, the accuracy rate of the proposed algorithm was above 80% for most of the time, and both comparison algorithms were below this level. Precision, recall, and F1 values are also important indicators for evaluating algorithm performance. After measuring the average data of 50 runs of the three algorithms, their above indicators were compared, as shown in Table 4.2. The performance of K-means algorithm and K-means++ were relatively close, and the two algorithms had advantages and disadvantages for each other under different data sets. The precision, recall, and F1 of the proposed algorithm were steadily higher than those of the other two algorithms under the three data sets. Its average precision, recall, and F1 values under the Iris dataset were 0.947, 0.931, and 0.928, respectively. In the same dataset, the recall rate and F1 values of K-means were 0.874 and 0.892, respectively, for K-means++, which are 0.893 and 0.895. Combining the accuracy test results, the proposed algorithm outperformed K-means and K-means++algorithms in clustering ability, which laid a solid data foundation for later correlation analysis.

For cluster analysis, the experiment also used its unique Silhouette Coefficients (SC) and Calinski Harabasz Score (CH) for evaluation. SC showed the differences within and outside the K-means cluster, and larger differences represented better clustering results. CH is an indicator obtained by calculating the ratio of inter cluster variance to intra cluster variance under certain conditions. The higher the value of this indicator, the better the clustering effect. The results of SC and CH evaluations are shown in Figure 4.2. Figure 4.2a shows the SC test results of the algorithm under different data sets, and Figure 4.2b shows the CH test results of the algorithm under different data sets. There are significant differences in the performance of each algorithm within different data sets depending on the different data sets. Observing within the same dataset, the SC



(a) Accuracy Under Iris

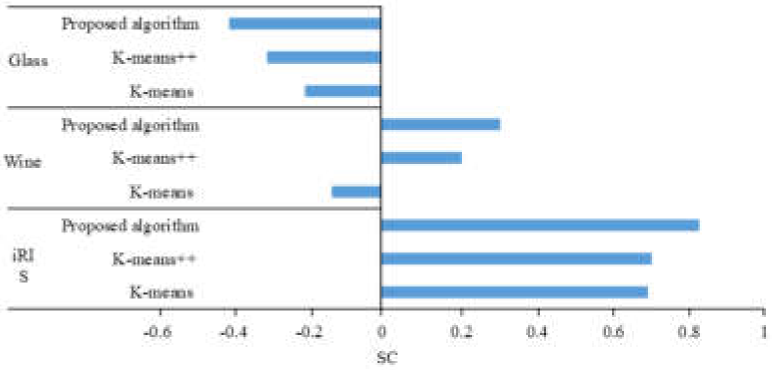


(b) Accuracy Under Wine

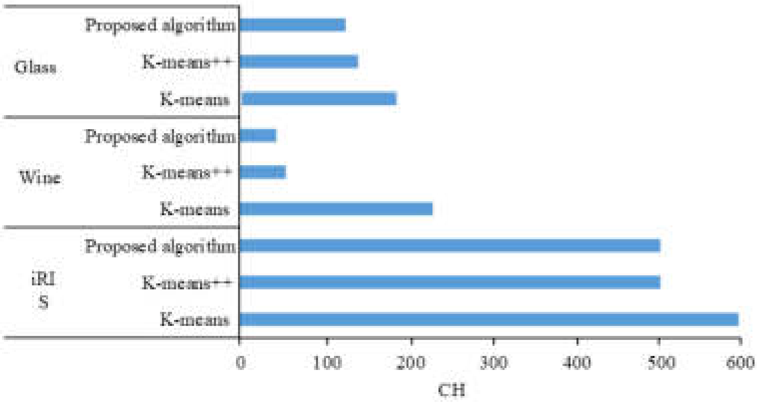
Fig. 4.1: Results of Accuracy Test under Different Data Set

and CH of the proposed algorithm were both higher than K-means and K-means++. Under the Wine dataset, the CH of the proposed algorithm was 240.3, while the CH of K-means and K-means++ were 63.4 and 67.2, respectively. Based on the evaluation results of SC and CH, the proposed algorithm had better clustering ability in general.

Due to the large amount of data on students' behavior data and sports performance data, the speed of algorithm operation was extremely important on the premise of ensuring correctness. The proposed algorithm mainly reduced the amount of item set processing and scanning times during Apriori's operation through pre pruning, thereby reducing the burden and cost of correlation calculation and increasing efficiency. To verify the



(a) SC Under Different Dataset



(b) CH Under Different Dataset

Fig. 4.2: SC and CH Evaluation Results of Different Algorithms

effectiveness of this optimization, the traditional Apriori, Dual Searching Apriori, and Pre Apriori proposed in this study were combined with K-means and the PSO-K-means proposed in this study. Their operational efficiency against the same dataset was tested. The results are shown in Figure 4.3. The curve of the proposed algorithm was below the other comparison algorithms, indicating that the proposed algorithm had the fastest running speed. When the support threshold was 0.4, the computation time was about 3400ms, significantly lower than other algorithms.

4.2. Correlation Algorithm Practical Application Test. After verifying the performance of the algorithm in the test dataset, it was also necessary to test it under a real student behavior dataset to verify its practical application value. A university was selected as the main venue for testing. The university has a high degree of intelligence in student management. Through the campus network management system, it was possible to query students' consumption on campus, the time to enter and leave the school and gym, the online duration of the campus network, dormitory water consumption, and light off time. This provided conditions for testing the correlation between student behavior and physical performance. After negotiation, the behavioral data and sports performance of 1540 students in a certain grade of this size were selected as data samples, and the samples were sent to the proposed algorithm for clustering analysis and mining related rules. According to the previous test results, the minimum support threshold of the algorithm was set to 0.2, and the minimum confidence threshold was set to 0.4. After mining, a total of 37 relevant rules were mined, as

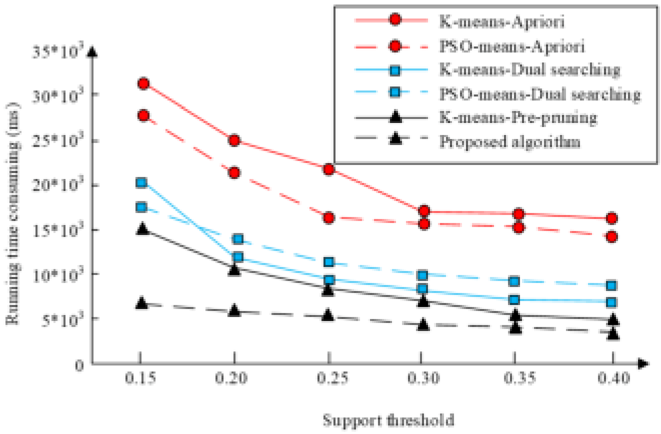


Fig. 4.3: Comparison Results of Running Time of Algorithms

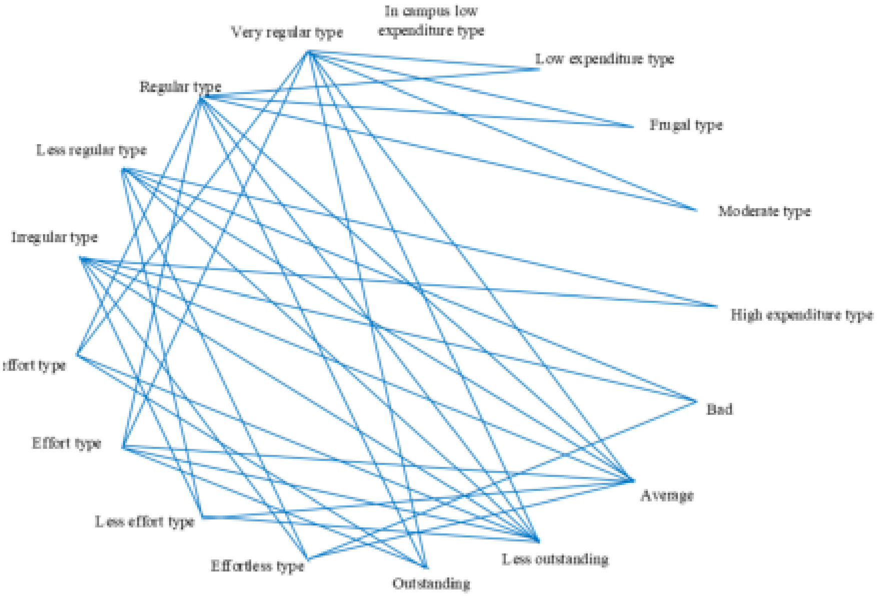


Fig. 4.4: Display of Association Rule Mining Results

shown in Figure 4.4. Through the images, there were more association rules between students' life behavior and sports behavior and their sports performance, while there were fewer association rules between consumer behavior and sports performance.

Table 4.3 presents the results of association rule mining in a data format, with support and confidence levels indicated after each rule. Two association rules treated sports as outstanding and less outstanding as post rules. Five association rules used average as a post rule for sports performance. Three association rules treated sports as bad as a post rule. Among the association rules mined, the highest support level was 0.550, and the lowest was 0.212. The highest confidence level was 0.923 and the lowest was 0.727. From the mining results of association rules, sports performance was related to life rules and sports habits. Students who had a regular, active lifestyle, and long-term exercise habits were more likely to have excellent physical performance.

Table 4.3: Association Rule Mining Results

Rule	Support	Confidence
Very regular type, very effort type: outstanding	0.451	0.862
Regular type, very effort type: outstanding	0.422	0.841
Very regular type, effort type: less outstanding	0.520	0.736
Regular type, effort type: less outstanding	0.513	0.727
Very regular type, less effort type: average	0.345	0.747
Less regular type, less effort type: average	0.424	0.912
Less regular type, less effort type, moderate type: average	0.235	0.785
Low expenditure type, less effort type: average	0.212	0.772
Frugal type, very regular type, less effort type: average	0.434	0.783
Effortless type: bad	0.546	0.923
Irregular type: bad	0.550	0.895
Effortless type, irregular type: bad	0.535	0.883

5. Conclusion. To improve the physical quality of contemporary college students and promote their comprehensive development, a mining algorithm based on Apriori for the association between college students' behavior data and sports performance is proposed to address the current lack of physical performance evaluation methods and effectiveness. The algorithm combines an improved K-means clustering and Apriori correlation algorithm. According to the experimental results, the average accuracy, recall, and F1 values of the proposed algorithm in the Iris dataset were 0.947, 0.931, and 0.928, respectively. The indicators of other algorithms in the same dataset were lower than those of the proposed algorithm. Under the Wine dataset, the CH of the proposed algorithm was 240.3, while the CH of K-means and K-means++ did not exceed 100. When the support threshold was 0.4, the computation time of the proposed algorithm was about 3400ms, significantly lower than other algorithms. In practical applications, the proposed algorithm had mined 12 association rules, with a maximum support level of 0.550 and a minimum confidence level of 0.212. The maximum confidence level was 0.923 and the minimum confidence level was 0.727. The results of association rule mining provided reference information based on quantitative data for the formulation of college physical training measures and the improvement of students' physical fitness. Although the proposed algorithm has already been practical, there is still room for improvement. The current construction of student behavior data system is not perfect. In further research, more important behavior data will be screened through controlled variable experiments, and less important behavior data will be eliminated.

Funding. The research is supported by The humanities and social science planning project of the Ministry of Education, "Research on the security mechanism of deepening the integration of sports and education to promote the implementation of the adolescent physical health policy", subject number: 21YJA890024..

REFERENCES

[1] Qi, S., Li, S. & Zhang, J. Designing a Teaching Assistant System for Physical Education Using Web Technology[J]. *Mobile Information Systems (*. **6** pp. 1-11 (2021)

[2] Wang, Z. Investigation on the Employment of Ex-Soldiers in Physical Education Major in Colleges and Universities [J]. *Open Access Library Journal.* **8**, 1-8 (2021)

[3] Zhang, Z., Liu, X., Li, Z. & Hu H. Outburst prediction and influencing factors analysis based on Boruta-Apriori and BO-SVM algorithms[J]. *Journal Of Intelligent And Fuzzy Systems.* **2021**, 1-18 (0)

[4] Aria, R. & Susilowati, S. Analisa Data Penjualan SaRa Collection menggunakan metode Apriori[J]. *Jurnal Teknik Komputer.* **7**, 68-73 (2021)

[5] Agalya, V., Kandasamy, M., Venugopal, E. & Others CPRO: Competitive Poor and Rich Optimizer-Enabled Deep Learning Model and Holoentropy Weighted-Power K-Means Clustering for Brain Tumor Classification Using MRI[J]. *International Journal Of Pattern Recognition And Artificial Intelligence.* **36**, 1-30 (2022)

[6] J., H. THE AUTOMATING OF COST ESTIMATES: APRIORI AND ITS SOFTWARE[J]. *F & M: Fabricating & Metalworking.* **21**, 44-471 (2022)

[7] Karthik, S. & Velu, C. Novel prediction of sales and purchase forecasting for festival season of Hypermarkets with customer dataset using Apriori algorithm instead of FP-Growth algorithm to improve the accuracy[J]. *ECS Transactions.* **107**, 12647-12659 (2022)

[8] Sornalakshmi, M., Balamurali, S., Venkatesulu, M., Krishnan, M., Ramasamy, L., Kadry, S. & Lim, S. An efficient apriori algorithm for frequent pattern mining using mapreduce in healthcare data[J]. *Bulletin Of Electrical Engineering And Informatics.* **10**, 390-403 (2021)

[9] Sun, Z., Hu, Y., Li, W., Feng, S. & Pei, L. Prediction model for short-term traffic flow based on a K-means-gated recurrent unit combination[J]. *IET Intelligent Transport Systems.* **16**, 675-690 (2022)

[10] Kumar, R. & Jeeva, J. Segmentation of Wound By K-Means Clustering and Automatic Prediction of Healing Time[J]. *ECS Transactions.* **107**, 20371-20376 (2022)

[11] Nyanjara, S., Machuve, D. & Nykanen P. Maternal and Child Health Care Quality Assessment An Improved Approach Using K-Means Clustering[J]. *Journal Of Data Analysis And Information Processing.* **10**, 170-183 (2022)

[12] Huang, Y. Integrating Online Teaching into Public Physical Education - Taking Vocational Colleges in Chongqing as Examples[J]. *JOURNAL OF CONTEMPORARY EDUCATIONAL RESEARCH.* **6**, 102-107 (2022)

[13] Guo, J. & Sun, C. Real-time monitoring of physical education classroom in colleges and universities based on open IoT and cloud computing[J]. *Journal Of Intelligent & Fuzzy Systems: Applications In Engineering And Technology.* **40**, 7397-7409 (2021)

[14] Wang, G. Exploration of Teaching and Education: Integration of Sports and Medicine in Physical Education in Colleges and Universities[J]. *JOURNAL OF CONTEMPORARY EDUCATIONAL RESEARCH.* **6**, 63-67 (2022)

[15] Warchalska-Troll, A. & Warchalski, T. The selection of areas for case study research in socio-economic geography with the application of k-means clustering[J]. *Wiadomosci Statystyczne.* **67**, 1-20 (2022)

[16] Majdina, N., Soeleman, M. & Supriyanto, C. Application of Particle Swarm Optimization (PSO) to Improve K-means Accuracy in Clustering Eligible Province to Receive Fish Seed Assistance in Java[J]. *IOSR Journal Of Computer Engineering.* **24**, 43-49 (2022)

[17] Qisman, M., Rosadi, R. & Abdullah, A. Market basket analysis using apriori algorithm to find consumer patterns in buying goods through transaction data (case study of Mizan computer retail stores) [J]. *Journal Of Physics: Conference Series.* **1722** pp. 1 (2021)

[18] Javed, M., Nawaz, W. & Hova-fppm, K. Flexible Periodic Pattern Mining in Time Series Databases Using Hashed Occurrence Vectors and Apriori Approach[J]. *Scientific Programming (.* **1** pp. 1-14 (2021)

[19] Muchlis, M., Fitri, I. & Nuraini, R. Rancang Bangun Aplikasi Data Mining pada Penjualan Distro Bloods Berbasis Web menggunakan Algoritma Apriori[J]. *Jurnal JTIK (Jurnal Teknologi Informasi Dan Komunikasi).* **4**, 26-33 (2021)

[20] Saha, S., Prasad, A., Chatterjee, P., Basu, S. & Nasipuri M. Modified FPred-Apriori improving function prediction of target proteins from essential neighbours by finding their association with relevant functional groups using Apriori algorithm[J]. *International Journal Of Advanced Intelligence Paradigms.* **19**, 61-83 (2021)

Edited by: Kumar Abhishek
Special issue on: Machine Learning and Block-chain based Solution for Privacy and Access Control in IoT
Received: Oct 7, 2023
Accepted: Dec 19, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1491–1498, DOI 10.12694/scpe.v25i3.2660

DYNAMIC SCHEDULING OF MULTI-AGENT ELECTROMECHANICAL PRODUCTION LINES BASED ON ITERATIVE ALGORITHMS

LULU YUAN*

Abstract. In response to the optimization scheduling problem in the dyeing production process, the author proposes a hierarchical scheduling method for dyeing vats based on genetic algorithm and multi-agent. In this method, a hierarchical scheduling algorithm is used to decompose production scheduling into static and dynamic strategies. The static strategy adopts a genetic algorithm that supports batch processing of multiple products, non equality of equipment, order delivery time, switching cost, and other constraints: Dynamic strategy is a coordinated dynamic optimization algorithm that uses multi-agent systems to support the running status of dye tanks based on static strategies. By solving the algorithm with multiple constraints and dynamic factors in the production process, the final result of the dyeing tank operation task is obtained. The simulation compared pure genetic algorithm with manual scheduling, and the results showed that the hierarchical dynamic scheduling strategy based on data-driven achieved the goal of optimizing the production scheduling of dyeing vats. The practical application results also demonstrate the feasibility of this method.

Key words: Multi-agent, Genetic algorithm, hierarchical scheduling, Agent

1. Introduction. The actual workshop production system is a dynamic production environment, and any changes in production plans, processing equipment, scheduling objectives, and other factors will cause changes in production scheduling[1]. In order to achieve coordinated operation of the workshop and achieve global optimization goals, in order to better complete workshop scheduling tasks, achieve reasonable resource allocation, and use ant colony algorithm to construct a dynamic scheduling algorithm for the workshop. The original meaning of Agent is ”agent”, and A agent is a computing entity with a problem reasoning and solving mechanism that can play its role autonomously. The reason why a multi-agent structure should be applied is because a multi-agent system is a network composed of multiple agents that can coordinate operations. Each agent in the computing system has its own different problem-solving methods and methods[2]. However, agents can dynamically schedule workshops through agreed unified communication protocols and make decisions through bidding, negotiation, and other means.Adopting a multi-agent structure can prevent the dynamic scheduling system structure of the workshop based on multi-agent in System Figure 1 from collapsing due to errors in a certain part of the system, which is beneficial for improving the stability of the system. At the same time, achieving distributed decision-making in the manufacturing system makes the system highly robust and scalable. The basic working principle is as follows: Firstly, the data collection agent plays a role in real-time production monitoring, collecting on-site data during the production process, such as machine tool status, route operation, etc., and providing raw data for the evaluation and decision-making agent in conjunction with the production plan. The integration agent combines various data sent by various data collection agents in an orderly manner and integrates them. Through the preprocessing algorithm in the agent, each data is converted into a unified format that the system can recognize, which is conducive to improving the computational performance of the entire system[3]. The main control agent not only needs to provide various data and evaluation and analysis algorithms available to the decision-making agent, such as correlation algorithms, mean value algorithms, etc., but also is responsible for real-time monitoring of the entire system's operation.If there is an error in the data, instructions can be issued to request the data collection agent to change the frequency of the data provided, thereby achieving global optimization goals based on monitoring. The evaluation and decision-making agent is the core of the entire system and the main executor of scheduling

*College of Mechanical and Electrical Automation, Henan Polytechnic Institute, Nanyang, Henan, 473000, China (Corresponding Author)

tasks. It can request the main control agent to provide real-time data and select appropriate algorithms for specific resource allocation in the workshop. Due to the role of the evaluation and decision-making agent as a workshop scheduling agent, it can make decisions through negotiation, bidding, and other means, and timely send the decision results to the appropriate processing unit agent. When problems occur during scheduling (such as the waiting time for a route to run is too long or rework phenomenon occurs), a processing command can be sent to the general control agent to make the system temporarily stop scheduling and processing, and error information can be timely called out from the shared knowledge base for analysis and resolution by the decision-maker.When the processing task to be processed is too large, the task can be decomposed and distributed by multiple evaluation and decision agents. Finally, the processing information is transformed into a unified decision result through preprocessing algorithms and published to the appropriate processing unit agent. With the continuous increase of decision-making events, the data information in the shared knowledge base will develop towards a direction that is more conducive to dynamic workshop scheduling[4,5]. The processing unit agent can continuously issue instructions to the evaluation and decision-making agent to receive processing tasks. If there is a problem with a certain machine tool during the processing, the equipment monitoring agent will determine that all processing routes passing through the machine tool are unavailable and promptly publish the information to the decision-making agent, so that the decision-making agent can re evaluate and make decisions on the real-time status in the workshop, and select a feasible path from the remaining feasible processing paths to complete the processing task, when the malfunction of the machine tool is eliminated, all processing routes passing through the machine tool will be redefined as available.

The production process of modern dyeing enterprises is mostly carried out in a small batch and multi variety manner. In this production method, achieving reasonable arrangement of production plans and tasks is a complex production scheduling process, which is difficult to solve with a scheduling strategy[6]. It often requires the use of composite scheduling strategies at different stages and conditions. There are currently many literature on this topic, and combining genetic algorithm with multi-agent is a new method that has emerged in recent years to solve complex scheduling problems. The scheduling methods for dyeing production workshops described in existing literature generally only consider delivery time and switching costs when designing models, and rarely design dynamic scheduling under complex and variable production conditions in the workshop. Based on real-time sample data collected by dyeing enterprises, the author studies a genetic algorithm for static scheduling that supports batch processing characteristics of multiple products, non equality of multiple dyeing tanks, pre order backlog, and order delivery time constraints, and research on dynamic optimization algorithms for coordination among multi-agent systems based on the running status of workshop dyeing tanks. By solving multiple constraints and dynamic factors in the production process in a hierarchical manner, a dynamic optimization strategy for dyeing tank operation tasks is obtained, achieving the goal of optimizing dyeing tank operation scheduling.

2. Methods.

2.1. Problem Description. There are M dyeing vats in the workshop, each of which can process any type of product, and the capacity of each dyeing vat varies[7,8]. The minimum and maximum production capacity of each dyeing vat are constrained by the bath ratio. The number of orders that require production and processing is N, and an order may require processing and production of one or more products. The quantity and delivery time required for each product may vary, and the production process may also vary.The time and cost of processing different product types in dyeing vats also vary. Dyeing processing also requires consideration of switching costs. Taking into account the above constraints and minimizing the total production cost under the conditions of meeting the delivery dates of each order as much as possible, the scheduling results are often not optimal due to the lack of consideration of real-time production changes. Therefore, it is necessary to find a dynamic optimization method based on this to achieve the optimization of dyeing tank scheduling.

2.2. Static scheduling model for dyeing vats based on genetic algorithm. In addition to the delivery time, time, and spatial constraints of general production processes, the dyeing production process also has industry characteristics such as the variability of processing equipment, switching costs, and additional resource consumption, which are complex nonlinear, stochastic, and uncertain, this makes it impossible for the dyeing scheduling model to copy the scheduling models in existing literature, and it is necessary to establish a

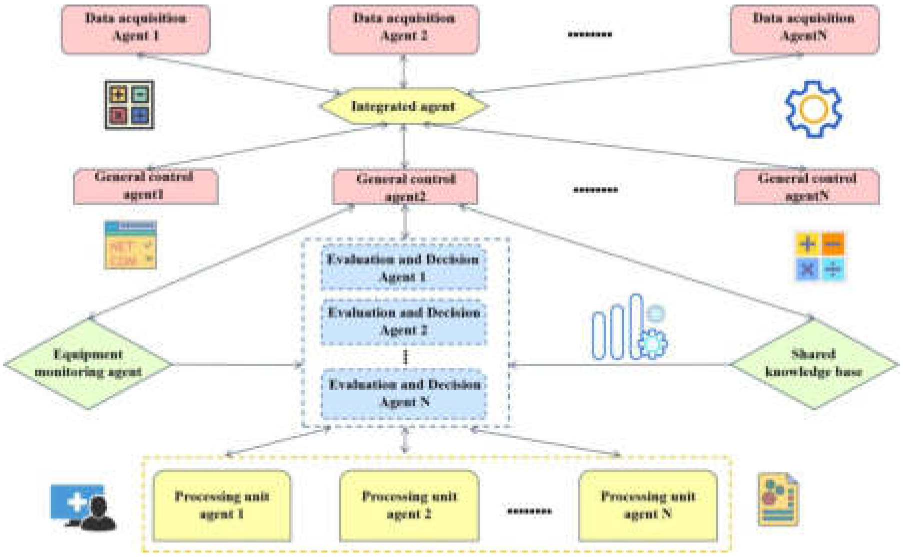


Fig. 1.1: Structure of a Multi Agent Based Workshop Dynamic Scheduling System

scheduling model that is suitable for the actual production status of printing and dyeing enterprises.

(1) Genetic Algorithm. After abstracting the above characteristics of dyeing production, the problem can be described as follows: Assuming that the production workshop will produce n independent products on m dyeing tanks during the planning period [1,T][9,10]. Based on the summary of orders, the contract delivery quantity $d_J(t)(j=1,2,...,T)$ for the jth product on day t can be obtained, and the daily available capacity range of the i-th dyeing tank is $[g_i, G_i](i=1,2,...,m)$, the switching time and cost of the dye tank are linearly related to the capacity of the dye tank. At the initial moment, the storage capacity of product j is $I_j(j=1,2,...,n)$, and the storage capacity I_j represents the delivery quantity of the jth product that was not completed during the previous planning period at the initial planning time[11]. This can be obtained by monitoring the operation process of the dyeing cylinder through the MES manufacturing system in the workshop. After obtaining the quantity of unfinished orders, according to the production schedule requirements, select the total amount of products to be processed in the ERP system order database, and complete the pull dynamic production scheduling within the planned number of days.For enterprises, delayed delivery requires payment of a penalty for breach of contract to customers, assuming that the penalty for delayed delivery per unit of product time is $\alpha_i(j=1,2,...,n)$.

(2) Design of Algorithm. Based on the characteristics of the problem model, in order to effectively reflect the order and quantity of products processed by each dyeing tank on a daily basis, a natural encoding method is adopted[12]. The specific scheme is as follows:

$$p_{11}(t), p_{12}(t), ...; p_{1n}(t), ...; p_{m1}, p_{m2}, ...; p_{mn}(t) \tag{2.1}$$

Among them, $p_{nm}(t)$ represents the output of the nth product on the mth dyeing tank on the th day. When initializing the population, if a random approach is used, it is difficult to guarantee that the resulting solution is feasible. Therefore, the method for initializing the population here is: For the i-th dyeing tank on the t-th day, generate a random number R_i within the interval $[g_i, G_i]$[13,14]. For the quantity of various products produced by this dyeing tank every day, use a random average distribution method and ensure $\sum\limits_{j=1}^{n} p_{ij}(t) = R_i$. Adopting this initialization method not only ensures the diversity of the population, but also ensures that the initial solution is feasible.

The purpose of selection is to select excellent individuals from the current population, so that they have the opportunity to reproduce as parents for the next generation. Based on the fitness values of each individual,

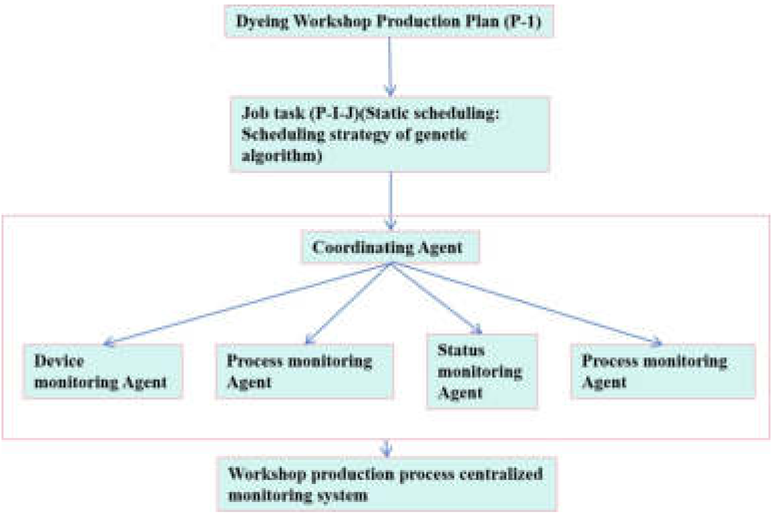

Dyeing Workshop Production Plan (P-1)
Job task (P-I-J)(Static scheduling: Scheduling strategy of genetic algorithm)
Coordinating Agent
Device monitoring Agent
Process monitoring Agent
Status monitoring Agent
Process monitoring Agent
Workshop production process centralized monitoring system

Fig. 2.1: Multi agent based workshop production scheduling model

select some excellent individuals from the previous generation population according to certain rules or methods to inherit into the next generation population. Compared with the simulation experiment results, we chose the operation method of uniform sorting, which sorts all individuals in the population according to their fitness size, and based on this sorting, assigns the probability of each individual being selected.

Cross operation is the most important genetic operation in genetic algorithms[15]. Through crossover operations, a new generation of individuals can be obtained, where each individual within the population is randomly paired and a portion of their chromosomes are exchanged with a certain probability for each individual. In order to meet the constraint requirements of impregnation production during crossover, the two-point crossover method is adopted, which randomly sets two crossover points in the individual coding string, and then conducts partial gene exchange.

Mutation operations are mainly used to adjust some gene values in an individual's coding string, which to some extent overcomes the situation of effective gene deletion and is beneficial for increasing population diversity. In order to ensure that the mutated chromosomes have a good individual coding structure, randomly select the mutated genes and use adjacency search method to insert the gene values that meet the constraint conditions (excluding the gene values that need to be mutated) into the selected mutated genes once, overwrite existing genes to generate new chromosomes, and select the best of them as the offspring of the mutation. And it can also effectively ensure the diversity of the population, the quality of mutated individuals, and the feasibility of producing individuals. Pre set a maximum number of evolution steps N_{max} , and if the maximum number of evolution steps is reached, terminate the algorithm process.

2.3. Multi agent based dynamic scheduling model for dyeing vats. Based on the characteristics of printing and dyeing processes and production processes, a dynamic model for workshop production scheduling based on multi-agent is constructed[16]. This model is based on the real-time operation status of the dyeing tank, combining static scheduling and dynamic adjustment, and achieving the goal of global optimization through the interaction between intelligent agents. The model has a four layer architecture: The first layer is composed of a workshop production planning layer; The second layer is the static scheduling layer; The third layer is the dynamic scheduling layer; The fourth layer is to control friction. As shown in Figure 2.1, it is a multi-agent based workshop production scheduling model.

2.4. Intelligent Agent Collaboration Process. Figure 2.2 shows the internal structure of the intelligent agent. This structure consists of modules such as communication interface, intelligent control, state saving, data collection, data processing, knowledge base, data view, and output, as shown in Figure 2.2 [17].

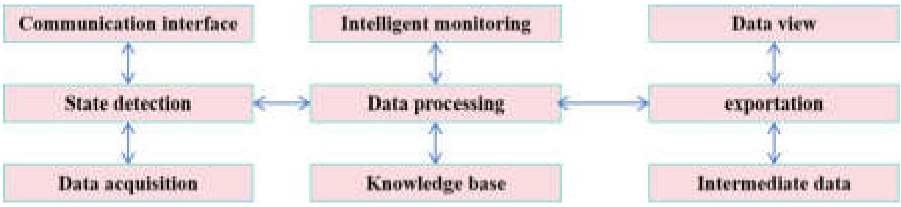


Fig. 2.2: Internal structure diagram of the intelligent agent

The collaborative process of intelligent agents can be divided into two types: Internal and external. Internal coordination is carried out in internal modules, such as communication interface message publishing, external data collection, data processing, intelligent analysis and control, and final data output. The intermediate data recording and working status saving module saves the intermediate results and system status of the data collection and processing process. The planning and control module is the coordinator and commander of various modules within the entire intelligent agent. It calls the corresponding modules for processing according to certain rules and requests from the communication interface module. The rule library stores the internal rules and control algorithms of the monitoring agent, while external coordination involves the interaction and coordination of information such as equipment operation, production process, process status, and operational status for optimizing scheduling. Due to the lack of consideration for the dynamic changes in the production site, the scheduling results obtained by static genetic algorithms are not optimal. Therefore, multi-agent technology needs to be added to reschedule individual scheduling tasks. Firstly, the coordinating agent determines the allocation of tasks on resources and the processing time on production equipment based on production plans and actual resource utilization. Afterwards, the equipment monitoring intelligent agent obtains processing task orders based on its own capabilities, and completes the optimization and scheduling of production tasks based on constraints such as meeting product delivery dates, processes, costs, energy consumption, and quality requirements, combined with the operational status of the production site dyeing vats. Coordination agents are also responsible for coordinating the behavior of various agents, resolving conflicts, synchronization, asynchrony, and other issues between agents, in order to ensure the coordinated operation of the production system.

3. Simulation calculation. Simulation of hierarchical scheduling using typical data samples from a dyeing enterprise to verify the feasibility of the scheme. Select two dyeing vats in a workshop group as the object, and set up four products with a production period of 1-15 days. Assuming that the penalty for one day of delay for each product is 0.1 yuan, the daily available capacity ranges of the two dyeing vats are [35, 45] and [40, 50], respectively[18]. According to the order information, the distribution of order quantities for the four products is shown in Table 3.1. Using the scheduling model and static genetic algorithm designed above, the calculation results were obtained by iterating around 1500-2000 times, with high computational efficiency. The optimal production schedule obtained by solving the algorithm is shown in Table 3.2, with a minimum production cost of 56.1. The planned period is 15 days, and ”A/B” indicates that the quantity of products with serial number A processed on this dyeing tank on that day is B. After encapsulating the static genetic algorithm in the inference mechanism of the agent in Table 3.3 and coordinating with multiple agents, the production schedule is formulated, and the production cost calculated by the model is 43.6. Compared with simply using static genetic algorithms, it saves 22.3% in production costs.

Table 3.1: Order Quantity of Four Products per Day

programme	Order quantity of product serial number			
	1	2	3	4
1	61	0	21	0
2	0	81	0	41
3	41	0	0	71
4	0	21	41	0
5	0	0	71	56
6	81	0	41	0
7	0	0	31	61
8	0	0	81	46
9	31	51	0	0
10	0	31	51	21
11	41	0	0	61
12	36	0	51	0
13	0	51	0	41
14	0	41	0	46
15	31	21	0	21

Table 3.2: Scheduling Results Based on Static Genetic Algorithm

fatalism	No.1 dyeing cylinder		No.2 dyeing cylinder	
	(capacity: 35 15)		(capacity: 10 50)	
1	2/23	4/21	2/24	4/23
2	2/24	3/20	1/22	2/10 4/17
3	1/24	3/21	2/24	3/24
4	2/24	4/22	2/7 3/24	4/20
5	1/21	2/22	3/23	4/26
6	2/22	4/25	2/24	3/24
7	1/24	3/23	1/29	3/22
8	1/23	3/22	3/24	4/24
9	1/21	4/25	1/4 2/25	3/21
10	2/1 3/22	4/24	1/21	2/26
11	3/21	3/23	1/17 3/9	4/17
12	1/24	4/23	3/21	2/25
13	1/21	2/23	1/24	4/25
14	1/15	2/23 3/4	2/24	3/21
15	2/23	4/21	1/3 2/23	4/25

Table 3.3: Scheduling results based on multi-agent and genetic algorithm

fatalism	No.1 dyeing cylinder (capacity: 35 45)		No.2 dyeing cylinder (capacity: 40 50)	
1	1/23	3/24	1/24	4/26
2	2/24	3/22	1/22	3/10 4/20
3	1/24	4/23	2/24	4/24
4	124	4/22	2/7 3/24	4/20
5	1/21	2/22	1/23	4/26
6	2/22	3/23 4/2	2/24	3/20 4/8
7	2/24	4/21	1/29	3/22
8	1/23	2/2 3/22	3/28	4/24
9	1/21	2/1 3/23	1/4 2/25	3/21
10	2/1 3/22	4/24	1/21	2/26 3/4
11	3/21	4/23	1/16 2/19	3/9 1/17
12	2/24	3/23	1/21	2/25 3/6
13	1/21	2/23 3/3	1/24	3/24 4/4
14	1/15	21/23 4/10	2/24	3/16 4/12
15	2/23	3/24	2/3 3/23	4/26

By comparing the scheduling results of static genetic algorithms, it can be concluded that using the multi agent and genetic algorithm proposed by the author to combine the hierarchical scheduling design of dyeing vats, considering the dynamic changes in the production site, the optimal solution for dyeing vat scheduling can be obtained[19,20]. After actual production workshop operation testing, the algorithm in this article has a good optimization efficiency when dealing with production lines with 60 dyeing tanks producing 8 product types.

4. Conclusion. The author analyzed the characteristics of dyeing production and established a hierarchical scheduling model that meets the actual production constraints. Due to the use of static genetic algorithm as the solution algorithm, it is difficult to consider the impact of actual changes in production site dyeing vats, yarn, and workshop personnel, and the scheduling plan made does not meet the optimal solution of the production site, therefore, the author proposes a hierarchical scheduling strategy for dyeing vats based on genetic algorithm and multi-agent. The comparison of simulation results with static genetic algorithm production scheduling verifies the effectiveness of the hierarchical design method proposed by the author. Through improvement, it can adapt to the formulation of production scheduling plans for large-scale dyeing enterprises. This has certain reference value for effectively solving the job scheduling problem of dyeing production workshops, achieving cost reduction and emission reduction goals.

REFERENCES

[1] Wang, Y., Yang, R. R., Xu, Y. X., Li, X., & Shi, J. L. (2021). Research on multi-agent task optimization and scheduling based on improved ant colony algorithm. IOP Conference Series: Materials Science and Engineering, 1043(3), 032007 (11pp).

[2] Xu, K., Wang, H., & Liu, P. X. (2023). Adaptive fixed-time output feedback formation control for nonstrict-feedback nonlinear multi-agent systems. International Journal of Systems Science, 54(11), 2281-2300.

[3] Wang, Q., Guo, F., Zhang, A. T., Guo, Y., & Qiang, Y. M. (2022). Control of the consensus of second-order multi-agent systems with time delay based on distributed pi. journal of unmanned undersea systems, 30(4), 457-464.

[4] Nitsche, B., Brands, J., Treiblmaier, H., & Gebhardt, J. (2023). The impact of multiagent systems on autonomous production and supply chain networks: use cases, barriers and contributions to logistics network resilience. Supply Chain Management: An International Journal, 28(5), 894-908.

[5] Wang, Y. (2021). Optimization of english online learning dictionary system based on multiagent architecture. Complexity, 2021(4), 1-10.

[6] Li, B., Song, C., Zhao, J., & Yu, J. (2023). Robust exponential stability analysis of switched systems under switching boundary mismatch. International Journal of Robust and Nonlinear Control, 33(11), 6459-6480.

[7] Wang, J., & Li, Y. (2021). Research on fault tolerance consistency of multi - agent based on event triggering mechanism. Journal of Physics: Conference Series, 1993(1), 012018-.

[8] Tao, M., Wang, Z., & Qu, S. (2021). Research on multi-microgrids scheduling strategy considering dynamic electricity price based on blockchain. IEEE Access, PP(99), 1-1.

[9] Zhao, M., & Li, D. (2021). Collaborative task allocation of heterogeneous multi-unmanned platform based on a hybrid improved contract net algorithm. IEEE Access, PP(99), 1-1.

[10] Kamruzzaman, M., Duan, J., Shi, D., & Benidris, M. (2021). A deep reinforcement learning-based multi-agent framework to enhance power system resilience using shunt resources. IEEE Transactions on Power Systems, PP(99), 1-1.

[11] Luo, X., Zhang, Z., Tang, D., Zhu, H., Zhou, T., & Pulido, A. S. R., et al. (2022). A practical approach for multiagent manufacturing system based on agent computing nodes:. Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science, 236(4), 1907-1930.

[12] Talmale, G., & Shrawankar, U. (2021). Dynamic multi-agent real time scheduling framework for production management. IOP Conference Series: Materials Science and Engineering, 1085(1), 012001 (6pp).

[13] Yang, Y., Liu, Q., Yue, D., & Han, Q. L. (2021). Predictor-based neural dynamic surface control for bipartite tracking of a class of nonlinear multiagent systems. IEEE Transactions on Neural Networks and Learning Systems, PP(99), 1-12.

[14] Chen, D., Liu, X., Yu, W., Zhu, L., & Tang, Q. (2021). Neural-network based adaptive self-triggered consensus of nonlinear multi-agent systems with sensor saturation. IEEE Transactions on Network Science and Engineering, PP(99), 1-1.

[15] Zhang, G., Li, X., An, J., Zhang, Z., Man, W., & Zhang, Q. (2021). Summary of research on satellite mission planning based on multi-agent-system. Journal of Physics: Conference Series, 1802(2), 022032 (5pp).

[16] Wang, T., & Yang, X. (2021). Optimal network planning of ac/dc hybrid microgrid based on clustering and multi-agent reinforcement learning. Journal of Renewable and Sustainable Energy, 13(2), 025501.

[17] Wang, J., Wen, G., Duan, Z., Hu, Y., & He, W. (2021). Distributed h∞ robust control of multiagent systems with uncertain parameters: performance-region-based approach. IEEE Transactions on Systems, Man, and Cybernetics: Systems, PP(99), 1-11.

[18] Hou, H. Q., Liu, Y. J., Liu, L., & Lan, J. (2023). Adaptive fuzzy formation control for heterogeneous multi-agent systems using time-varying iblfs. Nonlinear Dynamics, 111(17), 16077-16091.

[19] Chen, Z., Zhang, L., Wang, X., & Gu, P. (2022). Optimal design of flexible job shop scheduling under resource preemption based on deep reinforcement learning. Complex System Modeling and Simulation, 2(2), 174-185.

[20] Shan, H., Wang, C., Zou, C., & Qin, M. (2021). Research on pull-type multi-agv system dynamic path optimization based on time window:. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 235(7), 1944-1955.

Edited by: B. Nagaraj M.E.
Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing
Received: Oct 5, 2023
Accepted: Nov 21, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1499–1507, DOI 10.12694/scpe.v25i3.2661

APPLICATION OF DEEP LEARNING ALGORITHM IN OPTIMIZATION OF ENGINEERING INTELLIGENT MANAGEMENT CONTROL

SHUAI ZHOU*AND JING GUO†

Abstract. Currently, there are a series of problems in the management of the construction industry, such as resource waste, substandard quality, and low construction efficiency. In response to this phenomenon, the author proposes a multi-objective optimization control method for construction engineering management projects using deep learning algorithms. This method analyzes the relationship between cost, duration, and quality, and constructs an optimization management model for these three factors. At the same time, the improved SULSTM neural network algorithm is used to optimize the model parameters. The experimental results indicate that, when the value coefficient is 0.2211, the total investment cost and quality coefficient are 412700 yuan and 0.99496 yuan, respectively. When the value coefficient is 0.1976, the total cost and quality coefficient are 456300 yuan and 0.98798 yuan, respectively. When the value coefficient is 0.1990, the total cost and quality coefficient are 456300 yuan and 0.99496 yuan, respectively. Proved that the SUSTM neural network algorithm has faster convergence speed and lower loss values compared to the improved LSTM neural network algorithm. The cost of improving quality has a greater impact on the quality coefficient than the duration, and the total investment cost has a greater impact on the value coefficient than the quality coefficient.

Key words: Construction engineering, Multi objective optimization, Management efficiency, SUSTM neural network structure

1. Introduction.

At present, due to the influence of market environment and national control measures, the construction industry has encountered some problems in construction project management. After the country issued a series of market control measures, it is difficult for construction enterprises to grasp the latest national control policies and industry management decisions under the influence of market environment. It is difficult for enterprises to adjust in real-time according to national policies and changes in market environment during development, this leads to a mismatch between the company's own goal control and management level and socio-economic development, and a lack of timely response to market cyclical changes [16]. In fact, how construction enterprises improve management efficiency in the current market environment, and how to optimize management efficiency under multi-objective conditions will be important means to improve the overall efficiency of the enterprise, quickly adapt to external changes from within, and promote the healthy and stable development of the industry.

For a long time, homeowners have often established a temporary organization to carry out the design, procurement, construction, and daily quality and progress management of construction projects. If the owner has rich experience in engineering management, or if the project size is not large and the technical requirements are relatively simple, under such conditions, the traditional construction project management can achieve good results. However, in recent years, with the continuous improvement of the national economy and the continuous construction and development of municipal supporting facilities and public infrastructure, the number of large-scale engineering projects in related fields such as architecture, energy, and municipal highways has gradually increased, which has prompted homeowners to have higher requirements for various professional goals [3]. There are also higher expectations for the environment, quality costs, and progress. Corresponding construction project managers need to have richer management experience and more advanced construction techniques in order to keep up with the speed of industry development and meet the needs of industry development. Therefore, accurately grasping the overall goal of the project from the source, reasonably allocating various resources in the early stage of construction, promoting the improvement of project engineering quality, reducing construction investment to a certain extent, and shortening the construction period. The handling and decision-making of all these issues are of great significance for the entire construction project.

*Hebei Vocational University of Industry and Technology, Shijiazhuang, Hebei, 050091, China
†Hebei Vocational University of Industry and Technology, Shijiazhuang, Hebei, 050091, China (Corresponding Author)

Construction project management refers to the systematic regulation and management of projects based on management theories and methods, and the reasonable allocation of costs and construction periods while ensuring the expected functions of construction projects, ensuring quality assurance, and the healthy operation of daily project work. The management of construction projects has very obvious characteristics, which are particularity, one-time, strong constraint, and full lifecycle. For each construction project, the control of quality, cost, and progress is very important. By comprehensively controlling these issues, the expected management goals can be achieved. The regulation of these stages is comprehensive and holistic, with the goal of overall benefits. The allocation of various resources, reasonable construction period, cost investment, and guaranteed quality goals all comprehensively determine the accurate control of limited resources in the early stages of the project [8]. The construction unit, general contractor, and each construction unit have all undertaken corresponding tasks in each stage, and although there may be differences in tasks in each stage, the overall goal remains the same. Generally speaking, cost, schedule, and quality are the three basic objectives of project management. In the project management knowledge system PMBOK, the American Project Management Association has extended the project objective system to a certain extent after in-depth research and analysis, adding four additional contents, namely procurement, risk, communication, and integration. Based on these nine objectives, the basic framework of PMBOK has been formed. At present, research on construction project management is mostly based on this framework as the research premise. In future research and development, regardless of how the number of management objectives increases, progress, cost, and quality are still the three most important basic goals that have the greatest impact on the overall goal throughout the entire life cycle of the construction project. How to find a balance among the three, obtaining the optimal project execution effect remains the fundamental goal of construction project management.

2. Literature Review. At present, there are not many research results on the efficiency of construction project management in the country, and most of them are focused on a single goal. The author's research on the efficiency of construction project management starts from three aspects: progress, cost, and quality. Considering that these three goals have the greatest impact on the overall goals of the project, so through comprehensive optimization of this aspect, the overall level of construction project management can be improved. Moreover, the author's reasonable quantification of quality parameters has greatly improved the effectiveness of quality objective research. Through the construction of a comprehensive analysis model of progress cost quality, the current insufficient research on management efficiency under multi-objective construction engineering has been effectively improved, which can enrich the existing management efficiency optimization theory [1]. When evaluating the optimization effect of management efficiency, the author introduced value engineering as a measurement indicator and reasonably transformed the cost and quality objectives in construction engineering into value engineering, using the idea of high or low value to reflect the advantages and disadvantages of project multi-objective regulation. Not only does it provide an effective tool for optimizing the efficiency of construction project management, but it also facilitates the integration of management research with other disciplines. There are reference examples in the process of setting goals for construction project management. At present, the common problem in construction projects is that project managers do not attach importance to management objectives, and the allocation of various resources is not scientific and systematic enough [15]. Through the author's research, it can provide a certain reference for the implementation of construction project management. Construction project managers do not attach great importance to management efficiency from an ideological perspective. They often only focus on whether their management methods have been implemented in actual construction, and do not intuitively realize the effectiveness of improving management efficiency in improving the overall efficiency of the enterprise. The author's research can make up for this deficiency and effectively strengthen the importance of management efficiency by managers, furthermore, it ensures that the management costs invested in the implementation of construction projects can achieve maximum benefits.

Regarding the optimization of management efficiency in construction projects, foreign scholars have continuously increased their attention in recent years. Through the application of different research methods and theories, as well as attempts at various research objectives, a large amount of research has been conducted on the overall optimization of construction industry and construction project management efficiency, which has been continuously promoted and developed on the basis of previous research [7, 6].

With the continuous improvement of living standards and the increasing electricity load, the number of power transmission and transformation equipment is also rapidly increasing. The original maintenance mode is insufficient to ensure the safe operation of the huge power grid. This article mainly studies the research and application of machine learning based optimization technology for substation equipment maintenance decision-making. Liu, Z, based on the technical principles of online monitoring and status maintenance of substation equipment, combined with deep learning models, implemented an intelligent monitoring and maintenance early warning system. The main functions of this system include monitoring equipment management, operation monitoring, and comprehensive display, which can effectively carry out online monitoring and status warning for substation equipment [11]. Whale Optimization Algorithm (WOA) is a relatively novel algorithm in the field of meta heuristic algorithms. Compared with other mature optimization algorithms, WOA can demonstrate efficient performance, but there are still problems of premature convergence and easy falling into local optima in complex multimodal functions. Therefore, Guo, Y. K. Z proposed an improved WOA and proposed a new strategy of random jump change and a random control parameter strategy to improve the exploration and utilization ability of WOA. This article uses 24 well-known benchmark functions to test the algorithm, including 10 unimodal functions and 14 multimodal functions. The experimental results show that the convergence accuracy of this algorithm is better than the original algorithm on 21 functions, and better than the other 5 algorithms on 23 functions [5]. We combine Deep Gaussian Process (DGP) with multitasking and transfer learning for performance modeling and optimization of HPC applications. The deep Gaussian process combines the uncertainty quantification advantages of Gaussian processes with the predictive ability of deep learning. Multi task and transfer learning allow for improved learning efficiency when learning several similar tasks simultaneously, as well as when seeking models from previous learning to assist in learning new tasks separately. The comparison with state-of-the-art automatic tuners shows the advantages of our method in two application problems. In this article, Dongarra, J. combines DGP with multitasking and transfer learning, which can improve the adjustment of application parameters for problems of interest and predict parameters for any potential problems that the application may encounter [13].

In summary, domestic and foreign scholars have conducted research and analysis on the efficiency of construction project management from different perspectives, realizing the importance of optimizing management efficiency, and using various methods to establish systematic analysis models to improve the management effectiveness of construction projects. The proposed improvement measures also have practical guiding significance. Although current management efficiency has gradually been integrated into construction project management, the existing research literature mainly studies a single or two elements in the construction management process, and the essence of its research is still linear. From the perspective of managers, there is not much research on multi-objective comprehensive control in the construction process. Therefore, based on the dependency deep learning algorithm, the author establishes a multi-objective comprehensive analysis model to optimize the management efficiency of construction projects, and calculates and verifies it through examples, providing a new approach for subsequent research.

3. Methods.

3.1. Optimization Control Model for Construction Engineering Management Oriented to SUSTM. The optimal control model for construction project management is a management behavior that involves the entire production process of construction products, with characteristics such as one-time, comprehensive, and strong constraints. Construction project management includes quality management, schedule management, cost management, contract management, safety management, risk management, communication management, human resource management, information management, and environmental protection. The management project needs to ensure efficient management efficiency, and the recognized influencing factors at home and abroad include four aspects: Management process, management methods, manager quality, property rights, and responsibility system. Research the introduction of value engineering for multi-objective control optimization analysis of construction project management, replacing product functionality with quality, treating cost as contracting cost, and value coefficient as the ratio of quality to cost [14].

The flow chart of optimization control for construction projects is shown in Figure 3.1. Before the project officially starts, the first step is to establish a comprehensive model of schedule cost quality optimization for construction project management based on reasonable quantification of construction project quality. The second step is to optimize the three objectives and obtain corresponding target values, and then use the obtained target

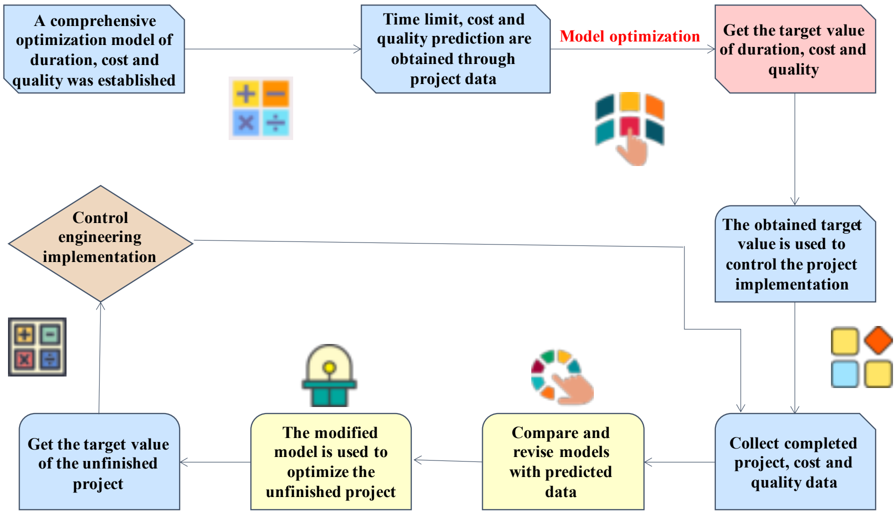

A comprehensive optimization model of duration, cost and quality was established
Time limit, cost and quality prediction are obtained through project data
Model optimization
Get the target value of duration, cost and quality
The obtained target value is used to control the project implementation
Control engineering implementation
Collect completed project, cost and quality data
Compare and revise models with predicted data
The modified model is used to optimize the unfinished project
Get the target value of the unfinished project

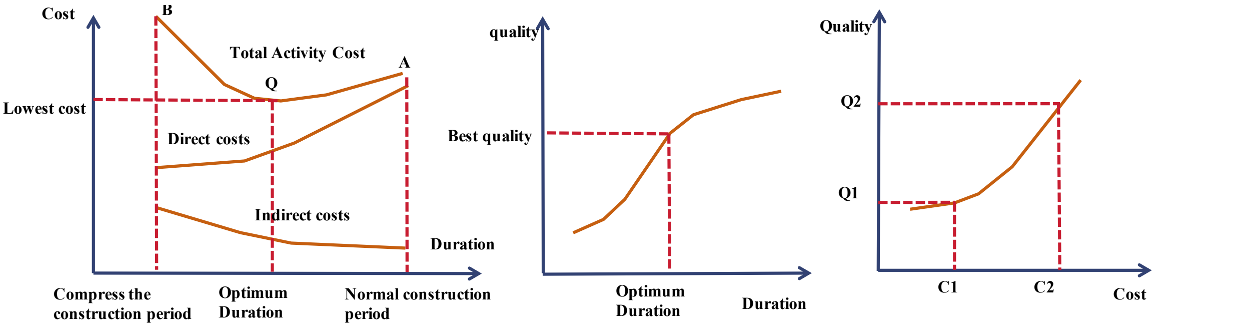
Fig. 3.1: Optimization control flow chart of the construction project

values to control the construction project [2]. After the project has been put into construction, data information on time, quality, and completed project costs is first collected. Then, these data are used to modify and optimize the comprehensive model. The next step is to use the modified model to optimize the unfinished projects and obtain three target values. Repeat the above steps until the optimal solution is obtained and the optimization process is completed.

The study introduces a quality coefficient for quantification, and the relationship between the quality coefficient and the six major quality characteristics can be expressed through calculation expressions. The method for determining quantitative quality based on the following characteristics of engineering projects, including applicability, safety, economy, safety and environmental protection, durability, and reliability, is represented by the letters a1-a6 [18]. Quality quantification can control the quality of construction projects during the construction process through quality coefficients. Different quality coefficients indicate inconsistent quality quantification values. When the quality coefficient is about 1, the minimum values of a1-a6 are 0.8, 0.9, 0.9, 0.9, 0.8, and 0.9, respectively. When the quality coefficient is about 0.9983, the minimum values of a1-a6 are 0.6, 0.7, 0.6, 0.6, 0.7, and 0.7, respectively. When the quality coefficient is about 0.97, the minimum values of a1-a6 are 0.5, 0.5, 0.3, 0.3, 0.5, and 0.5, respectively. When the quality coefficient is about 0.85, the minimum values of a1-a6 are 0.3, 0.3, 0.2, 0.2, 0.3, and 0.3, respectively. When the quality coefficient is about 0.47, the minimum values of a1-a6 are 0.1, 0.1, 0.1, 0.1, 0.1, and 0.1, respectively [21]. At the same time, the degree of satisfaction of the six quality characteristics of a1-a6 is as follows. In the case of slight satisfaction, the value range of a3-a4 is 0.1∼0.2, while the rest are 0.1∼0.3. Under basic conditions, the value ranges of a3-a4 are 0.2-0.4 and 0.2-0.3, respectively, while the other quality characteristic ranges are 0.3-0.5. In the case of satisfaction, the value ranges of a3-a4 are 0.4-0.6 and 0.3-0.6 respectively, while the other quality characteristic ranges are 0.5-0.6. In very satisfactory cases, the range of values for a3-a4 is 0.6-0.9, while the rest are 0.6-0.8. Under very satisfactory conditions, the range of values for a3-a4 is 0.9-1, while the rest are 0.8-1.

If the construction project includes a unit project with a total quantity of w, and there is a unit project involving a sub project with a quantity of r in the unit project, and the construction project also includes a sub project with a quantity of k. When the quality coefficient of the sub item is set, the corresponding satisfaction level value is set to 0. The calculation formula for the quality coefficient of construction engineering is:

$$\begin{cases} q^0 = 1 - \prod^0 (1 - a_m) \\ q^1 = 1 - \prod^k (1 - q^0 i) \\ q^2 = 1 - \prod^r (1 - q^1 i) \\ q = 1 - \prod^w (1 - q^2 i) \end{cases} \tag{3.1}$$



(a) The relationship between direct costs, indirect costs, and construction period(b) Relationship between construction period and quality(c) Relationship between cost and quality

Fig. 3.2: Progress cost quality relationship

In equation (3.1), $i = 1, 2, K$, q^0 refer to the quality coefficient of sub projects, q^1 and q^2 represent sub projects and unit projects, respectively, q is the quality coefficient of construction engineering. On this basis, the study establishes a quality cost schedule model and obtains the optimal quality coefficient, which is then compared and analyzed with the aforementioned quality characteristics and satisfaction level. Figure 3.2 shows the relationship between progress cost quality [12].

This includes both direct and indirect costs. The directly incurred expenses include the costs of raw materials, labor, and equipment. If there is a compression of the construction period, the direct costs incurred will also increase accordingly. Indirect costs mainly include enterprise management fees and training fees, which will continue to decrease with the acceleration of project progress, mainly manifested as the shortened usage time of leased equipment and prefabricated houses. Therefore, the optimal completion time can be determined by combining two types of costs. Figure 3.2 (b) refers to the relationship between project progress and quality. There is an opposing and unified relationship between progress and quality. The acceleration of progress is likely to cause a decrease in the overall quality level of construction projects [17]. If the balance and continuity of progress can be ensured, quality can meet the standards as much as possible. At the same time, the introduction of science and technology and the improvement of management level will also improve the overall progress and quality. Figure 3.2 (c) shows the relationship between cost and quality. Introducing advanced technology and equipment not only increases costs, but also increases the quality of construction projects. When the cost is higher than C1, the quality meets the minimum quality standard. When the cost is higher than C2, the quality level does not improve significantly, but the cost increases faster. Therefore, in the actual management process, it is advisable to choose a cost range of C1-C2 and a construction quality range of Q1-Q2.

C_1 and c_2 represent the cost of investment under the conditions of qualified quality and improvement, respectively. Before establishing a comprehensive model, the following conditions need to be established. C_1 and time t are inversely linearly correlated, with q_0 increasing with the increase of c_2 and q_0 decreasing with the decrease of t. The calculation expression for the decision variables of the model is equation (3.2).

$$\begin{cases} c_{1i} = \frac{\left(c_{1i}^C - c_{1i}^N\right)(t_i - t^c)}{t^C - t^N} + c_{1i}^C \\ c_i = c_{1i} + c_{2i} \\ q_i^0 = Ac_{2i}^2 + Bt_i \end{cases} \tag{3.2}$$

In equation (3.2), $A = \frac{t^N q^c - t^c q^N}{\left(c_2^N\right)^2 t^N - \left(c_2^c\right)^2 t^C}$, $B = \frac{\left(c_2^N\right)^2 q^N - (c_2^c)^2 q^C}{\left(c_2^N\right)^2 t^N - \left(c_2^c\right)^2 t^C}$, t_i^N and t_i^C refer to the normal and minimum durations of sub item i, respectively, c_{1i}^N and c_{1i}^C respectively refer to the lowest and highest costs after the building is qualified, c_{2i}^N and c_{2i}^C refer to the lowest and highest costs after improving quality, q_i^N and q_i^C refer to the minimum and maximum quality coefficients, and t_i refers to the actual duration. The constraint conditions are as follows: firstly, $\sum t_i \leq T_c$, where T_c is the calculated duration. Secondly, the minimum cost is less than or equal to the construction project cost, while the project cost is less than or equal to the project contract price. Thirdly, the quality coefficient is equal to or greater than the minimum qualification coefficient [10].

3.2. Selective Update Neural Network Structure Solving Model. The LSTM neural network structure is based on the classical recurrent neural network, introducing three logical structures: Input gate, output gate, and forgetting gate. Forgetting gate is the process of clearing all unimportant information in the unit state. Its input includes a specific time step x^t and the unit state of the previous hidden layer $h^{(t-1)}$. The control function determines whether the information is cleared or retained. The value interval of the vector $f^{(t)}$ for the final output unit state is $[0, 1]$. If the value is 1, then the input value is retained as a whole. If the value is 0, then the value is deleted as a whole. Input refers to determining whether information is added to the unit state for data update, using the Sigmad function to delete the information of x^t and $h^{(t-1)}$, and then calculating the current input unit state. After establishing the Tach function, a vector is selected with a value range of [-1,1], and finally calculating the unit state $c^{(t)}$ at the current time, this value is first multiplied by the previous unit state $c^{(t-1)}$ and the forgetting gate, then added to $c'(t)$, and multiplied by the input gate $i^{(t)}$ to obtain the final result. The output gate selects the valuable unit states that need to be presented for output, and its specific implementation process includes two steps, firstly, a filter $o^{(t)}$ is obtained by using x^t and $h^{(t-1)}$, and then the Tach function is selected to compress the values of the unit state vector into an interval of [-1,1]. At the same time, the result obtained by multiplying the vector and $o^{(t)}$ is used as the basis for determining the hidden information $h^{(t)}$. The selectively updated long-term and short-term memory neural network structure is based on the LSTM neural network structure, combining the forgetting gate and output gate as update gates, greatly reducing the training time of the neural network model and improving the learning efficiency of the algorithm. The update formula for the improved SUSTM neural network structure is equation (3.3).

$$\begin{cases} f^{(t)} = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_t\right) \\ o^{(t)} = \sigma\left(W_o \cdot [c_t, h_{t-1}, x_t] + b_o\right) \\ c'_t = \tanh\left(W_c \cdot [c_t, h_{t-1}, x_t] + b_c\right) \\ c_t = f_t \cdot c_{t-1} + (1 - f_t) \cdot c'_t \\ h_t = o_t \cdot \tanh\left(c_t\right) \end{cases} \tag{3.3}$$

W_f, W_c, W_o respectively represent the weights of the update gate, calculation input unit, and output gate, while b_t, b_c and b_o respectively refer to the biases of the update gate, calculation input unit, and output gate.

4. Analysis of the Optimal Control Model for Construction Engineering Management.

4.1. Model Optimal Parameters. The author takes the construction project undertaken by a certain construction company as an example. The project covers an area of 145.53 acres, has a plot ratio of 2.5, a total land cost of 470 million yuan, a floor price of 2114 yuan/m^2, and a building density of 30%. The experimental analysis includes a sales area with a building area of 1650m^2, a floor height of 8.6m, a contract period of 150 days, and a contract cost of 2.68 million yuan, the lowest quality coefficient is 0.8, and the lowest enterprise cost is 2.2 million yuan [20]. The maximum value coefficient calculated through the composition and parameters of the engineering project is 0.2375. The optimal parameters of the corresponding optimization control model for construction project management are shown in Table 4.1, and the quality qualification cost, improvement quality cost, quality coefficient, and duration of 11 sub projects are obtained. The experiment utilized a progress quality coefficient cost optimization model for data analysis, with an optimized cost of 2.2919 million yuan and a quality coefficient of 0.8325. Therefore, this optimization model can bring more economic profits to the enterprise while ensuring quality.

Figure 4.1 shows the model training loss results of deep learning neural networks before and after improvement. It can be seen from the figure that the training loss values of the two network structures continue to decrease with the increase of iteration times. Both LSTM and SUSTM neural network algorithms converge quickly at around 20 iterations, and the difference between the two algorithms is not particularly significant [9]. However, when the number of iterations ranges from 20 to 100, compared to the LSTM neural network algorithm, the convergence speed of the SUSTM neural network algorithm is faster, and the loss value tends to be more stable.

Table 4.1: Optimum parameters for optimal control of construction engineering management

Distribution entry	Distribution entry	Quality qualified fee / ten thousand yuan	Quality improvement cost / ten thousand yuan	Mass coefficient	Duration days / day
Foundation and foundation sub project	Earthwork project	10.48	7.46	0.832	8.12
	Reinforcement project	3.56	4.45	0.781	4.06
	Template Project	1.14	1.43	0.765	1.01
	Brick building project	7.25	9.32	0.833	6.11
	Waterproof project	2.39	4.22	0.791	18.26
Main structure sub projects	Template Project	22.87	27.31	0,761	6.12
	concrete	7.41	9.35	0.811	24.45
	Construction projects	30.5	35.61	0.801	12.11
Construction wall sub items from	Wall project	15.36	18.41	0.765	15.23
Decoration sub project	External wall plastering project	3.55	2.81	0.732	3.11
	Ground projects	9.75	12.56	0.912	8.05

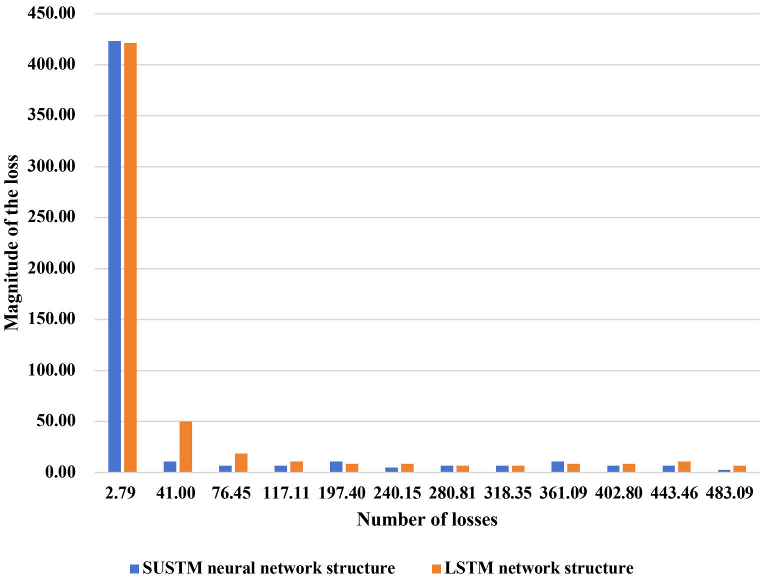


Fig. 4.1: Model training loss results

4.2. Sensitivity Analysis. The study further utilizes sensitivity analysis theory to determine the most important factors affecting the economic interests of enterprises, while verifying the correctness and rationality of the proposed model.

Based on the aforementioned content, the cost and duration of improving quality jointly determine the quality coefficient. The quality sensitivity analysis of the cost and duration of improving quality is shown in Figure 4.2. When the quality coefficient is 0.609, the duration and cost of improving quality are 8 days and 144000 yuan, respectively. When the quality coefficient is 0.7007, the duration and cost of improving quality are 7.2 days and 160000 yuan, respectively [19]. When the quality coefficient is 0.0716, the duration and cost of improving quality are 8 days and 160000 yuan, respectively. When the quality coefficient is 0.8343, the duration and cost of improving quality are 8 days and 176000 yuan, respectively. Therefore, the cost of improving quality has a greater impact on the quality coefficient of sub projects than the duration, which is consistent with the actual situation.

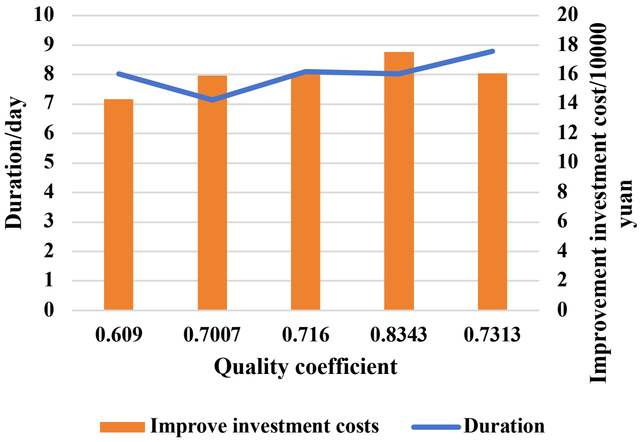


Fig. 4.2: Quality sensitivity analysis of improvement costs and duration

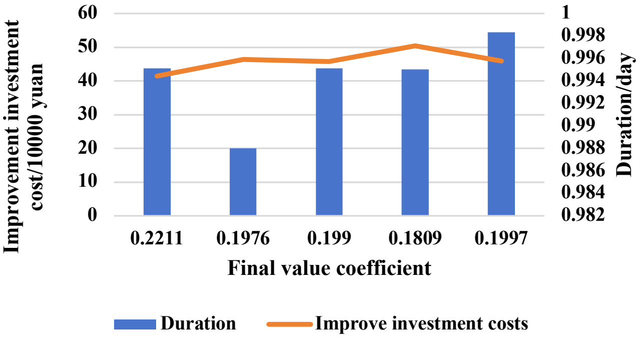


Fig. 4.3: Sensitivity analysis of total expenditure and quality coefficient

The sensitivity results of the total expenditure and quality coefficient to the value coefficient are shown in Figure 4.3 when selecting the projects of the foundation and foundation construction parts. When the value coefficient is 0.2211, the total investment cost and quality coefficient are 412700 yuan and 0.99496 yuan, respectively. When the value coefficient is 0.1976, the total cost and quality coefficient are 456300 yuan and 0.98798 yuan, respectively [4]. When the value coefficient is 0.1990, the total cost and quality coefficient are 456300 yuan and 0.99496 yuan, respectively. Therefore, compared to the coefficient of quality, cost has a greater impact on the value of buildings.

5. Conclusion. The optimization and control of construction project management is currently a topic of common concern among relevant experts and scholars. This study utilizes deep learning algorithms to achieve multi-objective optimization control, this method introduces quality coefficient and value engineering to quantify quality and construct an optimization control model, which is solved using the SULSTM neural network algorithm. The algorithm solving model results show that this method can obtain the optimal parameters for optimal control of construction project management. The loss values of LSTM and two neural network algorithms have the same trend in the first 20 iterations, but within the range of 20-100 iterations, the SUSTM neural network algorithm has faster convergence speed and more stable loss values. Sensitivity analysis shows that the cost of improving quality has a more significant impact on the quality coefficient, and the total investment cost has a more significant impact on the value coefficient. Therefore, the established optimization control model is feasible and practical.

REFERENCES

[1] R. M. Buvanesvari and K. S. Joseph, *An efficient secured pit management and attack detection strategy enhanced by csoa-dcnn algorithm in a named data networking (ndn).*, International Journal of Intelligent Engineering & Systems, 14 (2021).

[2] Y. J. Chen, J.-T. Tsai, W.-T. Huang, and W.-H. Ho, *Intelligent optimization in model-predictive control with risk-sensitive filtering*, Journal of Intelligent & Fuzzy Systems, 40 (2021), pp. 7863–7873.

[3] J. Du, Y. Xue, V. Sugumaran, M. Hu, and P. Dong, *Improved biogeography-based optimization algorithm for lean production scheduling of prefabricated components*, Engineering, Construction and Architectural Management, 30 (2023), pp. 1601–1635.

[4] A. Guo and C. Yuan, *Network intelligent control and traffic optimization based on sdn and artificial intelligence*, Electronics, 10 (2021), p. 700.

[5] Y. Guo, H. Shen, L. Chen, Y. Liu, and Z. Kang, *Improved whale optimization algorithm based on random hopping update and random control parameter*, Journal of Intelligent & Fuzzy Systems, 40 (2021), pp. 363–379.

[6] L. He, Z. Gu, Y. Zhang, H. Jing, and P. Li, *Review on thermal management of lithium-ion batteries for electric vehicles: Advances, challenges, and outlook*, Energy & Fuels, 37 (2023), pp. 4835–4857.

[7] Z. Hou, J. Guo, J. Xing, C. Guo, and Y. Zhang, *Machine learning and whale optimization algorithm based design of energy management strategy for plug-in hybrid electric vehicle*, IET Intelligent Transport Systems, 15 (2021), pp. 1076–1091.

[8] J. Jia, S. Yuan, Y. Shi, J. Wen, X. Pang, and J. Zeng, *Improved sparrow search algorithm optimization deep extreme learning machine for lithium-ion battery state-of-health prediction*, Iscience, 25 (2022).

[9] W. Li, G. Feng, and S. Jia, *Research on multi-energy management system of fuel cell vehicle based on fuzzy control*, Journal of Intelligent & Fuzzy Systems, 40 (2021), pp. 6205–6217.

[10] D. Lin, M. Li, Q. Zhan, X. Song, Y. Yang, and H. Li, *Application of intelligent logistics inventory optimization algorithm based on digital supply chain*, International Journal of Emerging Electric Power Systems, 24 (2022), pp. 61–72.

[11] Z. Liu, X. Zhu, J. Ma, C. Hu, H. Fu, and K. Zhao, *Application of optimization technology for overhaul decision of substation equipment based on machine learning*, in Journal of Physics: Conference Series, vol. 2066, Hangzhou, China, 2021, IOP Publishing, p. 012095.

[12] B. Luo et al., *A method for enterprise network innovation performance management based on deep learning and internet of things*, Mathematical Problems in Engineering, 2022 (2022).

[13] P. Luszczek, W. M. Sid-Lakhdar, and J. Dongarra, *Combining multitask and transfer learning with deep gaussian processes for autotuning-based performance engineering*, The International Journal of High Performance Computing Applications, 37 (2023), p. 10943420231166365.

[14] S. R. Nekoo, J. Á. Acosta, and A. Ollero, *A search algorithm for constrained engineering optimization and tuning the gains of controllers*, Expert Systems with Applications, 206 (2022), p. 117866.

[15] R. K. Prasad and T. Jaya, *Intelligent spectrum sharing and sensing in cognitive radio network by using aroa (adaptive rider optimization algorithm)*, International Journal of Computational Intelligence and Applications, 22 (2023), p. 2341007.

[16] G. Saravanan and N. Yuvaraj, *Cloud resource optimization based on poisson linear deep gradient learning for mobile cloud computing*, Journal of Intelligent & Fuzzy Systems, 40 (2021), pp. 787–797.

[17] H. Tian, C. Tian, C. Yuan, and K. Li, *Dynamic operation optimization based on improved dynamic multi-objective dragonfly algorithm in continuous annealing process*, Journal of Industrial and Management Optimization, 19 (2023), pp. 6159–6181.

[18] Y. Wang, R. Xie, W. Liu, G. Yang, and X. Li, *Modeling and optimization of nox emission from a 660 mw coal-fired boiler based on the deep learning algorithm*, Journal of Chemical Engineering of Japan, 54 (2021), pp. 566–575.

[19] S. Xiong, Y. Zhang, C. Wu, Z. Chen, J. Peng, and M. Zhang, *Energy management strategy of intelligent plug-in split hybrid electric vehicle based on deep reinforcement learning with optimized path planning algorithm*, Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 235 (2021), pp. 3287–3298.

[20] C. Yang, K. Liu, X. Jiao, W. Wang, R. Chen, and S. You, *An adaptive firework algorithm optimization-based intelligent energy management strategy for plug-in hybrid electric vehicles*, Energy, 239 (2022), p. 122120.

[21] D. Zhang, J. Zhao, Y. Zhang, and Q. Zhang, *Intelligent train control for cooperative train formation: A deep reinforcement learning approach*, Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering, 236 (2022), pp. 975–988.

Edited by: B. Nagaraj M.E.
Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing
Received: Oct 5, 2023
Accepted: Nov 14, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1508–1516, DOI 10.12694/scpe.v25i3.2679

A MULTI-LEVEL POWER GRID ENHANCED IDENTITY AUTHENTICATION DATA MANAGEMENT PLATFORM BASED ON FILTERING ALGORITHMS

CAIQIN NONG[*], HONG ZOU[†], JIAFA ZHANG[‡], JIAHAO LIU[§], PENGFEI SHAO[¶], JIAWEI JIANG[‖] AND JIAO XIE[**]

Abstract. In response to the optimal extraction of DCT coefficients in facial images, the author proposes a DCT coefficient extraction method based on discriminant analysis. Based on the discriminant analysis of DCT coefficients, the DCT coefficients with high discriminant values are selected as features. Comparing the DPA based discrete cosine coefficient selection method proposed by the author with the traditional Zigzag discrete cosine coefficient selection method, experiments were conducted on the ORL face database and the Yale face database, respectively. The recognition performance on the ORL face database was higher than that on the Yale face database, as the facial image expression and lighting changes in the ORL database were relatively few, making it suitable for extracting key features. In response to the problem that the speech parameter MFCC is greatly affected by noise and can only reflect the static characteristics of speech, the author extracted gamma pass filtering cepstrum coefficients with human auditory characteristics and gamma pass sliding differential cepstrum coefficients that can reflect the dynamic characteristics of speech based on gamma tone filters and sliding differential cepstrum. In the NUST603 speech database, under pure background, the recognition rate based on GFSDCC features reached 89.88%, and the recognition effect based on GFCC features was 87.52%, which is 4.66% and 2.36% higher than that based on MFCC features. In noisy environments, the average recognition rates of speaker recognition systems based on GFCC and GFSDCC are 56.06% and 59.07%, while the average recognition rates of speaker recognition systems based on MFCC speech features are 53.89%, 2.17% and 5.18% higher, respectively. The gain in this recognition effect comes from the characteristics of the auditory model, as the Gammatone filter effectively reflects the noise resistance of the human auditory system.

Key words: Filtering, Multi level, Enhanced identity authentication, Data management

1. Introduction.

The important significance of identity authentication technology in power information systems is reflected in its ability to ensure the security of the power information system, thereby ensuring the security of the entire power system. The current power information system is an important component of power system automation, including numerous automation equipment [10]. Automation equipment has both advantages and disadvantages. The advantage is that it liberates manpower from heavy labor, requiring only necessary monitoring, and plays a very important role in identifying and troubleshooting faults; But at the same time, it also creates a problem where once serious problems occur, the safety and stability of the entire system cannot be guaranteed, thereby affecting the entire power supply work. The current power information system is not perfect, and its identity authentication technology has not been widely applied in the entire industry, resulting in a series of problems. In the future, urban and rural electricity consumption will inevitably increase significantly, and the stable operation of the power system has become an urgent and important responsibility, which should be taken into account in every aspect. Identity authentication refers to the use of various means and methods to identify the identity of a person who wishes to obtain a certain permission. The identification of individuals mostly relies on visual memory, but if machines rely solely on visual recognition, it may result in significant costs [14, 16]. For example, the startup identity authentication of a certain computer can only be done with a simple password, as its security is not as high. If identification devices such as fingerprints and iris are installed on a regular computer, the cost will greatly exceed the budget.There are generally two types of identity

[*] China Southern Power Grid Digital Grid Group Information and Telecommunication Technology Co., Ltd., Guangzhou, Guangdong, 510670, China (Corresponding Author)

[†] China Southern Power Grid Digital Grid Group Information and Telecommunication Technology Co., Ltd., Guangzhou, China

[‡] China Southern Power Grid Digital Grid Group Information and Telecommunication Technology Co., Ltd., Guangzhou, China

[§] China Southern Power Grid Digital Grid Group Information and Telecommunication Technology Co., Ltd., Guangzhou, China

[¶] China Southern Power Grid Digital Grid Group Information and Telecommunication Technology Co., Ltd., Guangzhou, China

[‖] China Southern Power Grid Digital Grid Group Information and Telecommunication Technology Co., Ltd., Guangzhou, China

[**] China Southern Power Grid Digital Grid Group Information and Telecommunication Technology Co., Ltd., Guangzhou, China

authentication technology: Authentication based on relevant information, such as various passwords, certificates, etc; Authenticate based on relevant human characteristics, such as unique elements such as fingerprints and irises. Compared to the two, the former is simple and feasible, and can be relatively freely and widely used. In other words, as long as you master the password or certificate, you can obtain relevant permissions. However, the drawback of this method is also its low security performance, which can cause great inconvenience when relevant passwords and tokens are lost. The latter has the characteristic of uniqueness and is relatively secure, but its limitation is that authenticated users cannot stay away. For example, once users who have entered fingerprints and iris are on a business trip, the system cannot run smoothly. Extracting effective features of faces and speech is the key to completing facial recognition and speech recognition tasks. Although different features can represent facial images and speech signals, they reflect the different characteristics of faces and speech, and their suitable application backgrounds are also different, therefore, how to choose suitable and efficient feature extraction methods based on application needs, and how to improve and improve the performance of existing feature extraction methods are all worth further research. In response to this research issue, Chuang, C. W. et al. used a YOLOv2 model based on deep learning to jointly label the iris and sclera parts of human visible light images, and trained an identity classifier to infer the correct personal identity. The performance of the system was evaluated through a self-made visible light human eye image database, and the average accuracy (mAP) of the proposed iris sclera joint recognition based on deep learning can reach over 99%. In addition, compared to previous work, this design is more effective without the use of iris and sclera segmentation processes [5]. Braeken, A. et al. proposed the first non trivial IBI scheme with implicit authentication using the Elliptic Curve Curved Fanstone (ECQV) implicit authentication scheme. Compared to traditional identity based schemes, implicit certificate based methods can resist key escrow because trusted authorities only have a portion of the keys, which users use as input to construct their own user keys. According to Girault's definition, the scheme can achieve trust level 3 and requires fewer resources compared to certificateless identification. A corresponding formal security model has been defined, demonstrating the resistance of our proposed solution to simulated attacks. Compared with other Schnorr based IBI schemes, our proposed IBI scheme with implicit authentication outperforms other schemes in terms of storage, computing, and communication efficiency, thus providing a feasible solution for applications in the Internet of Things (IoT) environment [2].

Based on current research, the author first proposes a DCT coefficient selection method based on discriminant analysis from the perspective of selecting effective features for facial feature extraction. Secondly, for speech parameter extraction, a Gammatone filter and sliding differential cepstrum are used, We extracted static speech features based on human auditory characteristics, GammatoneFilterCepstralCoefficients (GFCC), and dynamic speech features, GammatoneFilterShiftedDeltaCepstralCoefficients (GFSDCC).

2. Methods.

2.1. Facial DCT feature extraction based on discriminant analysis.

Discrete Cosine Transform (DCT) is a common time-domain and frequency-domain transform in signal processing, and has been widely used in feature extraction in face recognition [12]. The DCT transform itself does not perform data compression, it only maps the image source data to another domain. How to select the most effective DCT coefficients as recognition features in the new data domain has become a key issue. The traditional DCT coefficient selection method selects low-frequency DCT coefficients as features in rectangular or Z-shaped order, and the extracted corresponding features often do not represent the best discriminative features. From the perspective of selecting effective features, a DCT coefficient selection method based on Discriminate Power Analysis (DPA) is proposed. Firstly, the DCT coefficients of each position in the facial image are calculated for their discriminative power values, and then the DCT coefficients with higher discriminative power values are selected as feature parameters.

(1) DCT coefficient. For an $M * N$ image matrix $f(x, y)$, its discrete cosine transform is defined as:

$$C(u,v) = a(u)a(v) \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \times \cos \frac{(2x-1)u\pi}{2M} \times \cos \frac{(2y-1)v\pi}{2N} \tag{2.1}$$

Among them, $C(u, v)$ is called the DCT coefficient of matrix $f(x, y), u = 0, 1\ M.\ V = 0, 1\ N.$ $a(u)$ and

$a(v)$ are defined as:

$$a(u) = \begin{cases} \sqrt{\frac{1}{2}}, u = 0 \\ 1, \text{ otherwise} \end{cases}$$
$$a(v) = \begin{cases} \sqrt{\frac{1}{2}}, v = 0 \\ 1, \text{ otherwise} \end{cases} \tag{2.2}$$

After DCT transformation, the two-dimensional DCT coefficients of the image form a matrix of the same size as the original image. The low-frequency coefficients are concentrated in the upper left corner of the matrix, which is the slow-moving part of the image. The high-frequency coefficients are concentrated in the lower right corner of the matrix, which is the detail and edge part of the image [19]. Facial feature extraction has two purposes: Firstly, in order to reduce the dimensionality of the image and the computational complexity during classification, and secondly, in order to select the most representative features to improve classification performance. A large DCT coefficient indicates that the frequency component changes significantly in the facial image. If the coefficient is small, it indicates that the frequency component does not change significantly in the facial image.

The traditional DCT coefficient selection methods, such as the rectangular method, the “Z” shape method, or their related improvement methods, are all based on the decision selection method and do not conduct relevant statistical analysis [9, 18]. Although these feature selection methods are simple and may be effective on certain data samples, they cannot guarantee that they are effective for all samples in the entire database. Based on the different discriminative abilities of each DCT coefficient in the DCT coefficient matrix, the author calculates the discriminative ability value of each DCT coefficient at each position based on discriminative ability analysis, with the aim of selecting DCT coefficients with strong discriminative abilities as features.

(2) Identification ability analysis. The DCT coefficient selection method based on discriminant analysis mainly relies on two assumptions: The coefficient has a large inter class variation and a small intra class variation, which can prove the strong discriminant ability of the coefficient [15]. Assuming the DCT coefficient matrix of a face image of size is:

$$X = \begin{pmatrix} x_{11}, x_{12}, \dots x_{1N} \\ x_{21}, x_{22} \dots x_{2N} \\ \dots \\ x_{M1}, x_{M2} \dots x_{MN} \end{pmatrix} \tag{2.3}$$

Assuming that the training samples have a total of C classes and each class has S images, the training samples have a total of $C * S$ images. Therefore, the calculation process of the discriminant ability value $D(i, j)$ of each DCT coefficient $x_{ij}(i = 1, 2 \cdots M.j = 1, 2 \cdots N)$ in the DCT coefficient matrix can be divided into the following steps:

By selecting the DCT coefficients at positions (i, j) in each DCT coefficient matrix, construct the discriminative ability matrix A_{ij}. The number of matrix A_{ij} is $M * N$, and its form is as follows:

$$A_{ij} = \begin{pmatrix} x_{ij}(1,1), x_{ij}(1,2) \dots x_{ij}(1,C) \\ x_{ij}(2,1), x_{ij}(2,2) \dots x_{ij}(2,C) \\ \dots \\ x_{ij}(S,1), x_{ij}(S,2) \dots x_{ij}(S,C) \end{pmatrix} \tag{2.4}$$

Calculate the average value M_{ij}^C for each type of sample:

$$M_{ij}^C = \frac{1}{S} \sum_{S=1}^{S} A_{ij}(s, c) \tag{2.5}$$

Calculate the intra class sample mean difference V_{ij}^C for each class:

$$V_{ij}^C = \sum_{S=1}^{S} \left(A_{ij}(s, c) - M_{ij}^c\right)^2 \tag{2.6}$$

Calculate the average value V_{ij}^W of the mean difference of samples within Class C:

$$V_{ij}^W = \frac{1}{C}\sum_{c=1}^{C} V_{ij}^c \tag{2.7}$$

Calculate the average M_{ij} of all samples:

$$M_{ij} = \frac{1}{S}\sum_{C=1}^{C}\sum_{S=1}^{S} A_{ij}(s,c) \tag{2.8}$$

Calculate the sample mean difference V_{ij}^B for all samples:

$$V_{ij}^B = \sum_{c=1}^{C}\sum_{s=1}^{S}\left(A_{ij}(s,c) - M_{ij}^C\right)^2 \tag{2.9}$$

Calculate the discriminative ability value $D(i,j)$ of position (i,j):

$$D(i,j) = \frac{V_{ij}^B}{V_{ij}^W} \tag{2.10}$$

The larger the discriminant ability value $D(i,j)$, the stronger the discriminant ability value of the DCT coefficient at position (i,j) in the DCT coefficient matrix, indicating that its corresponding DCT coefficient can be selected as a feature parameter [4]. The DCT coefficient selection method based on discriminant analysis is different from previous coefficient selection methods and is a statistical based selection method.

2.2. Static and dynamic speech auditory feature extraction based on Gammatone filter. The human auditory system is an extremely complex perception system. Studying the structure and function of the human ear can not only help us understand the perception process of the human ear, but also greatly assist us in designing automatic processing systems that simulate human ear function. The performance of the human ear auditory system is much more reliable than any automatic speech recognition system [13, 7]. In noisy environments, the speech parameter MFCC is greatly affected by noise and cannot effectively represent speech signals. Moreover, MFCC can only reflect the static characteristics of speech and cannot reflect the dynamic characteristics of speech. The author proposes a Gammatone Filter Cepstral Coefficients (GFCC) based on human ear characteristics based on Gammatone filters. Considering the temporal variation of speech spectrum structure, the author proposes a dynamic parameter for speech, Gammatone Filter Shifted DeltaCepstral Coefficients (GFSDCC), based on the Gammatone filter cepstrum coefficients using sliding differential cepstrum.

(1) Gammatone filtering cepstrum coefficient extraction. MFCC is currently the most commonly used speech feature parameter, among which Mel filter is used to smooth the amplitude square spectrum of speech signals using a triangular filter bank [1]. The author uses Gammatone filter banks instead of Mel filter banks to extract Gammatone filter cepstrum coefficients that can simulate human auditory characteristics. Figure 2.1 shows the extraction process of GFCC. Firstly, the speech signal is preprocessed, followed by Fourier transform of the speech frame to obtain the speech signal spectrum, by using a Gammatone filter, the linear spectral energy is converted into Gammatone spectral energy, and finally its cepstrum value is calculated. That is, the logarithm of the energy is calculated first, and then the discrete cosine transform is performed to obtain GFCC.

(2) Gammatone filtering sliding differential cepstrum. Although GFCC can accurately simulate the auditory characteristics of the human ear and outperform MFCC in recognition performance, like MFCC, it only reflects the static features of speech and does not consider the dynamic characteristics of speech [3]. The differences in people's speech are mainly reflected in the temporal changes in the spectral structure of speech. Shifted Delta Cepstral (SDC) uses a sliding differential cepstrum feature vector composed of several blocks of differential cepstrums spanning multiple frames of speech, allowing one frame feature to contain the acoustic information of multiple frames of speech before and after it, fully reflecting the dynamic characteristics of speech. On the

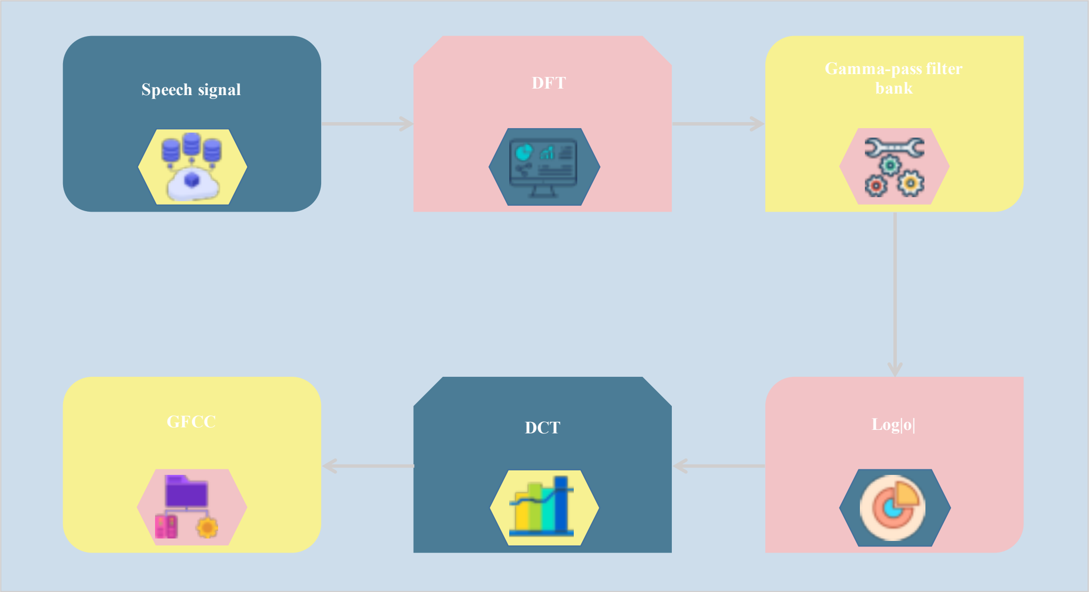


Fig. 2.1: Gammatone filter cepstrum coefficient extraction process

basis of GFCC and SDC, the author proposes a speech feature called Gammatone filtered sliding differential cepstrum coefficient that can reflect the dynamic characteristics of speech and accurately simulate human ear characteristics. By concatenating k-block differential cepstrum, the differential cepstrum is extended within one frame, with each block of differential cepstrum sliding backwards by p frames. During recognition, the author fuses GFCC and GFSDCC in the feature layer to form a fusion feature vector, which not only simulates the auditory characteristics of the human ear, but also comprehensively considers the static and dynamic characteristics of speech features to reduce the impact of external factors on speech signals.

3. Results and Analysis.

3.1. Comparison of different DCT coefficient selection methods. Experimental database: The ORL facial database consists of 400 grayscale facial images from 40 individuals, each with 10 images, and the size of the images is 92 * 112. The background of the image is black, and the facial expressions and details vary, such as whether to smile or not, whether to wear glasses or not, and the facial posture also changes. The depth and plane rotation can reach 20 degrees, and the size of the face can also vary by no more than 10%. In this experiment, each facial image was dimensionally reduced to a size of 46 * 56. The Yale facial database consists of 165 images from 15 individuals, each with 11 facial images, all of which are frontal facial images, with a size of 243 * 320. Facial images have facial expressions, facial details, and changes in lighting. In the experiment, each facial image was dimensionally reduced to a size of 60 * 80.

This experiment will compare the DPA based discrete cosine coefficient selection method proposed by the author with the traditional Zigzag discrete cosine coefficient selection method, and conduct experiments in the ORL face database and Yale face database, respectively. For each individual, the first 5 facial images will be selected as training samples, and the remaining images will be used as test samples. Assuming the training sample is $X_1 = [x_1, x_2, \cdots, x_n]$ and the test sample is $X_2 = [x_1, x_2, \cdots, x_n]$, the Euclidean distance between the two types of samples is as follows:

$$d(X_1, X_2) = \sum^{n} (x_i - x_j)^2 \tag{3.1}$$

Figure 3.1 shows the recognition results of two DCT coefficient selection methods on the ORL face database, while Figure 3.2 shows the recognition results of two DCT coefficient selection methods on the Yale database.

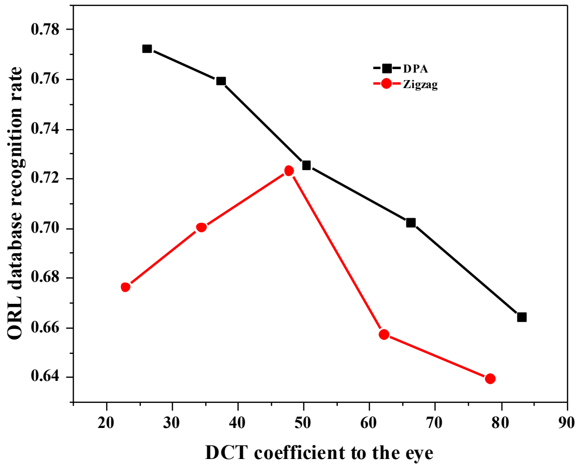


Fig. 3.1: Comparison of recognition rates of different DCT coefficient selection methods on ORL database

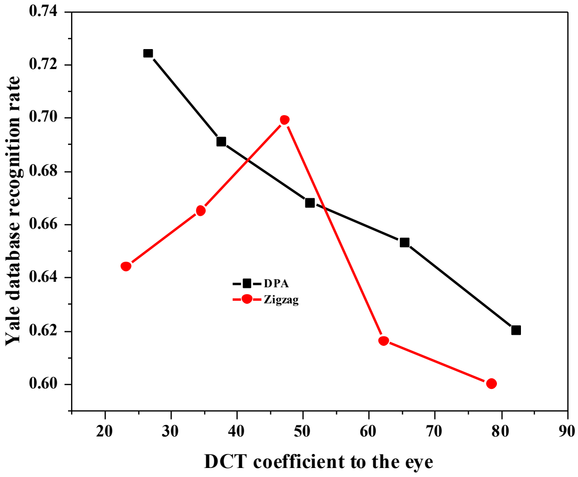


Fig. 3.2: Comparison of recognition rates of different DCT coefficient selection methods on Yale database

From the above two Figures, it can be seen that the DPA method performs better in selecting DCT coefficients than the Zigzag method, regardless of the experimental results in the ORL database or the Yale database. This is because the DCT coefficients selected using the Zigzag method are only low-frequency components in the DCT coefficient matrix, rather than selecting the most discriminative DCT coefficients from the entire coefficient matrix as features for recognition, just like the DPA selection method. From Figures 3.1 and 3.2, it can be seen that the higher the number of DCT coefficients, the higher the recognition rate. When using the DPA method to select DCT coefficients, the highest recognition rate is achieved when selecting 25 coefficients. However, when using the Zigzag method to select DCT coefficients, the best recognition effect is achieved when selecting 49 coefficients. And the recognition performance on the ORL face database is higher than that on the Yale face database, because there are relatively few changes in facial expressions and lighting in the ORL database, which is suitable for extracting key features.

3.2. Comparison of recognition performance of different speech parameters under pure background.

Table 3.1: Comparison of recognition effects of different speech features under pure background

Feature Type	Recognition Rate
MFCC	85.23%
GFCC	87.53%
GFSDCC	89.89%
GFCC+GFSDCC	93.05%

(1) Experimental steps. In order to verify the effectiveness of the speech parameters proposed by the author, the GMM model was used in the NUST603 speech library for validation. The experiment is divided into two parts:

Comparison of recognition performance of different speech parameters under pure background. In a pure background, compare the proposed speech parameters GFCC, GFSDCC, and their fusion features with traditional speech parameters MFCC. Verify the robustness of the speech parameters proposed by the author in different noise environments [6]. Compare the speech parameters GFCC, GFSDCC, and their fusion features proposed by the author with the traditional speech parameter MFCC under the background of White noise and Babble noise with different signal-to-noise ratios.

(2) Experimental database. In this section of the experiment, the author used the NUST603 speech library, which records pure speech in a quiet laboratory environment. The sampling frequency of the speech signal is 22.05KHz, with mono recording and 16Bit quantization. The voice data used in the experiment included 60 speakers, 28 females, and 32 males.

(3) Experimental parameter settings. The speech signal is first preprocessed in the preceding paragraph, using methods based on energy and zero crossing rate for silent detection. Then, a filter with a factor of 0.97 is used for pre emphasis. Then, a frame with a length of 20ms and a frame shift of 10ms is processed, and finally, a Hamming window is processed. Then extract 0-12 dimensional GFCC, totaling 13 dimensions. When extracting GFSDCC features, the selection of its parameter combination N-d-P-k will have a certain impact on the extraction of GFSDCC, among them, N is the number of cepstrum coefficients contained in each frame of speech, d is the time shift for calculating differential cepstrum, p is the sliding frame number of differential cepstrum blocks, and k is the number of differential cepstrum blocks contained in an SDC feature vector. Different parameter combinations have different recognition effects. According to the proposed mountain climbing optimization method, the author adopts a parameter combination of 13-2-3-3, resulting in a total of 39 dimensions of GFSDCC. When fusing the two features in the feature layer, a total of 52 dimension feature vectors are obtained.

In a pure speech environment, test the recognition performance of the GFCC features, GFSDCC features, and their combination features proposed by the author in a GMM model, and compare them with MFCC features. The results are shown in Table 3.1 [8]. From the data in Table 3.1, it can be seen that in pure backgrounds, the recognition performance based on GFCC speech features is better than that based on MFCC speech features, with a recognition rate of 2.3% higher. This is because GFCC features based on Gammatone filter banks have better distinguishability than MFCC features based on Mel filter banks.

GFSDCC features not only utilize the auditory characteristics of Gammatone filter banks, but also incorporate relatively long temporal information into a feature vector, effectively characterizing the dynamic characteristics of language features. In a pure background, the recognition rate based on GFSDCC features reached 89.88%, 2.36% higher than that based on GFCC features, and 4.66% higher than that based on MFCC features [11]. Both GFCC and GFSDCC only reflect one aspect of speech features. The author fused the two in the feature layer, taking into account the static and dynamic characteristics of speech features. In a pure background, the speaker recognition rate based on fused features reached 93.05%, which is 7.82%, 5.52%, and 3.16% higher than the recognition results based on MFCC, GFCC, and GFSDCC, respectively.

(4) Comparison of speech parameter recognition performance in different noise environments. In White noise and Bubble noise environments, under different signal-to-noise ratios, the GMM model was used to test the superiority of the GFCC features, GFSDCC features, and their combination features proposed by the

Table 3.2: Recognition rates using different acoustic features in various speech environments

Feature Type	Phonetic context	MFCC	GFCC	GFSDCC	GFCC+GFSDCC
	SNR				
Babble noise	0dB	34.73%	35.72%	36.65%	38.96%
	5dB	49.04%	50.41%	53.00%	58.85%
	10dB	52.256%	56.14%	59.49%	63.34%
	15dB	61.59%	64.67%	66.25%	71.25%
	20dB	63.57%	65.33%	68.23%	75.31%
White noise	0dB	36.64%	37.18%	44.02%	51.77%
	5dB	49.78%	50.18%	56.2%	60.73%
	10dB	56.14%	58.05%	61.01%	66.63%
	15dB	65.45%	67.28%	68.72%	73.18%
	20dB	69.8%	75.65%	77.25%	80.1%
average value		53.9%	56.06%	59.08%	64.01%

author in recognition, and they were compared with MFCC features. The results are shown in Table 3.2 [17]. From the experimental results in Table 3.2, it can be seen that in noisy environments, regardless of which Gammatone filter based speech feature is used, the recognition performance is higher than traditional MFCC features. In the Babble noise environment, when the SNR is 0, the recognition rate of the speaker recognition system based on MFCC features is only 34.73%, which is almost impossible to use, the speaker recognition rates based on GFCC and GFSDCC are 35.72% and 36.65%, respectively, while the recognition effect based on the fusion features of the two is 38.96%. In the White noise environment, when the SNR is 20, the recognition rate of the MFCC based speaker recognition system is only 69.8%. The recognition rates of GFCC based and GFSDCC based speakers are 75.65% and 77.25%, respectively, while the recognition effect based on the fusion of the two features is 80.1%. Due to the interference of the "cocktail party" effect, the performance of the speaker recognition system under Babble noise is lower than that under White noise environment. Overall, in a noisy environment, the average recognition rate of a speaker recognition system based on MFCC speech features is 53.9%. The average recognition rates of a speaker recognition system based on GFCC and GFSDCC are 56.07% and 59.08%, while the average recognition rate of a speaker recognition system based on the fusion of the two features is 64.01%, the average recognition rates of speaker recognition systems based on MFCC, GFCC, and GFSDCC speech features are 56.07% and 59.08%, respectively.The average recognition rate of speaker recognition systems based on the fusion of the two features is 64.01%, which is 10.11%, 7.94%, and 4.93% higher than that of speaker recognition systems based on MFCC, GFCC, and GFSDCC speech features, respectively.

From the experimental results in Tables 3.1 and 3.2, it can be seen that the acoustic features GFCC and GFSDCC based on Gammatone filter banks proposed by the author outperform the speech features MFCC in both pure and noisy environments. The gain of this recognition effect comes from the characteristics of the auditory model, as the Gammatone filter well reflects the noise resistance of the human auditory system [20].

4. Conclusion. The author mainly explores the optimal extraction of facial and speech features. Extracting effective features of faces and speech is the key to completing facial recognition and speech recognition tasks. Although different features can represent facial images and speech signals, they reflect the different characteristics of faces and speech, and their suitable application backgrounds are also different, therefore, how to choose suitable and efficient feature extraction methods based on application needs, and how to improve and improve the performance of existing feature extraction methods are all worth further research. The author first addresses the issue of facial feature extraction and proposes a DCT coefficient selection method based on discriminant analysis from the perspective of selecting effective features. After performing DCT transformation on the facial image, the DCT coefficient based discriminant ability values for each position in the image are calculated, and the DCT coefficient with the highest discriminant ability value is extracted as a feature parameter. Secondly, in order to address the issue of speech parameter extraction, static speech features based on human auditory characteristics, GammatoneFilterCepstralCoefficients (GFCC), and dynamic speech features, GammatoneFilterShiftedDeltaCepstralCoefficients (GFSDCC), were extracted using Gammatone filter and sliding differential cepstrum. The acoustic features GFCC and GFSDCC based on Gammatone filter banks proposed by the author outperform speech features MFCC in both pure and noisy environments. The gain of this recognition effect comes from the characteristics of the auditory model, as the Gammatone filter effectively reflects the noise resistance of the human auditory system.

REFERENCES

[1] J. M. Balmer and K. Podnar, *Corporate brand orientation: Identity, internal images, and corporate identification matters*, Journal of Business Research, 134 (2021), pp. 729–737.

[2] A. Braeken, J.-J. Chin, and S.-Y. Tan, *ECQV-IBI: Identity-based identification with implicit certification*, Journal of Information Security and Applications, 63 (2021), p. 103027.

[3] M. L. Charter, *Predictors of feminist identity utilizing an intersectional lens with a focus on non-hispanic white, hispanic, and african american MSW students*, Affilia, 37 (2022), pp. 97–117.

[4] J. Chia, J.-J. Chin, and S.-C. Yip, *A pairing-free identity-based identification scheme with tight security using modified-schnorr signatures*, Symmetry, 13 (2021), p. 1330.

[5] C.-W. Chuang and C.-P. Fan, *Deep-learning based joint iris and sclera recognition with yolo network for identity identification*, Journal of Advances in Information Technology, 12 (2021).

[6] R. Clément and B. Norton, *Ethnolinguistic vitality, identity and power: Investment in SLA*, Journal of Language and Social Psychology, 40 (2021), pp. 154–171.

[7] S. Frandsen and T. Huzzard, *Processes of non-identification: Business school brands and academic faculty*, Scandinavian Journal of Management, 37 (2021), p. 101157.

[8] Q. Gao, C. Zhong, Y. Wang, P. Wang, Z. Yu, and J. Zhang, *Defect analysis of the same batch of substation equipment based on big data analysis algorithm*, 651 (2021), p. 022093.

[9] Y. S. B. C. Z. W. J. S. F. Y. Jiangfeng Zhang, Feiyue Wang, *Research on power grid primary frequency control ability parallel computing based on multi-source data*, Acta Automatica Sinica, 48 (2022), pp. 1493–1503.

[10] X. Li, S. Wang, and Z. Lu, *Reverse identification method of line parameters in distribution network with multi-T nodes based on partial measurement data*, Electric Power Systems Research, 204 (2022), p. 107691.

[11] H. Liu, X. Huizhu, D. Dapeng, L. Fengming, S. Jian, M. Haofei, and W. Qiang, *Topology identification of low-voltage transformer area based on improved particle swarm algorithm*, 1972 (2021), p. 012049.

[12] X. Liu, M. Zhang, X. Xie, L. Zhao, and Q. Sun, *Consensus-based energy management of multi-microgrid: An improved SoC-based power coordinated control method*, Applied Mathematics and Computation, 425 (2022), p. 127086.

[13] O. Matuzkova, I. Rayevska, and O. Grynko, *Identification and identity: Differentiating the conceptual terms*, Wisdom, 1 (2021), pp. 44–52.

[14] F. Peng and J. Zhang, *The broken wires identification of wire rope based on multilevel filtering method using EEMD and wavelet analysis*, Journal of Failure Analysis and Prevention, 21 (2021), pp. 280–289.

[15] L. F. Rowe and M. J. Slater, *Will 'we'continue to exercise? the associations between group identification, identity leadership, and relational identification on group exercise class adherence*, International Journal of Sports Science & Coaching, 16 (2021), pp. 670–681.

[16] J. Song, Y. Jiang, X. Song, Z. Sheng, and Z. Meng, *User-transformer relation identification based on power balance model and adaptive AFSA*, 2195 (2022), p. 012042.

[17] B. Wang, F. Zhao, K. Xu, T. Wen, and L. Jiang, *An efficient multi-parameter synchronous identification method for fiber-reinforced laminated structure based on improved levenberg–marquardt algorithm and modal data*, Journal of Vibration Engineering & Technologies, 11 (2023), pp. 2505–2525.

[18] Y. Wang, W. Zhang, D. Huang, and Y. Liu, *Multi-level feature fusion and multi-loss learning for person re-identification*, Signal Processing: Image Communication, 94 (2021), p. 116197.

[19] L. Yang and T. Huang, *A vehicle reidentification algorithm based on double-channel symmetrical CNN*, Advances in Multimedia, 2021 (2021), pp. 1–6.

[20] Z. Zhang, Z. Wang, H. Wang, H. Zhang, W. Yang, and R. Cao, *Research on bi-level optimized operation strategy of microgrid cluster based on IABC algorithm*, IEEE Access, 9 (2021), pp. 15520–15529.

Edited by: B. Nagaraj M.E.
Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing
Received: Oct 8, 2023
Accepted: Nov 25, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1517–1523, DOI 10.12694/scpe.v25i3.2699

ENERGY OPTIMIZATION OF THE MULTI-OBJECTIVE CONTROL SYSTEM FOR PURE ELECTRIC VEHICLES BASED ON DEEP LEARNING

BUBO ZHU*

Abstract. Advancements in information technology have revolutionized multiple sectors such as healthcare, industrial control, and environmental monitoring. With the advent of smaller, more sophisticated and wireless sensors, their applications have expanded across various industries. These sensors offer numerous advantages like cost-effectiveness, easy setup, reliable transmission, and high capacity for data processing. However, despite their benefits, there are certain limitations to consider. The primary constraint that affects their lifespan is energy availability, as replacing or recharging power sources for nodes can be challenging or infeasible. The reliance on batteries hampers data analysis by network nodes, hindering the exchange of information. Hence, prolonging the network's overall lifespan is crucial for optimizing its performance. The existing approaches, with their tried-and-tested practices and heterogeneity, require enhancements to address specific characteristics. In every application, two critical aspects are the duration of network operation and energy consumption for data routing. Through comparative analysis, it is evident that various algorithms and techniques can reduce energy usage to different extents. Based on these findings, a recommended strategy is to achieve a significant 70% reduction in energy consumption.

Key words: Aggregated data energy balance, mobile detector, info linkage.

1. Introduction. Sensors are compact devices with communication and interaction functions, playing a crucial role in various application fields. They are usually equipped with WiFi communication components, allowing them to not only sense the surrounding environment, but also transmit data and interact with other devices. This versatility makes sensors widely used in monitoring, control, and data acquisition, providing critical support for modern technology and automation systems [4]. However, due to the limited power capacity of each sensor, managing their energy consumption becomes crucial, especially considering that these sensors often operate in remote or challenging environments, such as frontline areas or vacant plots. Consequently, it becomes necessary to replace the batteries of sensors located in significant local regions where numerous sensor nodes are densely deployed (up to twenty nodes per square meter). Developing an approach that can adapt the local sensor infrastructure without compromising the overall system's performance becomes vital, considering the aforementioned attributes of sensors. The majority of the plans do, however, take strength preservation into consideration. The direction-finding task is then formulated as a simple coding issue [7], and a cost-directing set of rules is then given, mostly based on link pricing. The resilience of the sensor network and the boundaries of what may be observed are the sole foundations of the suggested architecture for data alliance security. It will investigate which channels the MPMC rule set uses in a specific area. The results of the tests indicate that the placement of this rule is least comfortable with the sum of standard data.

Production and energy consumption are significant factors in domestic and global strategic choices. The short- and long-term sustainable development as it is intended in various countries must be closely monitored [8]. The meta-heuristic algorithm is one of the cutting-edge techniques and algorithms used in this prediction. Meta-heuristic algorithms can minimise errors and standard deviations when processing data. It is possible to analyse uncertainty and find any defects in the datasets using a variety of statistical techniques. Both the exponential and linear models employ each technique. The extent of error is about 3.7%. The winning model predicts that by 2030, global energy consumption would be at 459 terawatt-hours. Electricity-producing industries may be able to make erroneous predictions about future energy use thanks to meta-heuristic algorithms. To reduce this inaccuracy and produce a more accurate prediction, researchers should use various methods. Efficient energy management is essential for a nation's growth and development. Especially in the twenty-first century,

*Shaanxi College of Communication Technology, School of Automotive Engineering, Xi'an, Shaanxi, 710018, China (Corresponding Author, BuboZhu2@163.com)

electricity is one of the most significant energy sources. One may forecast electricity generation by calculating the output of existing power plants and speculative development projects [15].

Predicting power consumption, however, is a very challenging situation. There are both conventional and cutting-edge techniques for forecasting electricity usage. Prediction errors have been reduced thanks to contemporary techniques like neural networks and meta-heuristic algorithms. This article examined the IRO, CBO, and ECBO meta-heuristic algorithms, and six models were created. Two linear and one exponential mode were among the three methods used in these models. Calculations were based on the linear model of the ECBO algorithm, which had the smallest error of the six models. The Mean Absolute Percentage Error (MAPE) for the winning model was 3.7%. The trend of increasing power usage was estimated through 2030. 2030 will see the use of electricity reach 459 terawatt-hours. Four socioeconomical criteria have been taken into account to predict power usage. These four elements are the GDP, the population, the cost of electricity, and the consumption rate from the previous year. It is possible to plan for the country's power grid and important developments by using meta-heuristic algorithms and reducing the forecast error. By contrasting the results of several meta-heuristics, researchers might reduce this error level.

The authors of this study advise managers in the power generation sector to employ cutting-edge forecasting techniques like neural networks and meta-heuristic algorithms. Using unique meta-heuristic algorithms to address the proposed scheduling problem is advised since precise solution methods. At the same time, they can generate optimal answers in more execution time and need help analyzing large-scale issues and solving them logically. Additionally, MOKSEA algorithm performs substantially better than MOKA. Moreover, MOKSEA is acknowledged as the top approach for solving problems in the RAS, SNS, SM, and NPS indices. This study has offered some valuable managerial insights. One of the most crucial elements is that figuring out the best production schedule is very challenging. Additionally, there are frequent updates to this schedule. Therefore, finding a suitable solution quickly and within reason is essential by utilizing novel approximation techniques while considering numerous constraints [5, 1].

2. Literature Survey. According to the research results of coverage evaluation standards, the proposed technology shows significant advantages in terms of coverage, increasing by 12% compared to the FGOA method, 15% compared to GOA, and 16% compared to GSO. These data indicate that this technology has the potential to improve network coverage and can bring better performance to IoT systems. The Internet of Things (IoT) is a complex heterogeneous system that combines various communication technologies with data recording programs to collect, transmit, analyze, and store data. In the Internet of Things, edge nodes, such as RFID tags or sensor nodes, collect data through the network layer and transmit this data to customers or service providers. To achieve this goal, effective allocation of resources is crucial. By improving the performance of edge nodes and improving the service level provided by the network layer, the overall service cost can be reduced. This is crucial for the sustainability and service quality of the Internet of Things. However, further research is still needed to address the issues of network durability and service quality related to the Internet of Things. This will help ensure the stable operation of the Internet of Things system and provide high-quality services to meet the growing demand [14].

Most Internet of Things applications use distributed nodes with restricted power sources. As a result, it is urgently necessary to develop new techniques to stop energy loss, which reduces networks' lifespan. Due to these restrictions and the high network node density, designing and managing wireless networks has become difficult. All layers of the network protocol stack now require energy awareness. For instance, we urgently need to determine how to employ energy efficiency at the network layer to select paths and transmit data. IoT routing has grown in relevance and significantly impacts lowering energy usage, making it a significant research challenge. Energy-efficient routing is one method for reducing the amount of energy needed by selecting the best path. Three criteria were used to assess the effectiveness of the suggested method: network life, coverage rate, and residual energy [3].

Calculated optimization can be an effective way to reduce power consumption when working with low-energy buildings. This paper provides a strong combination strategy based on the bird of paradise optimization algorithm (POA) and one contender optimizer (SCO) to address issues with building energy optimization. The suggested hybrid algorithm (POSCO) makes use of the local solid search capability of the single candidate method and the efficient global search capability of the pelican upgrade. To optimize the building, the optimization technique was developed and integrated with the most recent edition of the Energy codes [13, 11].

The findings show that the POSCO technique outperforms a few cutting-edge methodologies and reduces building energy consumption at specific temperatures and lighting conditions. POSCO is contrasted with other algorithms, such as basic POA. In light of the information, The findings of the building energy optimization procedure for various climates demonstrate that modifications to the meteorological data did not considerably impact the efficiency of the process. This work presented a hybrid optimization technique based on the single candidate optimizer (POSCO) and pelican optimization to evaluate buildings' most minimal energy use. The strong exploratory capability of pelican optimization and the efficient local search capability of the single-candidate approach is used in the suggested methodology. Several unimodal and multimodal benchmark functions are used to evaluate the performance of the proposed method.The results demonstrate that POSCO outperforms conventional POA and other techniques in identifying the overall solution. The suggested solution outperformed other approaches in determining the global best for seven of the thirteen functions taken into account. It also resulted in better outcomes for the other functions. Each optimizer's performance has been unaffected by the change to the weather file. POSCO is a viable candidate method for BEO models because, in accordance with the outcomes of the competition simulation, it can accurately and dependably forecast the best design.

3. Materials and Methods.

3.1. Description of the genetic algorithm. The biological algorithm (GA), an evolutionary efficiency technique, has been successfully used in engineering. In this organized yet randomized search, mutation, crossover, and selection are handled by genetic operators. The core concept of GA includes the following key points: 1.Population: Firstly, GA will create a population consisting of multiple individuals, each representing a potential solution. 2.Selection: GA uses selection operations to evaluate individual adaptability, usually measured through a fitness function. The adaptability function evaluates the quality of each individual's solution in the problem space. Then, based on the size of adaptability, individuals who are more conducive to problem-solving are selected as parents. 3.Crossing: Selected individuals (parents) can generate new individuals (children) through crossing operations. Cross operation simulates the process of gene recombination in biology, by combining the chromosomal parts of two or more individuals to generate new ones. 4.Mutation: In some cases, GA may introduce mutation operations to introduce new randomness into the population. This simulates the process of gene mutation in biology, which helps to explore a wider range of problem spaces. 5.Iterative evolution: GA repeats the above steps for multiple generations, each generation generating new individuals. As algebra increases, better solutions have a higher chance of being preserved and passed on to the next generation. Let's look at a couple of jargon. A term that refers to genetic algorithms. The first suggested solution to the issue involves chromosomes. The genes or alleles on each chromosome must be the same size. Selection. The fundamental genetic process transfers genomes with greater quantity to the following generation. Position, constant state, and the gambling ball are other selection strategies in addition to elitism. Any selection method could be used, depending on the requirements of the application. Crossover. When a pair of adult genomes is chosen for bridging and some of their DNA is transported across, the chromosomes of the offspring are created [16]. 100,000 | 001,000 on the first chromosome.

Chromosome 2:... 000,100 | ... 001...

1... 100,000 | 000,001 offspring

2 offspring were born, 000,100 | 001,000....

How exercise serves a purpose. Chromosomes with more excellent fitness scores would result in more offspring than those with fewer points, according to the fitness function, which estimates these values. This article's health rating is the sum of every one of its components multiplied by the stated component proportion. Mutation. The chromosomes may change soon after crossing. The steady progress of the GA method is abruptly stopped. It is used to research solutions on a different website instead of seeking the most recent, best options.

... 10,001,000 ...

↓ mutation ...

00,010,001 ....

Because using complex genetic operators would make running the program more challenging, the suggested method avoids doing so.

(1) Start a group of people at random.
(2) Evaluate the original Solution using the fitness function.
(3) Go on as long as (1) is true
(a)Apply the elitist selection algorithm
(b) use person-point crossings
(c) use a specific mutation rate
(4) The number of individuals will be updated as more children are born.
(5) Conclude
(6) Form the cluster around the chromosome that fits the best.
GAECH is the first algorithm.

4. Experimentation and Results. The suggested installation approach has been implemented on a machine with a Core (TM) i5, 2.7 MHZ, 16 GRAM, and the latest version of Windows as the platform to demonstrate its viability and efficiency. As already stated, the target region is conceptualized as M N grid points. According to established principles, energy equilibrium and regulated boundaries collect various forms of data and create numerous bunch structures [10]. Its main objective is to significantly increase the processing, storage, and power of sensor layer hubs. I'm in charge of gathering the useful material for the sensor layer. Extension: Since each sensor layer corner covers multiple interest hubs, the ground unit receives information related to the intersection of interest features in a box. It is quite well-structured. Alternatively, you will see a list of several tasks between a specific objective hub and numerous monitoring centres. When the organization is prepared, these arrangements are appropriate. Despite this, there were better sites for the continued structure due to changes in network activity and the environment than this initial beginning strategy. Every hub in the organization has its power RPu set to 3 mW. This is 8 40t 2.0 10 regarding the functional relationship between communication power and sends distance [12]. A maximum communication intensity of 250 million watts and a maximum transmission distance of 20 m are both configured for the hub. Every seat in organization sends packages at four bundles per second once reproduction has started. The working principle of GA is shown in Figure 4.1.

Target hubs are arbitrarily selected from the group until the reenactment is through. Set the information transmission rate 6R = 10 pieces/sec and the size of the information bundle sent from the hub to L = 512 bytes and 512 bytes, respectively. As a result, the energy used by the transmitting hub u and the receiving hub v for each information packet delivered is P (u) * L/R and *RPvLR. The ND parcel's transmission time in the recreation is 4 seconds. The hub values for a 7-hub heterogeneous remote sensor organization. Every organization hub for the heterogeneous organizations transmits information with the highest possible communication power, independent of geography. The organization hub consumes power very quickly, and the power hub's excess energy is less than 100mJ. A few seats in the company are also rashly consuming a lot of force due to the unevenness in hub power use [6]. Nevertheless, this represents the cost incurred to modify the hub's energy usage, which lengthens the organization's existence.

4.1. Fitness performance. The fitness of biological inheritance, a measure of a person's survival and reproductive potential, determines their capacity [2] to pass on their DNA to others in a particular group. The genetic algorithm is used to evaluate each group member according to fitness. There is a purpose to wellness. The output from the input person can be the worth fitness has for the person in question.

4.2. Genetic programming. Genetic Programming (GP) is an evolutionary algorithm used to automatically develop computer programs to solve specific problems. Similar to genetic algorithms, genetic programming seeks solutions to problems by simulating the process of biological evolution, but its goal is to generate computer programs, not just optimize parameters. The biological evolution of animals is a lengthy and complex process that involves incremental optimization as lower species gradually develop into higher ones. The driving element behind the process is natural selection. Optimization techniques include copies, hybridizations, mutations, selection, and other operations. Academics have devised genetic algorithms (GA), which are based on the principles of biological evolution. The evolving algorithm offers a unique method of overcoming difficulties in search [9]. A genetic algorithm is made comprised of the three basic genetic operators of crossover mutation, selection, and inheritance.

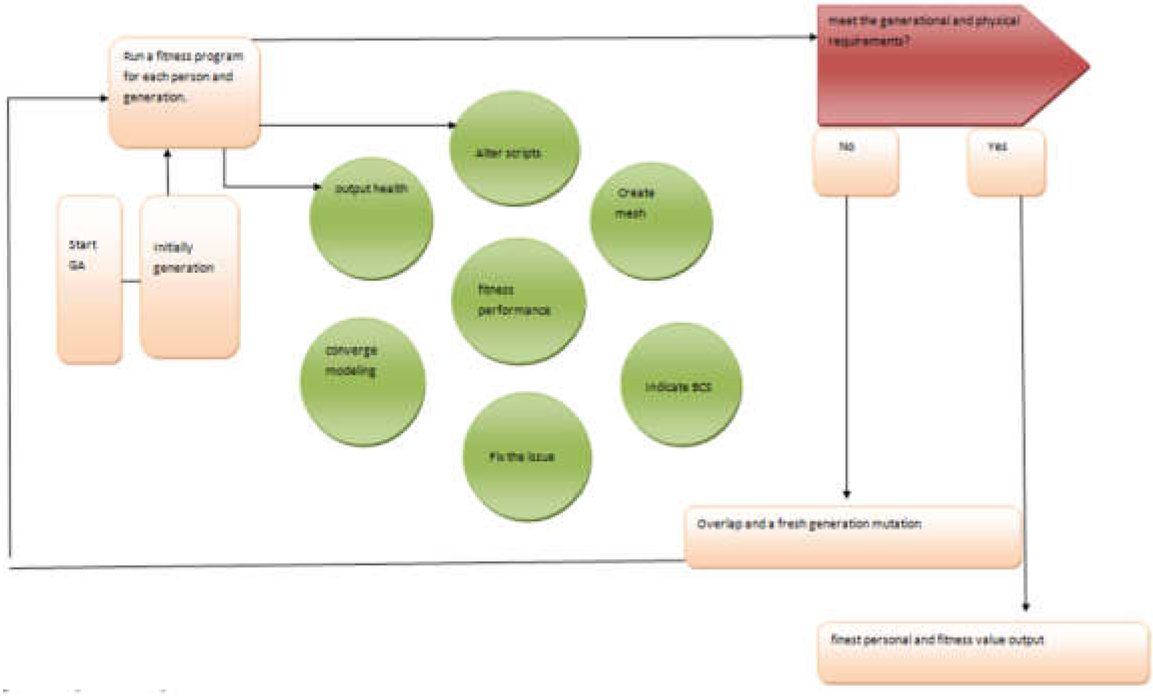

Run a fitness program for each person and generation.
meet the generational and physical requirements?
Alter scripts
output health
Create mesh
No
Yes
Start GA
initially generation
fitness performance
converge modeling
Indicate BCS
Fix the issue
Overlap and a fresh generation mutation
finest personal and fitness value output

Fig. 4.1: Working of GA

4.3. Genetic algorithm mathematical model for best solution. By choosing people or genomes in the shape of Formula from an interval, we will apply the computational method suggested by Michalewicz [12] to get the biggest value of Eq. (4.1).

$$b_1, b_2, \cdots, b_{n_2} = (\sum n(i = 0) b_{2i}) = X_i \tag{4.1}$$

There will be $A = 80$ chromosomes in this case, and every b-5 one will stand for an amino acid on one of them. Next, we convert x' into a number between 0 and 1, as shown in Eq. (4.2).

$$X_1 = L_{iow} + x_1 \frac{L_{up}}{2N - 1} \tag{4.2}$$

Chromosome with the best performance is 0110111100001110101010011000110011.
(3) Equation (4)'s best chromosomal codifies the answers to the previous equations. (2) and (3) as
$x = 0.404522352901455 97$ and
$y = 2.072061602212456$ for performance or fitness.

4.4. Schematic of a genetic algorithm. The crucial steps for the solution are as follows: Before turning the study's parameters into chromosomal code chains, identify the coding scheme used; the answer to each problem correlates to a signal string. 2) Initialization: Create a randomly chosen initial group with a population dimension P and a list of resolvable optimization problems. 3) The measure of fitness has been created, and the predicted fitness level for each person in the population is available. 4) Choose the action: The number of times an individual regenerates are determined by the fitness function chosen, and the optimized person is then included to the following cohort. Add new spouses or generations that have migrated to the following era. 5) cross: The preceding generation of the process is changed by removing two randomly chosen individuals from its organizational structure or content exchanger component. 6) Variation: The random chromosomal genes in the population are altered. Following selection, cross, and mutation procedures, group P will transform into a new

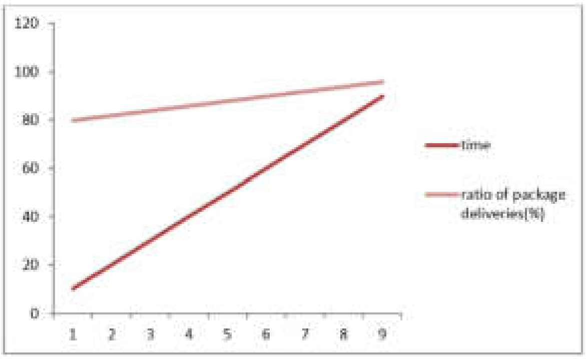


Fig. 4.2: Time taken for packet delivery within sensor node

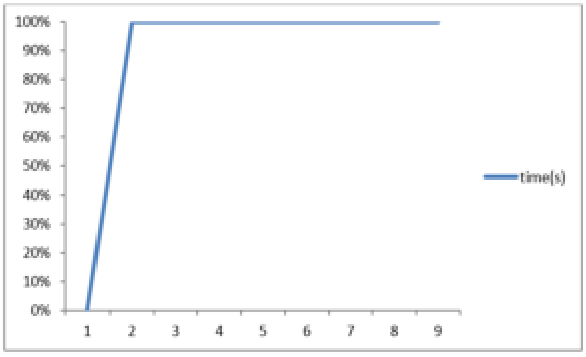


Fig. 4.3: Simulation outcomes of sensor nodes

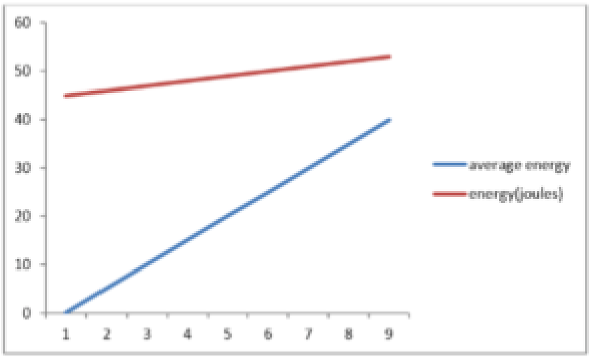


Fig. 4.4: Determining average energy of nodes as per GA

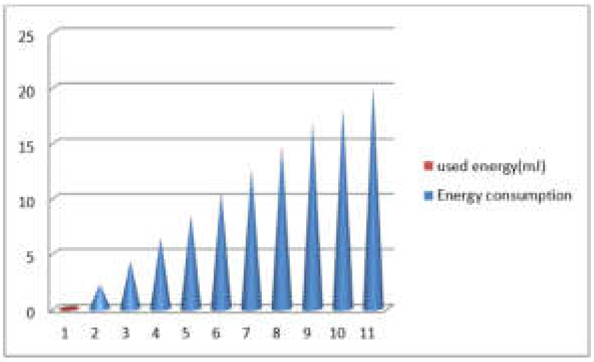


Fig. 4.5: Overall energy used by sensor node

group P based on a predefined probability of mutation Pm.17) Reiterate: If the best solution is identified, the process should be finished. If not, repeat step 3 to reassess, pick, combine, and manage the following groupings. This procedure is carried out again until the best fitness provider is found.

4.5. A crossover with sensor nodes. Selecting the optimum cross-operation phases to get to the fastest solution can speed up genetic processes. A new route is required since crossing in a genetic algorithm cannot change the initial good cluster path. Assume that accessing two current pathways from A to B is possible. The new path, which incorporates the benefits of the two ways, has been optimised in comparison to the two approaches employed in the previous generation because it is the common junction of the two paths mentioned above [6]. As shown in Figure 4.2, Figure 4.3.

4.6. A change of genetic algorithm with sensor nodes and packet transfer. A few gene loci on each specific chromosome sequence are altered as the main objective of the mutation operation. The strategy of the study, which also incorporates the substitution procedure, facilitates the cross-over operation. The final answer will be quite close to the ideal as a result, and a new information path called Ak2k4k5ki... B will appear. On this information path, the main operations will involve bridge auxiliary work and cross operation. If the chromosomes are modified to form a cluster k4, these changes will occur in k4 and have a corresponding impact on k6. In addition, it is expected that the mobile phone sink will use rechargeable technology, and it is unlikely to malfunction during testing. These detailed plans and operations will help ensure the smooth achievement of goals. As shown in Figure 4.4 and Figure 4.5.

5. Conclusion. In the article, a unique approach based on the Genetic Algorithm was developed to maximize area coverage. Using fewer randomly distributed data gathering nodes, this technique was created to increase the network's coverage. When selecting the initial population, it is critical to consider the diversity of the groupings, the calibre of the participants, and the likelihood that each group would succeed. In particular, competitiveness and genetic algorithms investigate these issues. It also shows that it is more dependable and stable than alternative methods. The simulation results evaluated the effectiveness of the suggested paradigm in terms of improved coverage and reduced network expenses. However, further work will be required to put the suggested strategy using a probabilistic detection model into practice.

REFERENCES

[1] P. Ajay, B. Nagaraj, J. Jaya, et al., *Smart spider monkey optimization (ssmo) for energy-based cluster-head selection adapted for biomedical engineering applications*, Contrast Media & Molecular Imaging, 2022 (2022).

[2] J. Anand, J. J. Tamilselvi, and S. Janakiraman, *Analyzing the performance of diverse leach algorithms for wireless sensor networks*, Int. J. Advanced Networking and Applications, 4 (2012), pp. 1610–1615.

[3] C. C. Da Ronco and E. Benini, *A simplex crossover based evolutionary algorithm including the genetic diversity as objective*, Applied Soft Computing, 13 (2013), pp. 2104–2123.

[4] G. Han, Y. Dong, H. Guo, L. Shu, and D. Wu, *Cross-layer optimized routing in wireless sensor networks with duty cycle and energy harvesting*, Wireless Communications and Mobile Computing, 15 (2015), pp. 1957–1981.

[5] G. Han, L. Liu, J. Jiang, L. Shu, and G. Hancke, *Analysis of energy-efficient connected target coverage algorithms for industrial wireless sensor networks*, IEEE Transactions on Industrial Informatics, 13 (2015), pp. 135–143.

[6] K. C. Rahman, *A survey on sensor network*, Journal of Computer and Information Technology, 1 (2010), pp. 76–87.

[7] T. Rault, A. Bouabdallah, and Y. Challal, *Energy efficiency in wireless sensor networks: A top-down survey*, Computer Networks, 67 (2014), pp. 104–122.

[8] A. Sharma, S. R. Kawale, S. P. Diwan, D. Gowda, et al., *Intelligent breast abnormality framework for detection and evaluation of breast abnormal parameters*, in Proceedings of the International Conference on Edge Computing and Applications, Tamilnadu, India, 2022, IEEE, pp. 1503–1508.

[9] A. Sharma, S. Reddy, P. S. Patwal, D. Gowda, et al., *Data analytics and cloud-based platform for internet of things applications in smart cities*, in Proceedings of the 2022 International Conference on Industry 4.0 Technology (I4Tech), Pune, India, 2022, IEEE, pp. 1–6.

[10] S. Tyagi and N. Kumar, *A systematic review on clustering and routing techniques based upon leach protocol for wireless sensor networks*, Journal of Network and Computer Applications, 36 (2013), pp. 623–645.

[11] S. Venkadesh, G. Hoogenboom, W. Potter, and R. McClendon, *A genetic algorithm to refine input data selection for air temperature prediction using artificial neural networks*, Applied Soft Computing, 13 (2013), pp. 2253–2260.

[12] S. Wu, *A traffic motion object extraction algorithm*, International Journal of Bifurcation and Chaos, 25 (2015), p. 1540039.

[13] M. Zbigniew, *Genetic algorithms+ data structures= evolution programs*, Comput Stat, (1996), pp. 372–373.

[14] X.-F. Zhang, G.-F. Sui, R. Zheng, Z.-N. Li, and G.-W. Yang, *An improved quantum genetic algorithm of quantum revolving gate*, Computer Engineering, 39 (2013), pp. 234–238.

[15] D. Zhao, H. Ma, and L. Liu, *Analysis for heterogeneous coverage problem in multimedia sensor networks*, in Proceedings of the IEEE International Conference on Communications, Kyoto, Japan, 2011, IEEE, pp. 1–5.

[16] M. Zhao, Y. Yang, and C. Wang, *Mobile data gathering with load balanced clustering and dual data uploading in wireless sensor networks*, IEEE Transactions on Mobile Computing, 14 (2014), pp. 770–785.

Edited by: B. Nagaraj M.E.
Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing
Received: Oct 16, 2023
Accepted: Nov 25, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1524–1532, DOI 10.12694/scpe.v25i3.2700

OPTIMIZATION OF UNMANNED AERIAL VEHICLE FLIGHT CONTROL SENSOR CONTROL SYSTEM BASED ON DEEP LEARNING MODEL

JI LIU*

Abstract. Based on data modelling strategies have created reliable classifier designs for various classes and other neural network applications. The fact that modelling complexity rises with the total number of groups in the system does is one of the approach's major shortcomings. No matter how well it performs, it could make the classifier's design ugly. This article discusses the development of a novel, logic-based Optimum Bayesian Gaussian process (OBGP) classifier to reduce the number of separate empirical models required to accurately detect various fault types in industrial processes. The precision of the OBGP classifier's defining faults also contrasts with the results of other approaches documented in the literature.

Key words: Classifiers with multiple levels, Fault identification and diagnosis, Regression with the Gaussian process, Ratio of generalized likelihood, Utilizing Bayesian analysis.

1. Introduction. Improvements in device learning and statistical methods allow the creation and execution of exact data-driven recognition and predictive models for several complicated, multifaceted programs, such as the treatment of wastewater or thermal power plants.When developing an accurate equation is challenging or prohibitively expensive, this is quite advantageous. Thus, data techniques are driving the procedure sector to make significant profits, particularly in the isolation and diagnosis of errors. Due to their straightforward design, ease of understanding and quick improvement, principal component analysis algorithms are disproportionately appealing for multidimensional applications [2]. They can also handle enormous numbers of numerical samples at a relatively low computing cost. Multiclass classification, a well-known area of research in machine learning, tries to provide an architecture capable of properly identifying several operating modes for the system in question. Because it can be hard to encompass every potential state available for a particular framework on execution, the challenge is frequently constrained. Although the fact that the linear layout of the IPCA encoder makes this result highly encouraging, there is a crucial caveat: its framework depends on the creation and contemporaneous deployment of several separate IPCA models among every single combination of categories. As a consequence, the amount of simulators and the various types of defects to be found rise rapidly [3]. The bare minimum of one IPCA model has been generated to create the proper conditional label for every particular bundle of erroneous. Upon determining the ultimate choice, each sample's category is separately estimated based on the result generated by every boolean learner in the design. This several-classes classification method is bipolar reduction or multiplication . It is usual to use two methods for theories can be applied [4]. OVA and AVA constitute two competing approaches. The outcomes of all the binary classifiers that comprise the overall several classes classifier are averaged by every approach to arrive at the ultimate result. A combination of assessments utilizing both practical and conceptual programs, the methodology above is shown to be no less than as precise as the OVA approach but using more processing resources. As a result, only the AVA technique will be employed in this study [5]. Depending on the use case, a classifier with multiple alternative models may provide the required autonomy. It also makes efficiency optimisation exceedingly challenging because the best classification performance can only be attained by manually improving each model. As a result, the main objectives are to develop an accurate classifier for multiple classes using the smallest theoretical models and an evaluation approach that improves the learner's classification skills using logic-based (i.e., model-free) parameters or rules [6]. The multiplex GP (MVGP) model used in the current study uses GP systems to forecast numerous outcome variables. In our indicated multi-class classifier (the Optimized

*Department of Electronic and Communication Engineering, Shanxi Polytechnic College, Shanxi 030006,China (Corresponding Author, JiLiu731@163.com)

Bayesian- Gaussian process, or OBGP), just one MVGP algorithm trained on a free-of-defects data class will be utilized. This MVGP algorithm's outputs are then queried with the help of an AVA binarization in order unit and identified with a reasoning-based decision scheme, accomplishing the objective above and lowering. In multiclass classification using GP models, the probability distribution of the expected results can be used to figure out the final class of information that is relevant [28, 7]. The proposed OBGP classifier intends to provide an organised technique for searching residual space utilising the more reliable GP models in order to manage the challenges connected with its application for challenging multimodal processes in the industry, such as the TE process. The format of the questioned paper is as follows. Historical data and previous studies are given in Section 2 to help understand the OBGP's architecture. The construction of the suggested OBGP classifier, the effectiveness of the MVGP approach, and the architecture for rationale-based choice are all covered in Section 3. In Section 4, the classification efficacy of the OBGP predictor's deployment to the TE process is contrasted with that of the IPCA learner and other research-related techniques. Part 5 contains concluding remarks, which is the final portion.

2. Literature Survey. The working atmosphere's volatility frequently leads to poor performance for a probabilistic PID control with a feedback system, which fails to satisfy the profit requirements. Outside noise, determining noise, and additional noises frequently occur in work environments. This study additionally considers and assesses the performance degradation brought on by Stochastic and non-Gaussian disturbances and measurement noise on a stochastic PID feedback system. The dynamic data reconciliation (DDR) technique has been invented for removing measurement noise and disturbances [8].

The results show that DDR has a stronger positive effect on output quality. In the traditional PID feedback control system, monitoring system performance is essential to maintaining optimum profitability. Regrettably, noisy data and distributed Gaussian/non-Gaussian disturbance are frequently observed in control systems.Since simulation results and study demonstrate that the offered method may filter Poisson and pseudo-Gaussian impediments. The term "positive effect" refers to elements of operation analysis such as variance, MSE, IAE, overshooting, and greatest tracker error, to name a few. This DDR filter reduces voltage output deviations for the DC-AC conversion case study when the structure has sinusoidal or faux-Gaussian disturbances. This shows how DDR could increase control effectiveness [9].

A poor strategy for distributed NMPC is described, using Poisson process representations of the motion of linked subsystems and taking account of the given constraints. The approach suggested relies on successive regression of the complex dynamics of the system, global iterations of the two-step accelerated slope method, and a poor solution to the resulting Quadratic Programming (QP) problem. The spread method has several advantages: It features a simple program design that allows the subsystems to calculate the unsatisfactory control inputs separately without centralized adjustment. The recommended method is illustrated through simulations of a basic sewer network concept. An optimal strategy for global GP-NMPC has been suggested based on the motion of the interrelated Gaussian process models of the systems. Due to the ease of its computer program implementation and the ease of performing the internet-based calculation, it is appealing for usage as an incorporated controller. Simulations of the sewer system's model show that the networked GP-NMPC technique produces realistic routes with acceptable suboptimality. Future innovations and their applicability to complex systems are intended [10].

Cybersecurity is a worry for governments and companies worldwide, but much is known about prospective laws that may be taken to stop and lessen risks to companies. In order to further enhance their execution and evaluation, it is essential to understand the efficacy of preventive tactics and policies. The study examines whether carrying out the suggested precautions is linked to more secure company conduct and whether the UK government's "Cyber Essentials" and "10 Procedures to Cybersecurity" initiatives, which encourage and support businesses to adopt security controls as well as policies, are associated with a reduced incidence of cybercrime harassment and its effects. With Bayesian network smoothing. The results indicate a link between improved computer security practises and knowledge of governmental activities. It has not been demonstrated that implementing the advised safety measures will decrease the likelihood of assault or damage to businesses [11]. Discussion is had regarding how the findings might be applied in practise, in policy, and in future studies. Despite the number of victims and the significance of assessments to understand what minimises cyber events and how they affect society, there are very few research on the effectiveness of legislative measures for mitigating cybercrime. Given the substantial commitments that governments and organizations are making to increase security online, Dupont (2019:513) finds it concerning that more mathematically rigorous attempts are required to identify which initiatives and initiatives are delivering tangible enhancements to the safety of our digital ecological systems. More research must be done on the efficacy of organizational safeguards and state cybersecurity initiatives. The primary benefit of the current study is filling in these significant gaps by applying a novel quantitative method underutilized in the social sciences. Our initial investigation question focused on the association between firms' awareness of government attempts to promote the adoption of security measures and their alleged acceptance of them [12].

The connection is likely very complicated given that other factors in our investigation were also related to following the Administration's guidelines, such as giving internet safety a high priority or adding board members to manage it. Businesses that prioritise cybersecurity are expected to put safety measures in place, appoint board members to oversee cybersecurity, and keep abreast of governmental initiatives. Future research should examine the relationship between changes in organisational behaviour and government cybersecurity activities, as well as the factors that may influence these projects' outcomes. Second, by adhering to government initiatives such as Cyber Essentials or the 10 Steps to Cyber Security, organizations are likely to implement best practices that are proven effective in protecting against cyber threats. These initiatives provide a framework and guidance for implementing security controls and policies, which can help organizations identify and mitigate vulnerabilities in their systems [1].

While these initiatives may help organizations reduce the likelihood of cyber incidents, they do not guarantee complete protection. Cyber threats constantly evolve, and attackers find new ways to exploit vulnerabilities. Organisations must therefore continuously evaluate their security posture and modify their controls and policies as necessary. Programmes for employee awareness and training are also essential for ensuring that people are aware of the hazards related to cyber events and have the skills necessary to spot and report suspicious behaviour (NIST, 2018) [13].

In conclusion, adhering to government initiatives such as Cyber Essentials and the 10 Steps to Cyber Security can reduce the likelihood of cyber incidents. These initiatives provide a framework for implementing security controls and policies that protect against cyber threats. However, it is essential for organizations to continually assess their security posture and adjust their controls and policies accordingly while also investing in employee awareness and training programs [15].

The complexity and variety of cyberthreats, as well as the difficulty in measuring the effectiveness of security measures, make it challenging to accurately assess the impact of cybercrime on businesses. Effective cybersecurity measures may reduce some attacks, but they may also expose new vulnerabilities or raise public awareness and incident reporting. In order to effectively defend themselves against cyber threats, organisations must therefore regularly evaluate and enhance their cybersecurity measures, while simultaneously recognising the shortcomings of such methods and the ongoing difficulty of managing cybersecurity risk [14].

Further research is needed to account for factors such as company size, industry type, and geographical location, which could impact a business's vulnerabilities and response to cybercrime incidents.

Moreover, the current study only examines the perspective of businesses and their experiences with cybercrime. Future research must also consider the perspective of law enforcement agencies and the challenges they face in investigating and prosecuting cybercrime cases. Furthermore, there is a need to explore the effectiveness of different policy and programmatic interventions in preventing, mitigating, and responding to cybercrime against enterprises [16].

In conclusion, the current study offers important new information about the frequency, nature, and effects of cybercrime incidents against businesses. These insights can be used by decision-makers and company executives to create more potent plans for boosting organisational resistance to cyberthreats. However, further work is required to enhance data collecting and analysis, take into consideration a variety of variables that affect firms' vulnerabilities, and investigate the efficacy of various policy interventions in combating cybercrime against enterprises.

These goals include improving production efficiency, reducing operational costs, enhancing product quality, and minimizing environmental impact. The proposed strategy enables autonomous learning and reasoning to enhance the capability of industrial systems, which is critical for meeting market demands and staying competitive. Moreover, it paves the way towards highly customizable and flexible manufacturing, ultimately leading to mass customization. Overall, this research emphasizes the importance of continuous self-improvement and adaptation in the era of Industry 4.0 and highlights the potential of agent-based optimization as a critical enabler of this paradigm shift [17].

This study presents a safe and efficient method for optimizing industrial processes through machine learning. The method utilizes a data discard strategy, local approximation methods, and an exploration-exploitation trade-off strategy to optimize process parameters while avoiding breaking industrial rules regarding process quality. The algorithm was applied to a saw blade straightening machine and showed compelling results in synthetic test situations. The software tool developed can be run as a software module directly in edge devices and carries out process control, allowing for direct communication as part of the shared ecosystem. Processing power is especially needed in high-dimensional contexts with large data budgets, and fog or cloud computing can widen the range of applications. The optimization may happen continually, be initiated by the user, or even be carried out automatically by online anomaly detection. The workflow presented can serve as the foundation for additional development, and midsized machine makers can leverage the potential of ML for their specific process optimization with modified optimization setups. This study provides a practical and efficient approach to optimizing industrial processes while ensuring process quality and safety [18].

3. Materials and Methods.

3.1. Model for Gaussian Process (GP).

1. Mean Function: This function defines the expected value of the Gaussian process at any point in the input space.
2. Covariance Function: This function calculates the covariance of the Gaussian process between any two points in the input space.

The hyperparameters of the mean and covariance functions are estimated using the training samples. Once these hyperparameters are learned, the GP model can predict the output for any new input by calculating the mean and standard deviation of the corresponding Gaussian distribution.

Non-parametric GP models are powerful because they do not make any assumptions about the underlying distribution of the data. Due to their flexibility, GP models are used in many machine-learning applications, including regression, classification, and data smoothing [19].

$$GP(x) \sim \mathcal{N}(\mu(x), \Sigma) \tag{3.1}$$

Making a good Gaussian process model requires careful consideration of the kernel since it has an impact on how well the model can predict new data based on training data. There are a number of common kernels in literature, but due to its versatility and longevity, the Gaussian kernel is frequently the most used. Custom kernels, however, can also be utilised for certain applications.

$$K(j,k) = \sigma^2{}_f \exp\left(-\left(x_j - x_k\right)\left(x_j - x_k\right)^{\mathrm{T}} / 2l^2\right) \tag{3.2}$$

Given a [np] matrix of input variables, X, where n and p are the sample and input variable counts, respectively, the Gaussian kernel matrix is defined as follows. Where f and l are the hyperparameters for the kernel, the former specifies the vertical span of prediction. At the same time, the latter illustrates how rapidly the correlation between two points decreases as their distance widens. The variables x_j and x_k also represent [1 p] samples in X, where j and k are positive integers in the range [1 n], and K is a [n n] symmetric matrix. In actuality, noise is almost always present in sensor-gathered data. The Gaussian kernel is then further modified to adapt to the situation. Here the hyperparameters used for the operating system are f and 1. The first figure suggests a vertical range of estimation, while the other demonstrates the rate that the relationship between the two locations drops as their distance expands. The variables x_i and x_k additionally correspond to [1 p] samples in X, where j and k are integers that are positive in the range [1 n], and K is a [n n] symmetrical grid. In actuality, noise is almost always present in sensor-gathered data. The Gauss kernel is subsequently altered to account for that:

$$K_y = K + \sigma^2{}_y I, \tag{3.3}$$

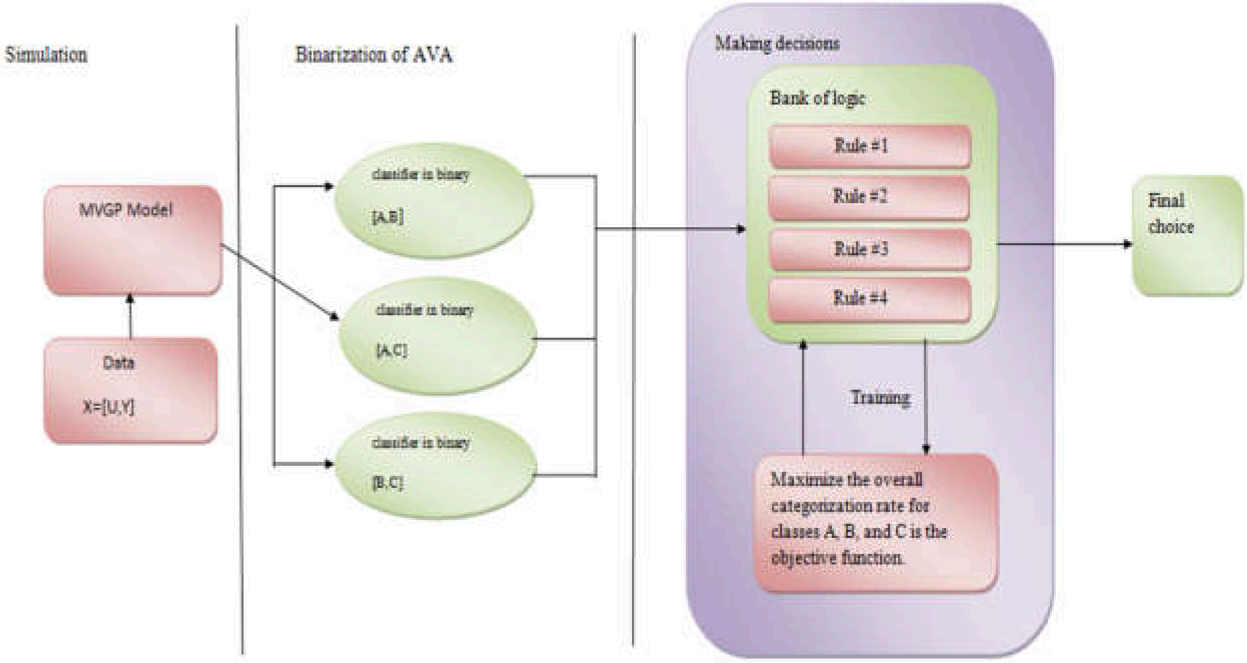


Fig. 3.1: Diagram of the OBGP classifier, which consists of three stages: Optimized Bayesian Logic-based decision making, multiclass polling, and modeling (MVGP model).

where the instructional and assessment samples' mean functions, X and Z, have been standardized to zero in everyday life. KX, X stands for the [n] kernel grid of the initial input, $K\ X, Z$ for the [n] kernel matrix produced between the testing and training input samples, and KZ, Z for the evaluation input.

$$E\left[y_z\right] = K^{\mathrm{T}}{}_{X,Z}\left(K_{X,X} + \hat{\sigma}2y\right)^{-1} y, \tag{3.4}$$

$$\Sigma\left[y_z\right] = K_{Z,Z} - \left(K^{\mathrm{T}}{}_{X,Z}\left(K_{X,X} + \hat{\sigma}^{2}{}_{y} I\right)^{-1} K_{X,Z}\right). \tag{3.5}$$

It follows the choice to optimize the combined chance at the projected mean value of yz. Because the likelihood of guessing based on the training data has already been maximized with the decision of optimum parameters, the issue then deteriorates into minimizing the odds of the expected median value of yz. The instructional data and test results are assumed to be different. Therefore, Schulz et al.'s (2018) calculations give the following projected averages and standard deviations for all samples in Z: The square root of the [n 1] diagonal of the [n n] variance matrix, where $E[yz]$ is the [n 1] prediction mean vector, thus qualifies for use to calculate the standard deviation. Figure 3.1 explains the diagram of the OBGP classifier, which consists of three stages: Optimized Bayesian Logic-based decision making, multiclass polling, and modeling (MVGP model) [30].

3.2. Optimization Using Bayes. Using the training data, the best hyper parameters for the Gaussian kernel are chosen so that the data is not over fitted and the system's behaviour under noisy conditions is adequately captured.

$$\theta = \{\sigma f, l, \sigma y\} \tag{3.6}$$

$$\theta = \arg_{\theta} \min[e(\theta)] \tag{3.7}$$

Furthermore, by evaluating the total amount of y with the predicted mean and spread of the method, the goal's error in the prediction function, e(), calculates the average of the squared error. To prevent excessive fitting of an initial set of data, the equation determines the error value using 4-fold cross-validation to.The dynamics of e() are first estimated using the GP model GPe (), the model's input.

$$a(\theta) = \max\left(0, GPe(\theta) - e_{\min}\right), \tag{3.8}$$

$$k(x_i, x_j) = \sigma^2{}_f \exp\left(1 + \sqrt{5}d/l + 5d^2 3l^2\right) \exp(-\sqrt{5}dl), \tag{3.9}$$

$$\text{s.t. } d = \|x_i - x_j\|_2 \tag{3.10}$$

4. Experimentation & Results.

4.1. Optimized Bayesian Gaussian Process.

4.1.1. Model for Multivariate GP (MVGP). A common strategy in data-driven modelling is to project elements from their initial area onto a latent (i.e., unseen) space with better algebraic features. Because it can improve prediction accuracy when there is chaos or irregular dynamics, this latent space is helpful for regression applications. This study also used the technique of principal component analysis (PCA), which is a well-known example of such a method. PCA is a linear data analysis tool used to reduce the dimensionality of a multivariate system by mapping the primary variables onto a latent space known as principle component (PC) space through a sequence of linear combinations. The frameworks for defect identification and diagnostics that utilise PCs benefit from their orthogonal and decor-related nature, which makes them more sensitive and precise [27].

$$1/nX^{\mathrm{T}}X = VDV^{\mathrm{T}} \tag{4.1}$$

Cumulative Percent Variance (CPV) metric determines which PCs to keep, represented by the letter l. As a result, l is computed as follows if 90% of the data's variability needs to be preserved:

$$l = \arg_l \min[\gamma[l] - 0.9] : \gamma[l] \geq 0.9 \& \gamma[l] = tr(D1 \to l)/tr(D) \tag{4.2}$$

Consequently, this would lessen the need for GP models to be created, which is also the primary goal of this effort. Furthermore, the outputs have no link because each GP model must forecast one of the retained PCs.

It implies that constructing the kernel matrix would be costly regarding memory and time for extensive data processes containing many variables and/or samples.multi-output GP model's tuning hyper parameters have increased. However, this does not necessarily mean the model is more accurate. This problem is especially pronounced in complicated structures like neural networks [20, 21]. Figure 4.1 defines the schematic representation of binary classifier for the classes A and B.

4.1.2. Classifiers in Binary. The suitable outcomes of other faulty categories are then identified by adding data used for training gathered from these problem classes to the MVGP model, Every chart receives values from the MVGP model. Movable Width Intermittent Aggregate, or MWIA, is the name of the aggregation technique. (MWIA uses a sliding window of sample size to aggregate conventional samples in real-time. The present recorded sample is added to a predetermined number of earlier observed samples to create the mean and standard deviation. The first one specifies the centres, whereas the second one specifies the radii. The window size is frequently restricted to a minimal amount to minimize any problems with inertial shifts that could occur with moving window techniques [22].

4.1.3. Multiple-class Classifiers. The BOGP classifier proposes four different criteria for the logic bank. Therefore, the criteria are defined as follows for each class pair "A, B" such that A B:

The binary classifier should employ the following logic-based decision schema.

Empirical thresholding of the chart statistics for class A. A sample is classified as class B if its statistics exceed the criterion. A sample is classified as class A if its statistics exceed the cutoff. It belongs to class B if not.

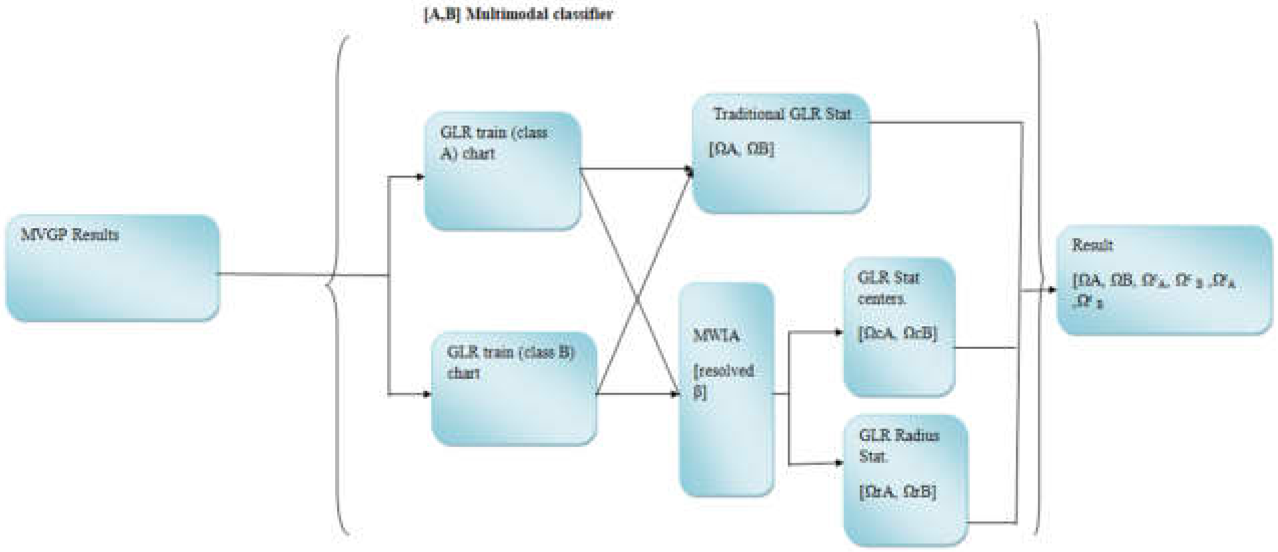


Fig. 4.1: Diagram of a binary classifier for the classes A and B.

Statistics from class A to B charts are empirically thresholded (A: B). The ratio A/B would be less than the threshold if a given sample falls under class A.

When the magnitude of the data on the two charts differs by an amount large enough to result in unexpected rounding errors during calculation, it might be advantageous. This design does not use empirical thresholding. 4. Real value with a 90% to 99.99% empirical threshold confidence level [23, 29, 24]. The threshold-free decision scheme is immune from this requirement. The best options from the training phase are applied for data validation and testing for each unique class pair. As a result, the objective function of the training phase is defined as follows:

$$f(c_1, c_2, \cdots, c_k) = \sum_{i=1}^{k} \alpha_{ci}\sigma_{ci} \tag{4.3}$$

As a result, optimization is carried out sequentially to choose the optimal alternatives for each binary classifier.

4.2. Analysis of the Findings. The findings support several essential conclusions. Furthermore, the OBGP classifier's optimization provided more trustworthy and uniform answers across all classes, even though it used more tiny training and validation datasets than the IPCA classifier. For the majority of industrial observations, data scarcity is often not a problem because advanced data-gathering techniques are easily accessible. It is a huge disadvantage when dealing with expensive or specialised operations. The availability of a method that more effectively employs a smaller dataset for training will help to reduce the cost of creating good models based on evidence and improve the modelling of complex applications when substantial sampling is not an option [25, 26].

5. Conclusion. The results show that, despite the IPCA classifier's earlier success in surpassing several informed by data and advanced learning methods acquired from prior research, the OBGP predictor was more accurate than the latter at classifying different errors in the Tennessee Eastman process. This outcome is an appropriate follow-up based on the information categorization effort.

Depending on the application, the OBGP classifier's logic-based design can also be modified later to accommodate more involved decision-making processes. Finally, the OBGP classifier outperformed those used in the literature despite having a significantly smaller training pool than earlier methods. This enormous difference in sample requirements was first caused by the GP's inability to build the core vector for large data sets.

Additionally, it provided an opportunity to demonstrate how successfully the OBGP algorithm replicated the complex irregularity of manufacturing.

REFERENCES

[1] P. Ajay, B. Nagaraj, B. M. Pillai, J. Suthakorn, and M. Bradha, *Intelligent ecofriendly transport management system based on iot in urban areas*, Environment, Development and Sustainability, (2022), pp. 1–8.

[2] E. L. Allwein, R. E. Schapire, and Y. Singer, *Reducing multiclass to binary: A unifying approach for margin classifiers*, Journal of Machine Learning Research, 1 (2000), pp. 113–141.

[3] M. Aly, *Survey on multiclass classification methods*, Neural Netw, 19 (2005), p. 2.

[4] N. Basha, M. Nounou, and H. Nounou, *Multivariate fault detection and classification using interval principal component analysis*, Journal of Computational Science, 27 (2018), pp. 1–9.

[5] E. Belasco, B. U. Philips, G. Gong, and P. Sanguansat, *The Health Care Access Index as a determinant of delayed cancer detection through principal component analysis*, Intech London, UK, 2012.

[6] A. Benaicha, G. Mourot, J. Ragot, and K. Benothman, *Fault detection and isolation with interval principal component analysis*, in Proceedings of the International Conference on Control, Engineering and Information Technology Proceedings Engineering and Technology, vol. 1, 2013, pp. 162–167.

[7] A. D. Bull, *Convergence rates of efficient global optimization algorithms.*, Journal of Machine Learning Research, 12 (2011).

[8] A. J. Burnham, J. F. MacGregor, and R. Viveros, *Latent variable multivariate regression modeling*, Chemometrics and Intelligent Laboratory Systems, 48 (1999), pp. 167–180.

[9] P. Cazes, A. Chouakria, E. Diday, and Y. Schektman, *Extension de l'analyse en composantes principales à des données de type intervalle*, Revue de Statistique appliquée, 45 (1997), pp. 5–24.

[10] L. H. Chiang, E. L. Russell, and R. D. Braatz, *Fault detection and diagnosis in industrial systems*, Springer Science & Business Media, 2000.

[11] S. X. Ding, *Data-driven design of model-based fault diagnosis systems*, IFAC Proceedings Volumes, 45 (2012), pp. 840–847.

[12] J. J. Downs and E. F. Vogel, *A plant-wide industrial process control problem*, Computers & chemical engineering, 17 (1993), pp. 245–255.

[13] R. Eslamloueyan, *Designing a hierarchical neural network based on fuzzy clustering for fault diagnosis of the tennessee–eastman process*, Applied Soft Computing, 11 (2011), pp. 1407–1415.

[14] P. I. Frazier, *Bayesian optimization*, in Recent advances in optimization and modeling of contemporary problems, Informs, 2018, pp. 255–278.

[15] J. H. Friedman, *Another approach to polychotomous classification*, Technical Report, Statistics Department, Stanford University, (1996).

[16] M. A. Gelbart, J. Snoek, and R. P. Adams, *Bayesian optimization with unknown constraints*, arXiv preprint arXiv:1403.5607, (2014).

[17] M. G. Genton, *Classes of kernels for machine learning: a statistics perspective*, Journal of Machine Learning Research, 2 (2001), pp. 299–312.

[18] J. P. George, Z. Chen, and P. Shaw, *Fault detection of drinking water treatment process using pca and hotelling's t2 chart*, International Journal of Computer and Information Engineering, 3 (2009), pp. 430–435.

[19] T. Hastie and R. Tibshirani, *Classification by pairwise coupling*, Advances in Neural Information Processing Systems, 10 (1997).

[20] S. Heo and J. H. Lee, *Statistical process monitoring of the tennessee eastman process using parallel autoassociative neural networks and a large dataset*, Processes, 7 (2019), p. 411.

[21] D. Hernández-Lobato, J. Hernández-lobato, and P. Dupont, *Robust multi-class gaussian process classification*, Advances in Neural Information Processing Systems, 24 (2011).

[22] C.-C. Hsu and C.-T. Su, *An adaptive forecast-based chart for non-gaussian processes monitoring: with application to equipment malfunctions detection in a thermal power plant*, IEEE Transactions on Control Systems Technology, 19 (2010), pp. 1245–1250.

[23] A. Hyvärinen, J. Hurri, P. O. Hoyer, A. Hyvärinen, J. Hurri, and P. O. Hoyer, *Independent component analysis*, Springer, 2009.

[24] I. T. Jolliffe, *Principal component analysis for special types of data*, Springer, 2002.

[25] I. B. Khediri and C. Weihs, *Process monitoring using an online nonlinear data reduction based control chart*, Frontiers in Statistical Quality Control 10, (2012), pp. 97–107.

[26] G. Kopsiaftis, E. Protopapadakis, A. Voulodimos, N. Doulamis, A. Mantoglou, et al., *Gaussian process regression tuned by bayesian optimization for seawater intrusion prediction*, Computational Intelligence and Neuroscience, 2019 (2019).

[27] A. Sharma, S. R. Kawale, S. P. Diwan, D. Gowda, et al., *Intelligent breast abnormality framework for detection and evaluation of breast abnormal parameters*, in Proceedings of the 2022 International Conference on Edge Computing and Applications, Tamilnadu, India, 2022, IEEE, pp. 1503–1508.

[28] A. Sharma, A. Singla, N. Sharma, D. Gowda, et al., *Iot group key management using incremental gaussian mixture model*, in Proceedings of the 3rd International Conference on Electronics and Sustainable Communication Systems, Coimbatore, India, 2022, IEEE, pp. 469–474.

[29] A. Tarek, W. Bougheloum, M. F. HARKAT, and M. Djeghaba, *Fault detection and isolation using interval principal component analysis methods*, IFAC-PapersOnLine, 48 (2015), pp. 1402–1407.

[30] S. H. Vahed, M. Mokhtare, H. A. Nozari, M. A. Shoorehdeli, and S. Simani, *Fault detection and isolation of tennessee eastman process using improved rbf network by genetic algorithm*, in Proceedings of the 8th European workshop on advanced control and diagnosis—ACD2010, no. FrA3, vol. 6, 2010, pp. 362–367.

Edited by: B. Nagaraj M.E.
Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing
Received: Oct 16, 2023
Accepted: Nov 25, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1533–1540, DOI 10.12694/scpe.v25i3.2703

BIG DATA ANALYSIS AND DEEP LEARNING OPTIMIZATION IN ARTIFICIAL INTELLIGENCE PRODUCTION OF INFORMATION ENTERPRISES

NA GAO*AND QIULING LU†

Abstract. Intelligent manufacturing technology is required to upgrade existing enterprises' management and production operations. To construct a ground breaking fusion structure, this project unites the theoretical underpinnings, technical breakthroughs, and applications of data analytics, optimisation, and intelligent production engineering. It is driven by China's desire of cutting-edge commodities and efficient growth methods. This research establishes the broad framework for merging optimisation and data analytics. There is a list of data analytics and system optimisation technologies that can address important challenges with intelligent manufacturing. By integrating data analytics and optimisation, businesses may better forecasting and management of new terrain, as well as reveal hidden information to increase decision-making efficacy.

Key words: Network optimizing, big data, smart industries, data mining strategies (DMTs), production control, intelligent manufacturing, statistic evaluation, finding of knowledge.

1. Introduction.

With the rapid development of information technology and the advent of the digital age, the demand for effective management and utilization of massive data in the business community is constantly increasing. Big data analysis and deep learning optimization have become key driving forces in modern enterprise artificial intelligence production, providing opportunities for enterprises to operate and innovate more intelligently and efficiently. In this context, big data analysis technology has become a key tool for extracting insights and knowledge from data. By delving deeper into data, enterprises can better understand key factors such as customer needs, market trends, and competitor behavior, thereby better formulating strategies and decisions. China, a major steel producer, is under pressure on two fronts. Traditional steel businesses must first modernise and reorganise in order to advance strategically. Second, new steel companies must strive for long-term growth. The most practical methods to do this are to minimise energy use, improve product quality, and boost competitiveness. The introduction of big data has had a huge influence on the industrial industry. For starters, a variety of common information and communication technologies (ICTs) have fundamentally altered how manufacturing is carried out [9, 14]. Enterprise information systems are critical in the Industry 4.0 era for realising smart manufacturing systems.

1. Needs to be more adequate information.
2. Limited business demands satisfied.
3. Lack of dynamic optimization, value-driven processes, business intelligence, and seamless integration.

Additionally, as effective, and efficient creative manufacturing systems have increased, so have their demands for knowledge, data-driven decisions, and information flow in corporate information systems. A new enterprise information systems framework is needed to close the gaps between the requirements for conventional production systems and intelligent manufacturing systems [1]. This new enterprise information system framework should have the following features and functions: (1) Data integration and interoperability: It should be able to integrate information from various data sources, including sensors, production equipment, supply chain, and market data, to support comprehensive data analysis and insights. (2) Real time and responsiveness: This framework should be able to monitor production processes in real time and respond quickly to events and issues to minimize production interruptions and efficiency losses. (3) Intelligent decision support: It should include advanced data analysis and machine learning algorithms to help enterprises make smarter and more

*Jilin Province Economic Management Cadre College, ChangChun, Jilin, 130021, China (Corresponding Author, NaGao9@126.com)

†Jilin Province Economic Management Cadre College, ChangChun, Jilin, 130021, China (QiulingLu5@163.com)

accurate decisions, thereby optimizing production processes and resource utilization. The procedure for this method consists of three steps.

It carries out three things:

- It proposes a new framework for enterprise information systems.
- It applies the TO-BE model to rethink six areas of corporate information systems.
- It uses the AS-IS model to establish requirements and collect best practices.

Finally, the proposed framework is validated using real-world examples. By incorporating six key EISs components, the issues of interoperability, uniformity, information and knowledge sharing, value creation, and data generating value were simultaneously addressed. Additionally, the functional structure looked at how business processes, information flows, and the future of data and knowledge innovation were related to the value-driven design of EISs. Using BPR and integrated lean thinking, this paper presented the process and principle for EISs. Using this technique, more conventional manufacturers can redesign their EISs to satisfy SMS standards [16]. In order to improve its production efficiency, competitiveness, and innovation ability. This empirical study provides strong guidance and reference for the future manufacturing and intelligent industries.

In the future study, software should be used to communicate the requirements analysis approach of EISs to improve intelligent manufacturing systems for conventional manufacturing processes. Applying this approach to other innovative manufacturing system operations and businesses is also essential. Developing tailored products to meet shifting consumer demands and a cooperative network to increase production efficiency are significant potential benefits of intelligent manufacturing. However, the automation of equipment in modern production processes and the digitization of industrial goods divide and disperse these technologies [18, 17].

The following are some key directions for future research: tailored products, and future research should focus on developing intelligent manufacturing systems to meet the constantly changing consumer needs. By combining big data analysis and adaptive manufacturing technology, manufacturing enterprises can better customize products, improve customer satisfaction, and achieve market competitive advantages. The establishment of cooperative networks and the development of intelligent manufacturing systems not only rely on internal technology and process optimization within the enterprise, but also require the establishment of a wide range of cooperative networks. Future research can focus on how to build supply chains, partners, and ecosystems to achieve improved production efficiency and resource sharing. Technology integration and interoperability will become key issues as different manufacturing systems and devices increase. Future research should seek standardized and universal interface solutions to ensure coordinated operation between various manufacturing systems.

2. Literature Survey. The suggested enhanced TCA tasks scheduling approach is preferable to FEF scheduling because it considers the tasks' minimal (optimal) time and precise decision (prediction) metrics. According to the experimental study, the revised PO-TCA scheduling method reduces hunger and dropout rates by 21% and 17%, respectively. In addition, our suggested strategy boosts machine utilisation by an average of 18% compared to the conventional scheduling method. This project aimed to develop a dynamic and persuasive work scheduling system based on predictive optimisation to manage resources in a well-lit manufacturing environment efficiently. The following is a summary of the main contributions of the planned study:

- An intelligent and dynamic P-TCA scheduling approach is created to improve the scheduler's decision-making skills. Additionally, to enhance the effectiveness of the TCA scheduler, a DNN-based prediction model is developed that incorporates specific decision-making skills.
- To balance workloads among symmetric production processes and enhance smart machine utilization, an integrated PO-TCA scheduling solution based on a predictive optimization mechanism is created.

With the proposed method, the task dropout rate is reduced dramatically, from 33% to 12% (a 21% improvement). As a result, the rate of tasks starting from 26% to 9% (a gain of 17%) is decreased by our suggested PO-TCA [19].

Additionally, PO-TCA has an average latency of only 16 ms, significantly shorter than FEF and TCA. The average latency for basic scheduling methods like TCA and FEF is 37.18 milliseconds and 49.59 milliseconds, respectively. Compared to the baseline plan, we suggest a scheduling method that averages an 18% improvement in machine utilisation to utilise the intelligent factory's resources efficiently. In addition, our system provided the ideal workload distribution across intelligent machines for achieving daily production objectives compared to the conventional method. We proposed that task scheduling performance is significantly improved by PO-TCA scheduling. It increases the scheduler's overall effectiveness by using data-driven and evolutionary approaches to assist and create intelligent and optimum scheduling decisions. Additionally, in the real-world setting of smart manufacturing, our proposed PO-TCA can be employed as the best scheduling technique for efficient resource management. Additionally, two effective ways to improve the suggested study are presented. Incorporating big data analytics enhances knowledge mining capabilities for the efficient operation of intelligent manufacturing, claim N. Iqbal et al. A proposed research that uses the block chain paradigm will also improve the privacy and transparency of data produced by intelligent manufacturing robots [7].

By using the potential and untapped knowledge value of precise industrial data, extensive data-driven analysis, one of the fundamental artificial intelligence technologies, improves the market competitiveness of the manufacturing industry. Additionally, it helps business executives make wise choices in a range of difficult industrial circumstances. This method provides novel solutions to difficult issues and suggests new lines of inquiry for this field of study. A comprehensive summary of crucial industrial data is given in this article. Next, it is addressed how big data-driven technologies are used in intelligent manufacturing. Finally, we discuss the problems and challenges this area is currently experiencing [20].

Using big data-driven analytics and dynamic perception, this method establishes a new paradigm for intelligent manufacturing that emphasises making the right decisions in production settings. The separation between the two research disciplines is this study's main flaw. First, the reliability issue relates to the exact sciences, such as engineering and mathematics. Big data's roots are, nevertheless, deeply ingrained in information technology. Using the conceptual framework of this new paradigm, the manufacturing system is introduced to industrial-intensive data-driven analysis. The validity and usefulness of this conceptual framework must be confirmed through additional study, even though this hypothetical big data analysis model was created in a perfect environment. The development of software systems and their application to industrial manufacturing will also be thoroughly studied in this study, along with the framework. Utilizing in-depth data analysis, this manufacturing system will also help design, implement, and manage manufacturing solutions. A popular and expanding study area is how extensive data analysis impacts manufacturing decision-making. Academics who want to study vast industrial data should find this helpful, systematic review. Petrochemical and other process-based product manufacturers may gain from it because, because of production optimisation, they can respond to market and environmental conditions more quickly. This article provides specific solutions to the challenges posed by expanding data dimensions, temporal gaps, and alignment between time series data, as well as the increased desire for quick results while considering ecological considerations. Then, a model was trained to generate intelligent production control based on real-time data using data from the industrial Internet of Things. A case study from the petrochemical sector illustrates the effectiveness of this strategy. Based on machine learning and industrial IoT, this article suggests a digital twin framework for optimising petrochemical production control. The recommended design includes practice loops, machine learning strategies, and crucial assessment indicators. The plan is a logical response to the environment's peculiar characteristics surrounding the petrochemical industry [21].

3. Materials and Methods.

3.1. Data analytics-based intelligent multi-objective optimization technique. Using data analytics to look at the interim outcomes of the evolution process, the program first dynamically estimates and builds the Pareto front of the optimization issue. The decomposition technique is put into practice on this basis [22, 23]. The method uses data analytics to map out the topography of the problem area and then uses that information to optimize the procedure. The provided form can resolve multi-objective industrial problems with outstanding results.

3.2. Multi-objective optimization-based machine learning.

3.2.1. Suggested technique. Ensemble learning is a hot topic in the field of machine learning. Examples of traditional methods are AdaBoost, Bagging, and Random Forest. These strategies employ a present framework for learning, which could cause an over-fit in actual situations. Machine learning based on multi-objective

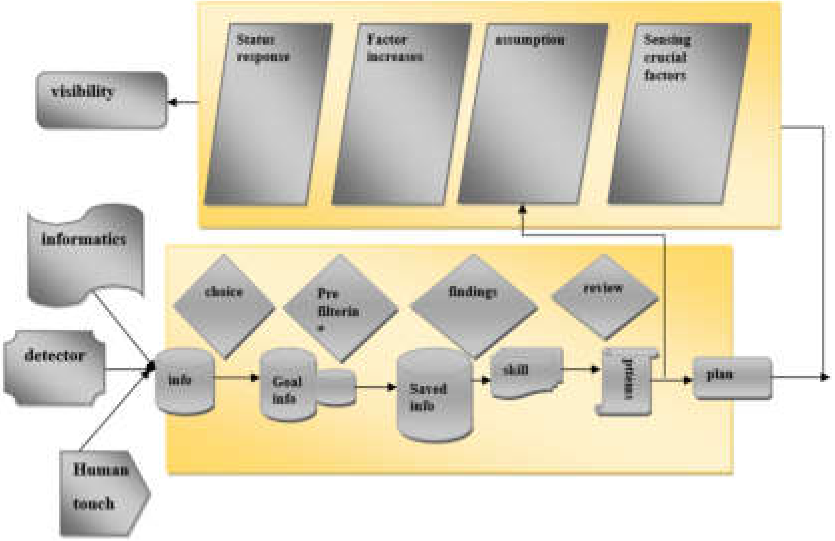


Fig. 3.1: The overall technique of data miming in production management

optimisation develops ensemble learning machines using evolutionary optimisation to balance accuracy with generalisation power [24, 25].

3.2.2. Understanding industrial processes and technology. Understanding industrial processes entails recognizing industrial images and videos, understanding industrial audio, and visualizing industrial processes. Recognizing pictures and videos is crucial for identifying and observing production processes. Skilled operators often perform it through actual image observation. Thus noise-to-text recognition and production mechanism modelling should be included in the knowledge of sound and voice technologies. To start, sound data from digital data are transformed for multidimensional monitoring. After reliable data and acquired sound signals have been assessed, the machinery and production lines' status is determined [26, 27].

The production process' dynamic essence is almost entirely restored by the industrial process visualisation. The three-dimensional simulation technique is part of the virtual reality-based production process model. The produced model is then used to visualise processes (such as the production of iron) using "black box" virtual reality technology. Additionally, the process model can be combined with essential production data by evaluating the operator, environment, and equipment state.

3.2.3. Technology for process observation and description. Monitoring and characterizing intricate industrial production processes are essential for ensuring safe manufacturing, energy conservation, and reduced emissions. Monitoring and description (such as the amount of energy and materials used at each production stage) are used to measure the manufacturing process. For instance, measuring issues in energy consumption can be categorized into three groups based on the multiple measurement objects: the product, the manufacturing process, and the medium. Each industrial process's specific media consumption and recovery rates are calculated statistically from the process dimension, resource consumption, and energy recovery [28].

3.2.4. Technology for inventory planning and the entire production process. Science and technological advancements have made collecting and storing precise local data easy. Data analytics may effectively extract vital information from vast quantities of inaccurate, noisy experimental data [6].

3.2.5. Technology for batching and scheduling in production and logistics. Customers, however, only need a small number of high-quality products. Production management has faced several difficulties as a result of the conflict between requirements for a wide variety of products and mass manufacturing. When examining the production characteristics of the steel industry, production/logistics batching and schedules refer

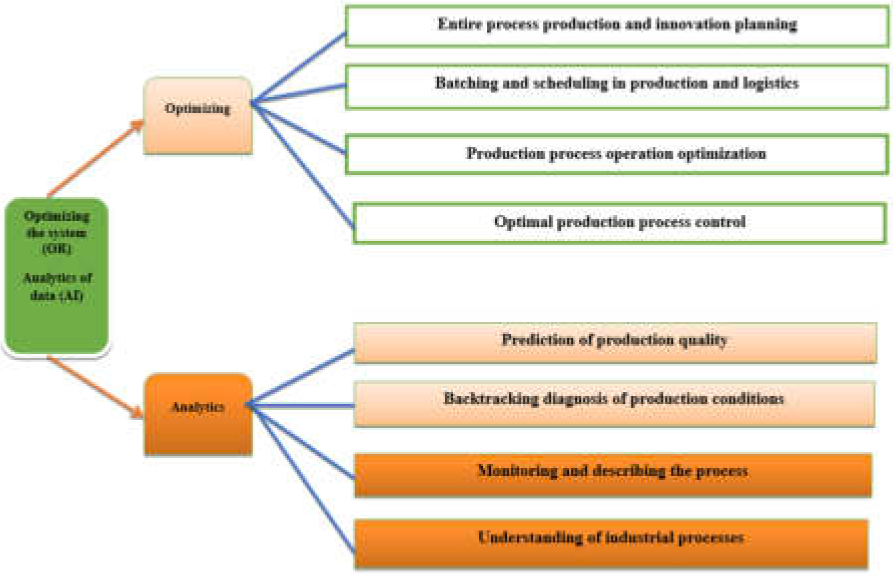

Optimizing the system (OR)
Analytics of data (AI)
Optimizing
Entire process production and innovation planning
Batching and scheduling in production and logistics
Production process operation optimization
Optimal production process control
Analytics
Prediction of production quality
Backtracking diagnosis of production conditions
Monitoring and describing the process
Understanding of industrial processes

Fig. 4.1: In order to solve urgent issues, the smart industry uses data analytics and optimization

to the assignment of works with identical or equivalent parts to batches of sufficient size. Most batch-scheduling issues in production and logistics were resolved by utilizing deterministic parameters. Stochastic optimization is the most popular approach for solving problems with uncertain parameters [12]. Figure 3.1 describes the overall technique of data miming in production management.

4. Experimentation and Results.

4.1. Engineering Data analytics and optimization technology implementation in smart industries. Product quality is anticipated at the discovery phase following a comprehensive production process analysis. Operation optimisation and ideal control follow from this. Eventually, the scheduling and production planning decision-making procedures are improved to match the intelligence business' capabilities. System optimisation focuses on action and judgement, whereas data analytics depends on perception and discovery [3]. By implementing engineering data analysis and optimization technologies, the intelligent industry can better utilize data resources, improve decision-making accuracy and efficiency, drive technological innovation, and achieve more sustainable business operations. This will help the intelligent industry maintain competitiveness and achieve sustainable growth in the constantly changing market.

4.2. Awareness level. An intelligent industry's perceived level is its bedrock. At this stage, the critical analytics concerns are understanding industrial data and monitoring and describing processes. Understanding involves distinguishing between industrial data (such as pictures, sounds, and text) and the virtual reality representation of black-box technology.

4.3. Knowledge level. Management, machinery, control systems, and manufacturing methods all significantly impact the level of innovation in the intelligent sector. Three key analytics issues are addressed: process diagnostics, product quality forecasting, and technological knowledge mining. A thorough analysis of the production process may also show the amount of technical proficiency supporting the levels of execution and decision-making. A scientific basis for corporate production planning and management strategies is provided by prediction, which tries to demonstrate the quality of products based on the present production conditions and previous data [8, 15, 29, 11]. As shown in Figure 4.1.

4.4. Execution quality. At the execution level, system optimisation techniques like manufacturing process optimal control and operation optimisation are needed. Operational optimisation controls the production

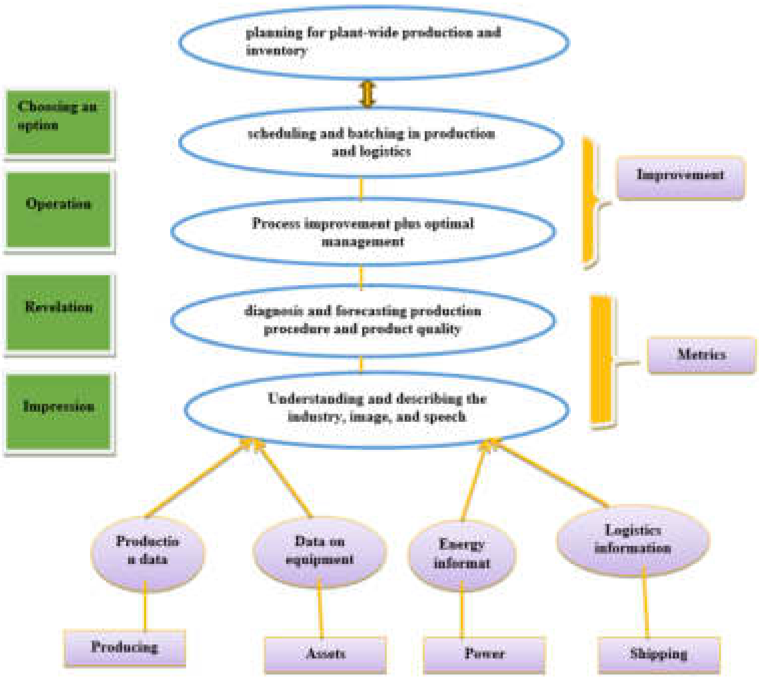

planning for plant-wide production and inventory
Choosing an option
Operation
Revelation
Impression
scheduling and batching in production and logistics
Process improvement plus optimal management
diagnosis and forecasting production procedure and product quality
Understanding and describing the industry, image, and speech
Improvement
Metrics
Productio n data
Data on equipment
Energy informat
Logistics information
Producing
Assets
Power
Shipping

Fig. 4.2: Portrays a multiple structure for information analytics and enhancement software in smart industry sectors

process by using a mechanical or data analytics model to describe the quantitative link between the operating parameters and relevant economic indicators. In other words, system activities are monitored. At the same time, appropriate process parameters (such as temperature, pressure, and flow) are established without changing the process flow or adding more production equipment. The goals are increasing product quality, making money, and streamlining the production process [4, 10, 13, 2, 5, 30].

4.5. Level of decision-making. Engineering management decision-making is the most important in the ecosystem of the intelligent industry. Production/logistics batching and scheduling and whole-process production and inventory planning are two key optimization concerns identified as having the potential to alter the production process and improve resource, energy, and equipment consumption. Optimizing the output of each production unit and the quantity between two successive cycles and the inventory, from raw materials to semi-finished goods to finished goods, is a part of the problem of whole-process production and inventory planning.

5. Conclusion. Finally, this understanding demonstrates a four-level framework for the intelligence industry. Due to the restrictions of the research topic, this study might only touch on a small portion of the intelligence industry. Industrial intellectualization is a field that is constantly evolving. More incredible information about how products are made can be collected and stored thanks to current manufacturing control solutions. Data analysis methods can therefore be applied automatically. The results of the previous study suggest that there may be restrictions on the processing and mining of detailed data, intelligent mining process enhancement, assessment of the quality of excavation, expression and preservation of information, and other conditions.

Acknowledgement. 2021 Project of the 14th Five Year Plan for Education Science in Jilin Province: Innovative exploration and practical research on the education management mode of higher vocational colleges based on artificial intelligence under the background of Educational Informatization (GH21493).

REFERENCES

[1] P. Ajay, B. Nagaraj, B. M. Pillai, J. Suthakorn, and M. Bradha, *Intelligent ecofriendly transport management system based on iot in urban areas*, Environment, Development and Sustainability, (2022), pp. 1–8.

[2] A. Azadeh, J. Seif, M. Sheikhalishahi, and M. Yazdani, *An integrated support vector regression–imperialist competitive algorithm for reliability estimation of a shearing machine*, International Journal of Computer Integrated Manufacturing, 29 (2016), pp. 16–24.

[3] D. Ding, Z. Pan, D. Cuiuri, H. Li, S. Van Duin, and N. Larkin, *Bead modelling and implementation of adaptive mat path in wire and arc additive manufacturing*, Robotics and Computer-Integrated Manufacturing, 39 (2016), pp. 32–42.

[4] S. Du, C. Liu, and L. Xi, *A selective multiclass support vector machine ensemble classifier for engineering surface classification using high definition metrology*, Journal of Manufacturing Science and Engineering, 137 (2015), p. 011003.

[5] S. Du, L. Xi, J. Yu, and J. Sun, *Online intelligent monitoring and diagnosis of aircraft horizontal stabilizer assemble processes*, The International Journal of Advanced Manufacturing Technology, 50 (2010), pp. 377–389.

[6] S. Du, X. Yao, and D. Huang, *Engineering model-based bayesian monitoring of ramp-up phase of multistage manufacturing process*, International Journal of Production Research, 53 (2015), pp. 4594–4613.

[7] R. S. Kumar, B. Nagaraj, P. Manimegalai, and P. Ajay, *Dual feature extraction based convolutional neural network classifier for magnetic resonance imaging tumor detection using u-net and three-dimensional convolutional neural network*, Computers and Electrical Engineering, 101 (2022), p. 108010.

[8] S. Leo Kumar, J. Jerald, and S. Kumanan, *Feature-based modelling and process parameters selection in a capp system for prismatic micro parts*, International Journal of Computer Integrated Manufacturing, 28 (2015), pp. 1046–1062.

[9] L. Lopez, M. W. Carter, and M. Gendreau, *The hot strip mill production scheduling problem: A tabu search approach*, European Journal of Operational Research, 106 (1998), pp. 317–335.

[10] M. W. Milo, M. Roan, and B. Harris, *A new statistical approach to automated quality control in manufacturing processes*, Journal of Manufacturing Systems, 36 (2015), pp. 159–167.

[11] P. Podržaj and A. Čebular, *The application of lvq neural network for weld strength evaluation of rf-welded plastic materials*, IEEE/ASME Transactions on Mechatronics, 21 (2015), pp. 1063–1071.

[12] M. Ranjit, H. Gazula, S. M. Hsiang, Y. Yu, M. Borhani, S. Spahr, L. Taye, C. Stephens, and B. Elliott, *Fault detection using human–machine co-construct intelligence in semiconductor manufacturing processes*, IEEE Transactions on Semiconductor Manufacturing, 28 (2015), pp. 297–305.

[13] P. K. Rao, J. Liu, D. Roberson, Z. Kong, and C. Williams, *Online real-time quality monitoring in additive manufacturing processes using heterogeneous sensors*, Journal of Manufacturing Science and Engineering, 137 (2015), p. 061007.

[14] S. Sahay and P. Kapur, *Model based scheduling of a continuous annealing furnace*, Ironmaking & Steelmaking, 34 (2007), pp. 262–268.

[15] N. Sahebjamnia, I. Mahdavi, and N. Cho, *Designing a new model of distributed quality control for sub-assemble products based on the intelligent web information system*, Journal of Intelligent Manufacturing, 21 (2010), pp. 511–523.

[16] A. Shao, *Can industrial intelligence promote industrial transformation?—-case of mining enterprises*, Frontiers of Engineering Management, 4 (2017), pp. 375–378.

[17] A. Sharma, S. R. Kawale, S. P. Diwan, D. Gowda, et al., *Intelligent breast abnormality framework for detection and evaluation of breast abnormal parameters*, in Proceedings of the International Conference on Edge Computing and Applications, Tamilnadu, India, 2022, IEEE, pp. 1503–1508.

[18] L. Tang and P. Che, *Generation scheduling under a co 2 emission reduction policy in the deregulated market*, IEEE Transactions on Engineering Management, 60 (2013), pp. 386–397.

[19] L. Tang, F. Li, and Z.-L. Chen, *Integrated scheduling of production and two-stage delivery of make-to-order products: Offline and online algorithms*, INFORMS Journal on Computing, 31 (2019), pp. 493–514.

[20] L. Tang, J. Liu, A. Rong, and Z. Yang, *A review of planning and scheduling systems and methods for integrated steel production*, European Journal of Operational Research, 133 (2001), pp. 1–20.

[21] ———, *Modelling and a genetic algorithm solution for the slab stack shuffling problem when implementing steel rolling schedules*, International Journal of Production Research, 40 (2002), pp. 1583–1595.

[22] L. Tang, J. Liu, F. Yang, F. Li, and K. Li, *Modeling and solution for the ship stowage planning problem of coils in the steel industry*, Naval Research Logistics, 62 (2015), pp. 564–581.

[23] L. Tang, P. B. Luh, J. Liu, and L. Fang, *Steel-making process scheduling using lagrangian relaxation*, International Journal of Production Research, 40 (2002), pp. 55–70.

[24] L. Tang, Y. Meng, Z.-L. Chen, and J. Liu, *Coil batching to improve productivity and energy utilization in steel production*, Manufacturing & Service Operations Management, 18 (2016), pp. 262–279.

[25] L. Tang, D. Sun, and J. Liu, *Integrated storage space allocation and ship scheduling problem in bulk cargo terminals*, IIE Transactions, 48 (2016), pp. 428–439.

[26] L. Tang, G. Wang, and Z.-L. Chen, *Integrated charge batching and casting width selection at baosteel*, Operations Research, 62 (2014), pp. 772–787.

[27] J. Tao, K. Wang, B. Li, L. Liu, and Q. Cai, *Hierarchical models for the spatial–temporal carbon nanotube height variations*, International Journal of Production Research, 54 (2016), pp. 6613–6632.

[28] K. Wang and F. Tsung, *Recursive parameter estimation for categorical process control*, International Journal of Production Research, 48 (2010), pp. 1381–1394.

[29] S. M. Weiss, A. Dhurandhar, R. J. Baseman, B. F. White, R. Logan, J. K. Winslow, and D. Poindexter, *Continuous prediction of manufacturing performance throughout the production lifecycle*, Journal of Intelligent Manufacturing, 27 (2016), pp. 751–763.

[30] W.-A. Yang, *Monitoring and diagnosing of mean shifts in multivariate manufacturing processes using two-level selective ensemble of learning vector quantization neural networks*, Journal of Intelligent Manufacturing, 26 (2015), pp. 769–783.

Edited by: B. Nagaraj M.E.
Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing
Received: Oct 17, 2023
Accepted: Nov 25, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1541–1548, DOI 10.12694/scpe.v25i3.2704

AN INTELLIGENT NETWORK METHOD FOR ANALYZING CORPORATE CONSUMER REPURCHASE BEHAVIOR USING DEEP LEARNING NEURAL NETWORKS

QIUPING LU[*]

Abstract. Earth system models (ESMs) are our key tools for analyzing the planet's existing state and predicting its evolution in the next continuing human-caused events. However, the use of artificial intelligence (AI) approaches to augment or even replace conventional ESM functions has expanded in recent years, raising hopes that AI will be able to overcome some of the major difficulties in climate research. We address the advantages and disadvantages of neural ESM neurons, as well as the unsolved question of whether AI will eventually replace ESMs. Dynamic geophysical events are the foundation of Earth and environmental studies. Given the widespread acceptance of AI and the growing amount of Earth data, the geoscientific community may wish to seriously explore using artificial intelligence (AI) approaches at a much deeper level. Although it is a tall ambition to integrate hybrid physics and AI approaches from a fresh perspective, geology has yet to figure out how to make such methods feasible. This research is an important step towards realising the concept of combining physics and artificial intelligence to address problems with the Earth's system.

Key words: Earth system modelling, long short term memory, artificial intelligence, environmental sciences, geology

1. Introduction. In geosciences, applying AI approaches has a lengthy history. For instance, Abbott (1991), who coined hydro informatics 30 years ago, characterized it as combining computational hydraulics and artificial intelligence. Nonetheless, the mainstream geoscientific community is still cautious about embracing AI approaches, in large part because an AI model is believed to be a "black box," offering few mechanistic explanations beyond its capacity to fit, while some scientists have made an effort to explain black-box models, doing so instead of first developing interpretable models is likely to result in bad practices be perpetuated. With AI models, the geoscience community has increasingly considered the efficiency of the two paradigms as an appealing study area [30, 6].

AI is used to create a proxy model, identify and repair the discrepancy between physical models and observations, and other potential ways of physics-AI efficiency in geoscience were outlined. Comparatively speaking, less research has been done on the hybrid modelling method, which tries to add several physical layers to a network of neurons (NN) to make it more materially realistic. The geoscience industry has grown more interested in studying the effectiveness of the two approaches due to the relative benefits of physical procedures and AI models [12]. Given the relative benefits of biological processes and AI models, the geoscience community has grown more interested in studying the effectiveness of the two paradigms. Due to the employment of a single, integrated AI architecture throughout the process, the hybrid modelling method more closely matches the possibility of raising geoscientific awareness of AI systems [2].

Since the nineteenth century, geoscientists have extensively used ODEs to explore geosystem undercurrents, such as signal processing and global climate modelling. The proposed work is developed a innovative style is utilizing runoff simulation. The primary function of hydrology is catchment runoff modelling. Hydrology is an entire field in geosciences. In a watershed, water intake, outflow, and storage all change completed period. In this work, the LSTM layer in a DL architecture incorporates a conceptual hydrologic model, resulting in hydrology-aware DL models. Overall, our work shows that when adequately trained, AI may similarly acquire biological knowledge to humans [25].

2. Fundamentals of Earth System Modelling. Based on Navier-Stokes equivalences, which explain the atmosphere's - fluid dynamics and seas, are examples of simple physical equations of motion explicitly

[*] Henan Polytechnic Institute, Nanyang, Henan, 473000, China (Corresponding Author, QiupingLu9@126.com)

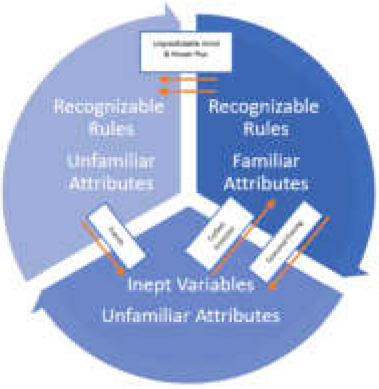


Fig. 2.1: Representation of components of earth system model

known for Earth system components (Figure 2.1). It is practically impractical to resolve all pertinent dynamics scales quantitatively. Hence approximations must be made.

The complication of the ESM makes it difficult to easily infer macroscopic occurrences from tiny scales that may or may not be understood, is primarily to blame for this. For these situations, parameterizations of potentially critical processes must also be approximated. Such parameterizations create free parameters in ESMs, regardless of the process, for which fair values must be determined empirically [26]. Modern ESMs are so large that most systematic calibration techniques, such as those based on Bayesian inference, are impractical. As a result, the models are frequently adjusted by hand.

Even if they are required, parameterizations can generate biases or structural model errors. Furthermore, it is envisaged that the model's representation of the Earth system will become more accurate if significant advancements are resolved plainly. Despite the huge success of ESMs, problems and uncertainty persist.

1. A large range of equilibrium climate sensitivity still exists in current ESMs. Between CMIP5 and CMIP6, the range of expected symmetric weather warmth increased from 1.9-4.5 °C to 1.5-6.6 °C. Losing such reservations is one of the key issues in developing ESMs.
2. Numerous Earth system subsystems may swiftly and gradually induce alterations, according to theoretical considerations and paleoclimate evidence. Many clear evolutions have been found in the CMIP5 models' predictions of the future after a comprehensive investigation. But due of the extremely risky events, it is still unclear if ESMs are reliable in predicting them.
3. Using the present ESMs is still necessary to assess the efficiency or environmental impact of CO2 removal methods and crucial mitigation options for putting the Paris Agreement16 into practice. ESMs also need to do a better job of capturing basic environmental processes like the carbon cycle, the availability of water and nutrients, or the connections between land use and climate [24, 10].
4. The distributions of the time series encoding the dynamics of the Earth system frequently include heavy tails. Severe weather has a very detrimental socioeconomic effect. Because human climate change is still occurring, such events are expected to get worse. There is still space for improvement when representing extremes, even though modern ESM are too competent at predicting usual climatic quantities.

3. Literature Review. Following this line of thinking, we introduce the term "Neural Earth System Modelling" (NESYM) and emphasize the need for a detail explanation forum that brings organized professionals in AI, extensive data analysis, and Earth and climate science. The possibilities and potential problems of NESYM and talk about the uncleared queries of AI is neither only permeate but ultimately replace ESM.

Process-based models were once considered vital resources for comprehending the intricate relationships between the coupled Earth system's components and predicting how the Earth system will react to human-induced weather modification. The startling idea that Earth system models (ESMs) would become obsolete when new artificial intelligence (AI) capabilities are developed has caused a gold rush-like feeling and ridicule among the scientific communities On the other hand, the majority of neural networks lack actual process knowledge and are trained for discrete applications [18]. Yet, the daily expanding Earth system observation (ESO) data streams, growing processing power, and the accessibility and availability of potent. We emphasize the need for fresh transdisciplinary cooperation between the concerned communities to address the arising problems.

It is not simply a fun exercise; it is crucial for applying AI to creating and using NESYM. Earth and climate scientists can contribute to creating uniform standards that compare the geophysical consistency of stand-alone ML and NESYM hybrid models. However, the AI community's assistance is required to tackle additional recently noted ML issues. For instance, it is creating new ways to recognize and prevent shortcut learning in NESYM hybrids.In conclusion, the evolution of neural earth system modelling will only occur through joint cooperation. The development of techniques will be further stimulated by problems unique to the Earth system, and we offer the following four leading suggestions [20].

As a result, we suggest testing the efficacy of machine learning methods using produced fictional data. It is used to assess actual data utilizing a range of dynamics that complex physical models simulate. When training data is provided and extrapolation issues are taken into account, it is crucial. Future models should employ process-driven and machine-learning methods of learning, according to our recommendation. Although data-driven machine-learning technologies will greatly improve and supplement physical modelling, it will still play a vital role in geoscientific research. Additionally, the neuro sciences will contribute to the development of reliable physically grounded linkages for machine learning research [4].

Since physics constrains the search parameter space and eliminates implausible models, hybridization has an intriguing regularization effect. Hence, physics-aware machine learning models need less training data, are simpler (sparser), and better combat overfitting to attain similar performance. Overall, the hybrid modelling framework represents a new line of inquiry that should be intensified and continued [5].

Despite its widespread success in other fields, The Transformer as a new DL architecture has yet to receive much acceptance in this one. In this study, we suggest Earth former, a space-time Transformer for predicting the behaviour of the Earth system. The concept is to apply parallel cuboid-level self-attention while decomposing the data into cuboids. A group of global vectors connect these cuboids in more detail. To test the efficacy of cuboid attention and determine the ideal architecture of Earth former, we do tests on the Movingness dataset [8].

This paper proposes an Earth former, a space-time transformer, to forecast how the Earth system would behave. Cuboid Awareness is a flexible and useful construction material that forms the basis of the Earth. We obtained SOTA on Movingness, our recently proposed N-body MNIST, SEVIR, and ICAR-ENSO. There are certain limitations to the job we do. Initially, the Earth model is a mechanical version without an uncertainty model. By forecasting the average of all potential futures, the model can deliver foggy forecasts with poor perceptual quality and require additional beneficial small-scale characteristics. More suitable methods must be taken to evaluate the uncertainty in Earth system forecasting models. Extending Earth's historical forecasting model to a probabilistic one represents an exciting future direction. We plan to investigate ways to include biological data into Earth's past atmosphere in the future. [22, 9].

4. Materials and Methods.

4.1. LSTM Architecture. A unique variety of recurrent neural networks (RNN), known as the LSTM architecture, was created to address the typical RNN's inability to learn long-term dependencies. The typical RNN can only remember sequence 10, as Bengio et al. (1994) demonstrated. It would indicate that for daily streamflow modelling, we could only utilize the past ten days (about one and a half weeks)' worth of input taken from climatological data to forecast.

We unfold the network's recurrence into a directed acyclic graph to illustrate how the RNN and the LSTM function. The input $m = m_1, m_2, \cdots, m_n$ consists of the preceding n repeated period stages of self-determining variable star and is processed sequentially to forecast the output at a particular period. The internal processes of the recurrent cell, and these processes distinguish the LSTM from a standard RNN.Old RNN cells have single internal state, l_t, which is recalculated at each time step using the equation below.

$$l_t = s(Vm_t + Yi_{t-1} + bias) \tag{4.1}$$

Also, the input gate of the second gate computes which (and to what extent) information effect is utilized. The current time step's cell state should be updated:

$$i_t = \sigma(Vi_{xt} + Yi_{ht-1} + bias) \tag{4.2}$$

The following equation updates the cell state c_t.

$$c_t = f_t(c_{t-1} + i_t c_t) \tag{4.3}$$

where denotes multiplication by elements. Eq. (4.1) applies because both entries in the vectors f_t are in the range (0, 1). Like that, it determines which newly stored infect information will be discarded. (The value of it of approx. 0).

Output gate calculation:

$$o_t = \sigma(VWx_t + Yoh_{t=1} + b_0) \tag{4.4}$$

V_0,Y_0 and b_0 are a set of learnable parameters defined for the problem, and ot is a vector with values between (0, 1) output control. It is determined from this vector (4.5)

$$h_t = \tanh(c_t)o_t \tag{4.5}$$

It can maintain the integrity of the information stored across many time steps because of its straightforward linear interactions with the remaining LSTM cell. This property assists in preventing the issue of exploding or disappearing gradients during training. The final discharge prediction is computed by a single output neuron conventional dense layer. The following equation provides the viscous layer calculation:

$$y = V_d h_n + bias \tag{4.6}$$

V_d is the weight matrix, bias is the bias term, h_n is the output of the final layer in LSTM at the previous time step, and y is the final discharge, all derived from Eq. (4.6).

Finally, Algorithm 1 displays the complete LSTM layer's pseudocode. When there are numerous stacked LSTM layers, the output $h = [h_1, h_2, \cdots, h_n]$ of the first layer serves as the input for the subsequent layer. Eq. (4.6) is then used to determine the discharge, the final output, where ht is the final output of the last LSTM layer [11, 13, 3, 15, 19].

4.2. LSTM Layers Description. The standardization process had comprised a predetermined number of recapitulations in which the full calibration period is reproduced using a particular set of model parameters. The network's adaptable (or learnable) parameters, including its weights and biases, are altered when an LSTM is trained based on the particular loss function of each iteration step. As a result, the gradient loss function including the network metrics may was evaluated .

Figure 4.1 depicts the LSTM training and standardization process for one iteration phase graphically. A batch or mini batch of the available training data is typically used for one iteration of LSTM training. A hyperparameter is anything preset, such the 512 samples per batch. One discharge value from a certain day plus weather information from the n days before that day make up each sample. The loss function is computed as the average of the MSE of the simulated and real runoff for each of 512 samples in each iteration step [14]. Each piece inside a batch can be made up of randomly selected time steps, which are unnecessary to be ordered chronologically because the discharge of a certain time step is just a function of the meteorological inputs of the prior n days. Convergence can be hastened even with random samples included in the batch [23].

Given an optimization procedure without a convergence condition, the number of iteration steps affects the overall number of model runs performed during calibration for conventional hydrological models. Neural networks are referred to as epochs. Epochs are the intervals at which a training sample updates a model parameter. If the data set included 1000 training samples and the batch size was 10, an epoch would have 100 iteration steps (the quantity of training samples divided by the quantity of samples per batch). In each iteration step, 10 of the 1000 samples are taken without a replacement, continuing until all 1000 samples have been used. The discharge time series of the training data is accurately replicated once [16].

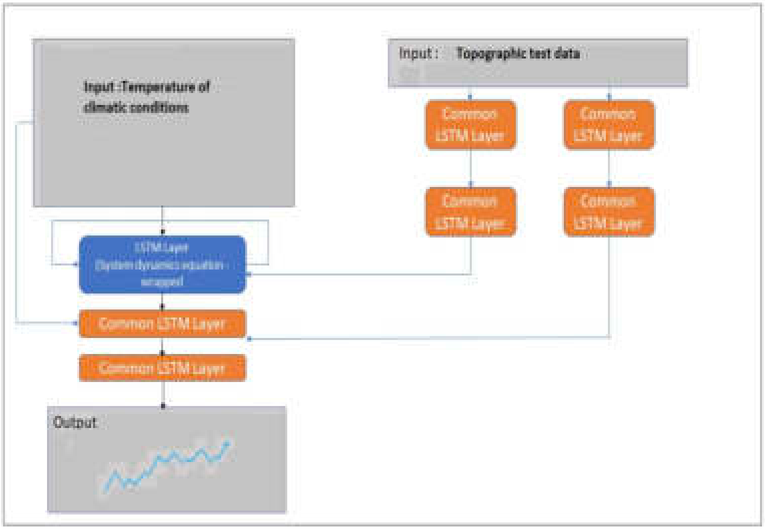

Input :Temperature of climatic conditions
Input : Topographic test data
Common LSTM Layer
Common LSTM Layer
Common LSTM Layer
Common LSTM Layer
LSTM Layer (System dynamics equation wrapped)
Common LSTM Layer
Common LSTM Layer
Output

Fig. 4.1: LSTM architecture based on earth system modelling

For a conventional hydrological model, it is comparable to one calibration iteration, with the crucial difference being that each sample is generated independently of the others. The LSTM's learning process throughout several training epochs. Despite having to learn the complete rainfall-runoff relation from scratch for each period, the network may better capture the discharge dynamics (grey line of random weights).

5. Experimentation and Results.

5.1. Dataset. The GSDE is created using a variety of regional and national soil databases or soil maps, as well as the 1:5 million scale Digital Soil Map of the World (DSMW), which serves as a fundamental soil map. In the accompanying information, specifics regarding the data sources are provided. One or more components make up the soil mapping units in the soil maps. Each element takes up a specific portion of the mapping unit, although it is not evident where they are. In most cases, the components share the same soil type or a mix of soil type and additional taxonomy data, such as land use and texture class. The FAO-74 legend is used to construct the DSMW. Europe and northern Eurasia are covered by the 1:1 million ESDB, which uses FAO-90 soil categorization data. Using the soil polygon linkage approach and the Genetic Soil Classification of China (GSCC), the soil database for the land surface modelling in China was created [Shangguan et al., 2013]. To fill in the gaps in the SOTER attribute data at scales between 1:250,000 and 1:5 million, the soil attributes of the SOTWIS are based on the FAO-90 categorization [1, 17, 27, 28].

5.2. Results and Discussion. The State Soil Geography (STATSGO) dataset was replaced by the GSM of the U.S. at a scale of 1:250,000 using the Soil Taxonomy (S.T.) [Soil Survey Staff, 1999]. However, the available properties are significantly diverse and only partially cover the soil maps. These two profile databases were integrated into a single data structure. Ten thousand two hundred fifty-three profiles containing FAO-74 and FAO-90 legends were stored in WISE 5.1, released in 2005. Around 1900 of the 81,218 profiles in the NCSS were gathered outside of the United States. The NCSS uses the ST to refer to dirt. After deleting soil profiles lacking soil classification or soil property measurement, 71 339 profiles remain. Using an LSTM approach, Figure 5.1 shows the dataset's mean, median, and mode [21, 29, 7]. Geospatial data is present in 60,638 of the 89,592 profiles in the WISE and NCSS. Local soils are typically more accurately represented in soil profiles with greater density. Multiple techniques were employed in the lab and throughout time to measure the soil properties in WISE and NCSS. The accuracy of the information of the NCSS is greater because the soil investigations in the NCSS adhered to established protocols. In contrast, soil analyses in the WISE were carried out in at least 190 laboratories worldwide using a variety of approaches. [Batjes, 2008a]. For deep soils,

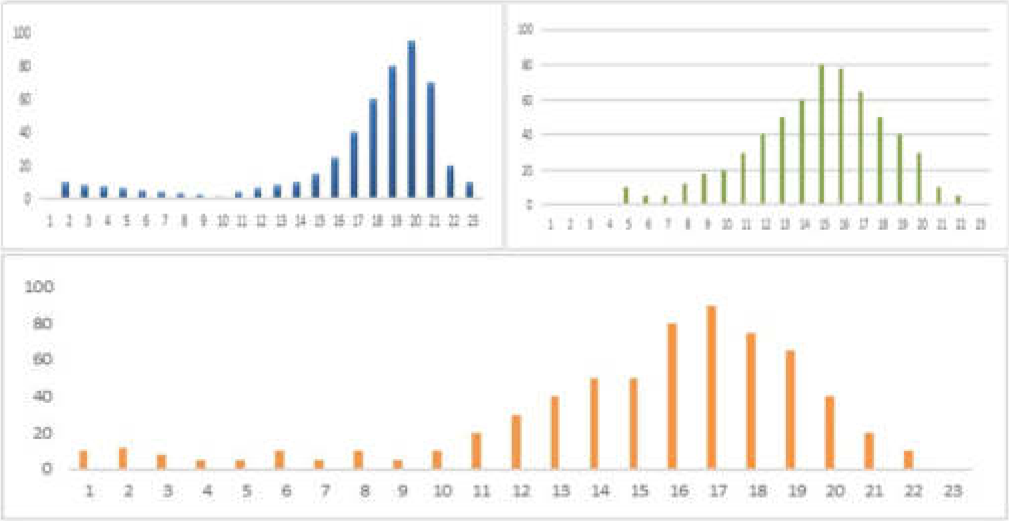


Fig. 5.1: Determination of mean, median and mode of LSTM in ESM

in particular, the characteristics of a soil profile are only sometimes known for each horizon. Regarding soil properties, different soil classes are represented differently.

The two pipelines, in this instance, are used for runoff modelling and its parameterization, respectively, in the generic design. As required by the LSTM, the climatic forcing variables P, T, and Lday, shortwave downward radiation SRad and vapour pressure VP are the leading pipeline's inputs. A two-layer standard NN block is supplied with the five input variables and a preliminary runoff estimation Q* from the model-wrapped LSTM(Feng et al., 2019). Conv1D layer has been used in research for data-based hydrologic modelling because it can handle the lagged impact through a one-direction convolution operation. Through the Conv1D layers, the physical approach's approximation errors are fixed, and the final runoff Q is achieved. Like the main pipeline architecture, the "hybrid DL model" blends physical principles (represented by the LSTM) with data-driven components (i.e., Conv1D layers). Although there are many potential traps and dead ends in this research field, a significant amount of risk is involved. The promise that artificial intelligence (AI) will assist in resolving the main problems in Earth and climatic sciences is now required. Some of these challenges were highlighted at the beginning of this Viewpoint. In addition, it is unlikely that AI will be able to solve the issue of climate prediction on its own at this time. Therefore, the science of the Earth system will be able to advance through AI, transcending the current uproar. The chance of the next evolutionary step will, however, improve if we can create interpretable and geophysical consistent AI technology and find solutions to the limitations mentioned above. The goal of Reichstein et al. (2019) to use hybrid physics and AI methodologies to address Earth system challenges has been advanced by this study.

Moreover, the parameterization pipeline provides the main pipeline's catchment awareness, which has dual blocks of completely associated layers that supply for the LSTM layer and N for the Conv1D layers. The parameterization pipeline allows to change with physiographic features across many catchments. Figure 5.2 shows yearly based data of ESM analysis.

6. Conclusion. Our Perspective is a reaction to the recent request for cooperation from the AI community as well as the description of a workable scientific approach to better comprehend the present and future conditions of the Earth. The artificial neural network framework suggested in this study can correctly infer information about occurrences that are not experienced, as demonstrated via runoff modelling. The revolutionary design provides a practical method for appropriately guiding AI using geoscientific data. We foresee future studies that will extend the developed framework to accommodate the deployment of increasingly sophisticated AI systems to advance geoscience research and apply it in a variety of geoscientific situations.

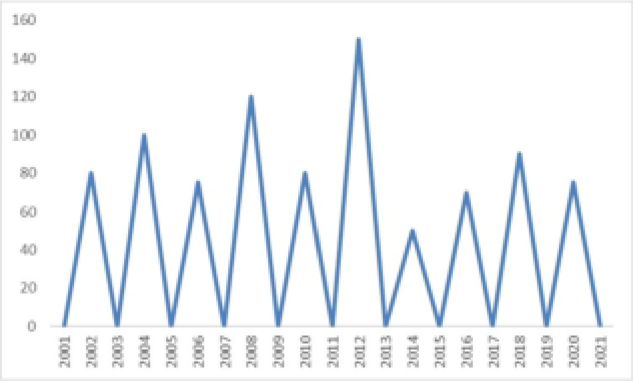


Fig. 5.2: Yearly based ESM analysis

REFERENCES

[1] *Chapter 12: Zonal Energy Budget*, pp. 141–157.

[2] A. Agapiou, *Remote sensing heritage in a petabyte-scale: satellite data and heritage earth engine© applications*, International Journal of Digital Earth, 10 (2017), pp. 85–102.

[3] P. Ajay, B. Nagaraj, R. A. Kumar, R. Huang, and P. Ananthi, *Unsupervised hyperspectral microscopic image segmentation using deep embedded clustering algorithm*, Scanning, 2022 (2022).

[4] M. Alvarez, D. Luengo, and N. D. Lawrence, *Latent force models*, in Artificial Intelligence and Statistics, PMLR, 2009, pp. 9–16.

[5] A. C. Antoulas, *Approximation of large-scale dynamical systems*, SIAM, 2005.

[6] P. Bauer, A. Thorpe, and G. Brunet, *The quiet revolution of numerical weather prediction*, Nature, 525 (2015), pp. 47–55.

[7] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., *Language models are few-shot learners*, Advances in Neural Information Processing Systems, 33 (2020), pp. 1877–1901.

[8] S. L. Brunton, J. L. Proctor, and J. N. Kutz, *Discovering governing equations from data by sparse identification of nonlinear dynamical systems*, Proceedings of the National Academy of Sciences, 113 (2016), pp. 3932–3937.

[9] G. Camps-Valls, L. Martino, D. H. Svendsen, M. Campos-Taberner, J. Muñoz-Marí, V. Laparra, D. Luengo, and F. J. García-Haro, *Physics-aware gaussian processes in remote sensing*, Applied Soft Computing, 68 (2018), pp. 69–82.

[10] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, *Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization*, Geoscientific Model Development, 9 (2016), pp. 1937–1958.

[11] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, *Transformer in transformer*, Advances in Neural Information Processing Systems, 34 (2021), pp. 15908–15919.

[12] S. Hantson, A. Arneth, S. P. Harrison, D. I. Kelley, I. C. Prentice, S. S. Rabin, S. Archibald, F. Mouillot, S. R. Arnold, P. Artaxo, et al., *The status and challenge of global fire modelling*, Biogeosciences, 13 (2016), pp. 3359–3375.

[13] W. H. Heiss, D. L. McGrew, and D. Sirmans, *Nexrad: next generation weather radar (wsr-88d)*, Microwave Journal, 33 (1990), pp. 79–89.

[14] J. Herman, P. Reed, and T. Wagener, *Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior*, Water Resources Research, 49 (2013), pp. 1400–1414.

[15] J. Ho, N. Kalchbrenner, D. Weissenborn, and T. Salimans, *Axial attention in multidimensional transformers*, arXiv preprint arXiv:1912.12180, (2019).

[16] C. Hulbert, B. Rouet-Leduc, P. A. Johnson, C. X. Ren, J. Rivière, D. C. Bolton, and C. Marone, *Similarity of fast and slow earthquakes illuminated by machine learning*, Nature Geoscience, 12 (2019), pp. 69–74.

[17] A. Karpatne, G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar, *Theory-guided data science: A new paradigm for scientific discovery from data*, IEEE Transactions on Knowledge and Data Engineering, 29 (2017), pp. 2318–2331.

[18] R. Knutti, M. A. Rugenstein, and G. C. Hegerl, *Beyond equilibrium climate sensitivity*, Nature Geoscience, 10 (2017), pp. 727–736.

[19] A. Krishnapriyan, A. Gholami, S. Zhe, R. Kirby, and M. W. Mahoney, *Characterizing possible failure modes in physics-informed neural networks*, Advances in Neural Information Processing Systems, 34 (2021), pp. 26548–26560.

[20] G. A. Meehl, C. A. Senior, V. Eyring, G. Flato, J.-F. Lamarque, R. J. Stouffer, K. E. Taylor, and M. Schlund, *Context for interpreting equilibrium climate sensitivity and transient climate response from the cmip6 earth system models*, Science Advances, 6 (2020), p. eaba1981.

[21] S. Penny, E. Bach, K. Bhargava, C.-C. Chang, C. Da, L. Sun, and T. Yoshida, *Strongly coupled data assimilation in*

multiscale media: Experiments using a quasi-geostrophic coupled model, Journal of Advances in Modeling Earth Systems, 11 (2019), pp. 1803–1829.

[22] A. Rajendran, N. Balakrishnan, and P. Ajay, *Deep embedded median clustering for routing misbehaviour and attacks detection in ad-hoc networks*, Ad Hoc Networks, 126 (2022), p. 102757.

[23] A. Sharma, D. Gowda, A. Sharma, S. Kumaraswamy, M. Arun, et al., *Priority queueing model-based iot middleware for load balancing*, in Proceedings of the 6th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2022, IEEE, pp. 425–430.

[24] T. F. Stocker, D. Qin, G.-K. Plattner, M. M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, *Climate change 2013: The physical science basis. contribution of working group i to the fifth assessment report of ipcc the intergovernmental panel on climate change*, (2014).

[25] M. Stockhause and M. Lautenschlager, *Cmip6 data citation of evolving data*, Data Science Journal, 16 (2017), pp. 30–30.

[26] K. E. Taylor, R. J. Stouffer, and G. A. Meehl, *An overview of cmip5 and the experiment design*, Bulletin of the American Meteorological Society, 93 (2012), pp. 485–498.

[27] P. J. Van Leeuwen, H. R. Künsch, L. Nerger, R. Potthast, and S. Reich, *Particle filters for high-dimensional geoscience applications: A review*, Quarterly Journal of the Royal Meteorological Society, 145 (2019), pp. 2335–2365.

[28] S. Vetra-Carvalho, P. J. Van Leeuwen, L. Nerger, A. Barth, M. U. Altaf, P. Brasseur, P. Kirchgessner, and J.-M. Beckers, *State-of-the-art stochastic data assimilation methods for high-dimensional non-gaussian problems*, Tellus A: Dynamic Meteorology and Oceanography, 70 (2018), pp. 1–43.

[29] A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis, et al., *Deep learning for computer vision: A brief review*, Computational Intelligence and Neuroscience, 2018 (2018).

[30] N. Wiener, J. von Neumann, M. . Mead, G. Bateson, W. S. McCulloch, W. Pitts, K. I. Lewin, F. S. C. Northrop, M. R. Harrower, L. S. Kubie, R. A. Bjork, and W. E. Bijker, *Predicting the future, leo howe and alan wain. cambridge university press, new york, ny. 208 pages. isbn: 0-521-41323-0. $29.95*, Bulletin of Science, Technology & Society, 14 (1994), pp. 240 – 240.

Edited by: B. Nagaraj M.E.
Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing
Received: Oct 17, 2023
Accepted: Nov 25, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1549–1556, DOI 10.12694/scpe.v25i3.2709

APPLICATION OF CLOSED-LOOP THEORY IN DEEP LEARNING TRAINING GUIDED BY HIGH-STRENGTH INTELLIGENT MACHINERY

ERFU GUO*

Abstract. Artificial intelligence (AI) algorithms and continuous monitoring technologies have the potential to transform the way chronic illnesses are managed. We will also talk about the problems and potential that AI technology presents for CGM in individualised and preventive medicine. Furthermore, we assessed the AHCL system's usefulness in patients with impaired awareness of hypoglycemia (IAH) and those who correctly recognised hypoglycemia symptoms. The participants' ages varied from 37 to 15, and they had received diabetes medication for an average of 20 to 10 years. IAH was seen in 12 individuals (27%) with a Clarke's score of less than 3. Patients with IAH were older than those who did not have IAH. The baseline CGM readings and A1c were the same, but the estimated glomerular filtration rate (eGFR) was lower. Despite prior insulin treatment, the AHCL system resulted in an overall drop in A1c (from 6.9 0.5% to 6.7 0.6%, P 0.001). Only three patients (7%) received Clarke's three scores after six months on the AHCL system, resulting in a 20% absolute risk decrease for IAH (95% confidence interval: 7-32).

Key words: Artificial intelligence, machine learning, glucose monitoring in a closed loop.

1. Introduction.

In recent years, deep learning technology has made significant breakthroughs in various fields, from natural language processing to computer vision, as well as autonomous driving. However, to make machines more intelligent, more high-strength intelligent machinery is needed to guide the deep learning process. In this context, the application of closed-loop theory has become particularly important. Closed loop control is a method of controlling a system by continuously monitoring and adjusting the system's output to achieve specific goals or maintain the system's operation in the desired state. This theory has been widely applied in automation, engineering, and control systems, but its application in the field of deep learning is still in the exploratory stage. Diabetes mellitus is a worldwide and chronic disease caused by a difficulty with glucose metabolism. By 2030, it is expected that there will be 439 million adult diabetes globally, costing roughly $490 billion USD. Diabetes and its complications are mostly caused by abnormalities in glucose metabolism. However, such monitoring cannot immediately detect the functions of hyperglycemia. The growth of continuous glucose monitoring (CGM) is a developing area of interest, which will rely heavily on technological advancements that have taken decades to perfect. Wearable CGM biosensors have recently experienced tremendous growth in popularity, with sales exceeding $1 billion. When controlling diabetes, CGM has a few advantages to finger stick blood glucose monitoring [8, 9]. First, CGM has several advantages over the traditional capillary blood glucose measuring method, including the removal of psychological and physical pain. Despite the rapid advancement of CGM technology, several barriers, like cost, lag time, the need for calibration, and others, still prevent its widespread usage. The main topics have been well-reviewed. The evolution of CGM and wearable CGM biosensors is examined in this article. Unfortunately, patients with IAH are underrepresented in clinical studies, and more research is required on how to use automated insulin delivery systems to restore hypoglycaemia awareness. To better understand this, we examined a prospective cohort of 46 T1D patients who had transitioned to an autonomous insulin administration device and then underwent CGM or flash glucose monitoring (FGM) [7, 4].

Several types of type 2 diabetes mellitus are available. They can be used in various ways, including evaluating long-term clinical outcomes, assessing the costs of clinical trials, and assisting in choosing the most appropriate interventions for these populations. The Dexcom system was regarded as a helpful study aid. The effectiveness of a new, powerful algorithm-equipped Dexcom was evaluated. Suitable modelling techniques are used to anticipate future glucose concentrations. The high-precision and real-time data transmission of the

*Shijiazhuang University of Applied Technology, Shijiazhuang, Hebei, 050081, China (Corresponding Author: ErfuGuo3@126.com)

Dexcom system helps to ensure the accuracy of research data. This is crucial for the reliability of the research results. The Dexcom system not only provides current blood sugar levels, but also monitors blood sugar trends. This is very useful for studying the triggering and duration of hypoglycemic or hyperglycemic events. The real-time data transmission of the Dexcom system allows patients to more easily monitor their blood sugar without the need for frequent fingertip blood sugar tests. This can improve patient engagement and research execution. With time, the continuous glucose monitor's accuracy and consistency increased, with the most significant improvement. The results show that signal processing-induced time delays have been reduced, and low plasma glucose performance has been improved [12, 1, 6, 22, 10]. Performance enhancements for sensor systems are envisaged as a result of these upgrades. Thanks to this application, which keeps blood glucose levels steady, reaction times are slowed when glucose is consumed. The results of this study can also be used to improve closed-loop systems and provide data for insulin pumps. Future work may have a foundation thanks to these findings. People with long-term diabetes become more mindful of how low blood sugar affects their bodies. This programme can assist in preventing hypo and restless periods since it can identify hypo and hyper periods even before the drop or rise in blood glucose level has fully begun. It can, however, also result in an unexpected change in consciousness. This application successfully maintains a constant blood glucose level because glucose eating reduces blood glucose level reaction times. After testing this application with real-time data from a continuous glucose monitor, the algorithm will be enhanced. A single programme that covers every available sensor would also be the ideal use case. The outcomes of this study can also be used to teach insulin pumps and improve closed-loop systems [18, 11, 20, 13].

2. Literature Survey. An autoimmune condition known as type 1 diabetes mellitus necessitates ongoing patient care. We demonstrated the viability of a model-based Reinforcement Learning strategy for a fully automated artificial pancreas that is safe for humans. The architecture used can control blood sugar levels without the requirement for meal notification because it doesn't need to know how much CHO was consumed. The average results demonstrate that the created controller can automatically and effectively regulate blood glucose levels for simulations lasting up to 12 hours and incorporating two meals. It is feasible to look into this exploratory work further [2].

This study differs from others in that it makes use of artificial intelligence. Technology in a supply chain and reverse logistics' garbage recycling section. In this paper, a design for CLSC pomegranates is proposed. The corresponding logistics network, developed for years, includes producers, distribution centres, customers, and compost end consumers. In the current study, a MOO model of a sustainable CLSC is presented. Reduced network costs and energy consumption are the chain's reverse logistic operations' goals, including rubbish recycling. An effort has been made to evaluate and validate the identified issue through a case study on pomegranates in Iran. The fruit is transformed as planned along the supply chain into food, pharmaceuticals, and concentrate after delivery and distribution. Through reverse logistics, the pomegranate waste is also transformed into recycled goods like compost, an organic fertiliser, and ethanol, which may be used as a sustainable energy source and alternative fuel for vehicles—several methods used to produce pomegranates. Automation cuts personnel costs and shipping expenses, which significantly lowers chain pricing. Using image processing to diagnose pomegranate quality ensures compliance with international standards. Manufacturers have also experienced severe problems with waste and damaged products. The organization's economic, social, and environmental aims can all be accomplished through this framework. The findings of this study may be helpful to businesses and managers who work with food, crops, and other items that have a chance of failing [25].

Two meta-heuristic methods are employed. The three answers are compared, and the problem is also addressed using the GAMS program.

- This will modify the research's findings. The mathematical model is then built using the data that was collected.

Also, here are a few suggestions for future research:

- Using the fuzzy set technique to estimate the degree of ambiguity in pomegranate demand
- Solving the given model using a sound optimization technique and treating it as a scenario-based model.
- Consider a cooperative game between pomegranate growers and the government from the game theory perspective [5].

Based on these forecasts, patients choose the optimal strategy to control their blood sugar levels, considering things like insulin dosage and other relevant factors. Machine learning (ML) techniques can be used to model glucose level trends to forecast this variable accurately. It is challenging to directly run complex machine learning algorithms on restricted devices due to their poor processing capabilities. Machine learning (ML) technology can be used to simulate trends in blood sugar levels to accurately predict this variable. The ML algorithm can train models based on past blood glucose data, lifestyle factors, dietary habits, medication treatment, and other factors to predict future blood glucose levels. This model can help patients with diabetes better manage their blood sugar, predict potential hyperglycemia or hypoglycemia events, and take corresponding measures to avoid dangerous situations. Performance, edge computing, and the usage of lightweight machine learning techniques. Despite these limitations, feature extraction and pre-processing allow machine learning techniques to be applied to constrained devices. It is crucial to compare the computational needs of machine learning methods for forecasting as this might significantly impact how well-suited they are to restricted devices. A capable device could easily handle the computational requirements of the random forest technique for short datasets. Many times, with some limitations, random forest-based forecasting tasks can be managed by restricted devices. Given the characteristics of the volunteers, the present study, which collected measurements from 40 patients with diabetes, may have some limitations. Although the glycaemic control levels in our sample ranged widely, it is possible that individuals who performed inaccurately could hurt the standard of diabetes management. It might lead to more individualised and accurate glucose level projections, enabling more effective management of diabetes [19].

3. Materials and Methods.

3.1. Research plan. In our prospective study, 46 T1D patients who frequently attended the diabetic outpatient clinic at an academic hospital in Madrid, Spain, and were monitored by CGM or FGM took part. (ClinicalTrials.gov identifier: NCT04900636). Extending invitations to all T1D patients who met the research eligibility requirements, we used consecutive sampling to reduce sample bias. Approved national laws carried out the study, the 1964 Helsinki Declaration, and its following amendments [17].

3.2. Study participants. Consecutively, we sought out adults at least 18 years old, had had T1D for more than a year, and were closely watched by CGM or FGM. Prior episodes of ketoacidosis and diabetic autoimmune disease were required for the diagnosis of T1D, and the usage of insulin was essential for survival after them American Diabetes Association standards. The following were listed as exclusion criteria:

1. Identifying forms of diabetes mellitus outside type 1 diabetes (T1D).
2. Lack of ability to receive the instruction or learn the information necessary to operate a computerized insulin delivery device [15, 14].

4. Experimentation and Results.

4.1. Data gathering and evaluation.

4.1.1. Hypoglycaemia with diminished awareness as measured by the Clarke score. A hypoglycaemia incident is a hypoglycaemic episode requiring outside help to administer therapy. A hypoglycemic event, also known as a hypoglycemic episode, refers to a decrease in blood sugar levels to a dangerous level that requires external intervention or treatment to correct the condition. Under normal circumstances, the blood sugar level in the human body fluctuates within a certain range, but when the blood sugar level is too low, it may lead to a series of physical and neurological reactions that may endanger the patient's health. The Clarke questionnaire has already been verified using hypoglycaemic clamping, both prospective and retrospective records of extreme hypoglycaemia in the T1D patient group, according to L. Nattero-Chavez et al. Diabetes Research and Clinical Practice 199 (2023) 110627. The participants' knowledge of hypoglycaemia symptoms was assessed using a reliable and validated 8-item survey, and the results were used to determine Clarke's score. The frequency of hypoglycaemic episodes that participants had encountered in the two months prior and their symptoms' during hypoglycaemia was disclosed. Each response received either an "A" for awareness or an "R" for cognitive impairment. Each R response was worth 1, compared to each A [3, 16, 21, 23, 24]. Figure 3.1 describes the schematic of artificial intelligence used to manage diabetes. Continuous glucose monitoring, or CGM.

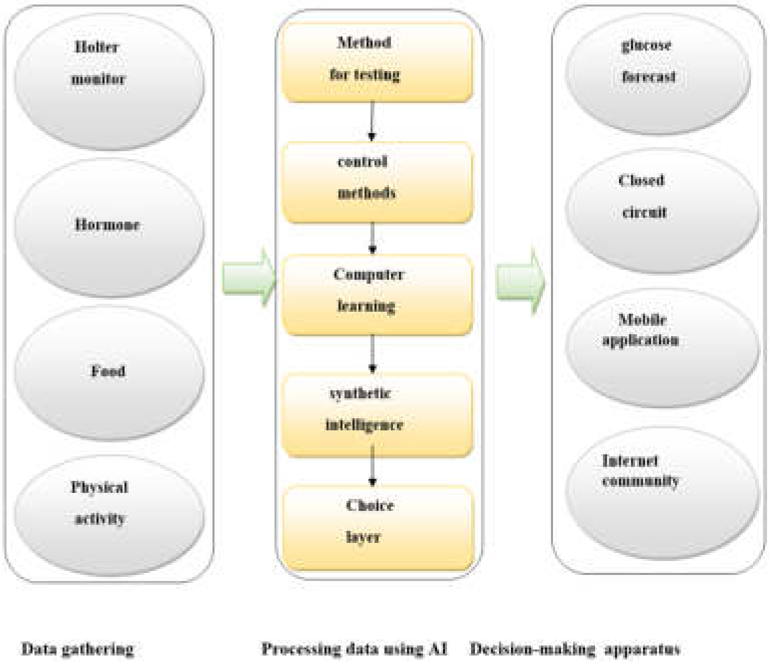


Fig. 3.1: Schematic of artificial intelligence used to manage diabetes. Continuous glucose monitoring, or CGM.

4.1.2. Artificial intelligence for CGM biosensors. Instead of conventional screening methods, artificial intelligence is being utilised to diagnose diabetic macular oedema and moderate diabetic retinopathy. Several diabetes control application scenarios have shown potential for combining CGM and machine learning. The artificial intelligence application of continuous glucose monitoring (CGM) biosensor is an important technology that can help diabetes patients better manage their blood glucose levels. The resources for patient self-management, an automated retinal screening, closed-loop control, and calibration are all included. A biosensor with artificial intelligence that continuously monitors the amount of glucose is shown in Figure 4.1. Patients put on a continuous glucose monitor (CGM) sensor that wirelessly feeds data to a smartphone while continuously checking glucose.

4.2. Algorithms for closed-loop controls. The closed-loop control algorithm, also known as feedback control algorithm, is an automatic control method in a control system. Its basic principle is to adjust the input based on the output feedback information of the system to achieve the desired goal or maintain stable operation of the system in the expected state. These algorithms are widely used in automation, engineering, medical equipment, and other fields, where real-time adjustments are required to the system to respond to changes or maintain performance. People with T1D must use insulin therapy and frequently adjust their dosage to meet their glycemic goals [10]. Hence, closed-loop control systems. The MPC algorithm predicts glucose levels and modifies insulin delivery using a dynamic model based on fictitious output data [6, 10]. Even though a closed-loop technology that automatically regulates blood glucose has been commercially accessible, some patients may still find using an artificial pancreas unreliable and potentially stressful. Figure 4.2 defines an advanced hybrid closed-loop system for hypoglycaemia awareness was implemented, and baseline and six months later results were compared and Figure 4.3 deals with an variations in the concentration of glycated haemoglobin (A1c) were seen six months after switching to an advanced hybrid closed-loop system (means SD).

Data represent mean SD.

Observable variations between T0 and T2 -x

There are notable distinctions between T0 and T6-y

Differences that are statistically significant between T2 and T6-z

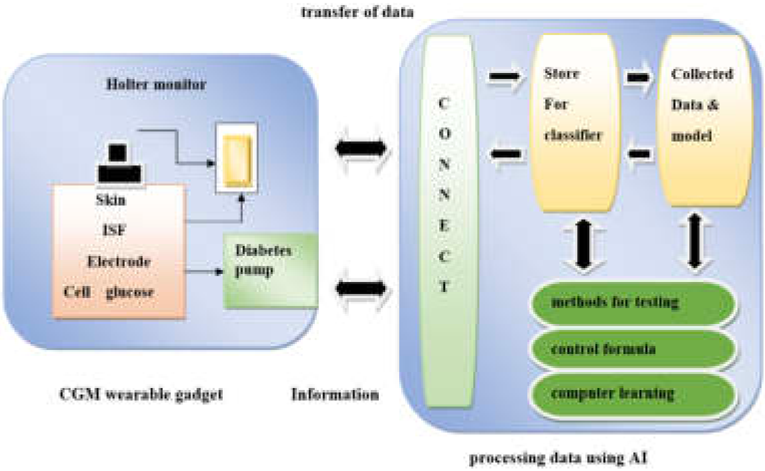


Fig. 4.1: A biosensor with artificial intelligence that continuously checks the level of glucose. Patients apply a continuous glucose monitor (CGM) sensor to their skin, which wirelessly transmits data to a smart phone while constantly monitoring glucose l.

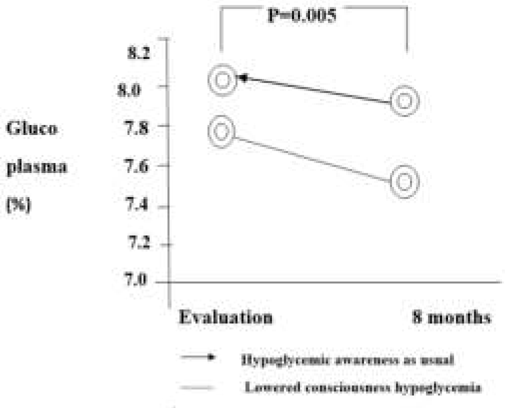


Fig. 4.2: An advanced hybrid closed-loop system for hypoglycaemia awareness was implemented, and baseline and six months later results were compared.

Substantial variations between the IAH group and the regular IH group=w

5. Conclusion. In its conclusion, this research analysed the advancement of CGM and emphasized the interaction between CGM performance and AI. CGM biosensors aim to revolutionize patient care for treating diseases like diabetes. The primary barriers to the widespread use of CGM biosensors are the cost of supplies (35.3%), accuracy (30.1%), and discomfort with having devices on one's body (29.7%). However, the evolution of CGM technology is moving in the direction of adaptability, downsizing, and long-term closed-loop systems. The goal of developing AI-powered CGM biosensors is being met. First, it should be emphasised that the physiological lag time and the CGM sensor's effectiveness impact the well-known lag between blood glucose and ISF glucose. An additional factor for T1D is the price. Closed-loop decision-making based on CGM sensors, as well as data adaptation and learning, are essential. But even before the next technology revolution, AI is being created for biosensors, opening the door to medical advancements. We've already covered the enormous amounts of data created by continuous monitoring and the three ways AI enhances CGM performance. They

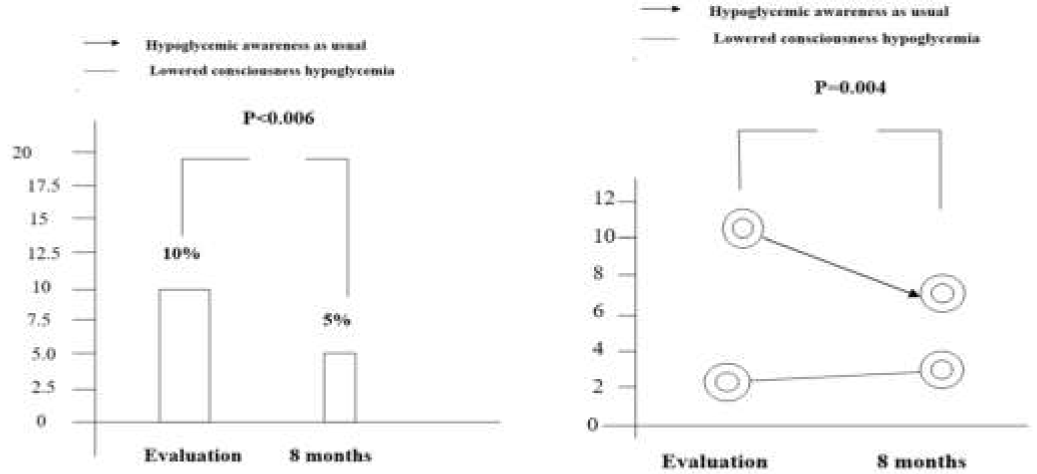


Fig. 4.3: Variations in the concentration of glycated haemoglobin (A1c) were seen six months after switching to an advanced hybrid closed-loop system (means SD).

Table 4.1: Based on the presence or absence of impaired awareness of hypoglycemia (IAH), maintains glucose monitoring metrics, biochemical parameters, and insulin dosages at baseline (T0), two months later (T2), and six months later (T6). After switching to an advanced hybrid closed-loop system, this is done.

Factor	Baseline(T0)		After 3 months (T2)		After 8 months (T6)		P
	Normal IAH n=18	AH n=44	Normal IAH n=18	AH n=44	Normal IAH n=18	AH n=44	
variables for continuous glucose tracking							
TBR <80 mg/dL	0.4±0.9	0.9±0.12	0.7±0.89	0.12±0.12	0.2±0.89	0.2±0.1	NS
TBR <100 mg/dL^{x,y}	10±15	9±9	80±19	9±7	60±1	4±4	0.02
Total TBR <100 mg/dL^{x,y}	10±15	9±9	18±65	5±9	8±5	5±9	0.013
TIR 100 and 210 mg/dL^{x,y}	90±10	80±20	50±40	40±90	90±0.10	4±20	<0.004
TAR >210 mg/dL^{x,y}	30±8	35±20	50±7	33±70	5±9	33±70	<0.004
TAR >300 mg/dL	6±9	4±9	8±10	3±10	6±40	1±40	NS
Total TAR >210 mg/dL^{x,y}	40±20	34±23	80±30	3±7	8±0.3	3±67	>0.003
Other results							
Average sensor use(%)	120±10	100±4	0.9±0.56	02±0.12	0.9±0.56 90±49 38±85 70±20 7±6 4±1 90±40	0.3±0.2	0.003
GMI(%)	8.1±0.6	29±4	90±49	5±7		4±70	0.001
ordinary sensor glucose^{x,y}	200±23	45±5	38±85	2±7		4±7	0.004
sensor glucose CV(%)^{x,y}	28±20	67±9	70±20	50±97		67±7	NS
sensor glucose SD(mg/dL)^{x,y}	53±13	78±12	7±6	35±80		5±0.4	0.007
closed-loop system duration(%)	-	4±6	4±1	3±90		6±0.60	0.008
everyday carbohydrate intake(gr/d)^{W}	-	-	90±40	7±7		4±2	NS
Biochemical factors							
fasting blood sugar(mg/dl)	230±45	56±6	-	-	40±5	56±6	0.05
eGFR(mL/min/2.35 m^2)^{y}	120±5	67±6	-	0.5	10±70	67±6	0.065
UACR(mg/g)	15±3	7±3	-	1	1±2	7±3	NS
A1c(%)$^{y^{W}}$	9±0.9	35±8	-	0.2	67±012	35±8	NS
BMI & daily insulin dosage							
Daily total insulin dose(U/kg)^{z}	0.66±0.32	0.6±0.78	0.6±0.32	0.1±0.38	0.6±0.2	0.6±0.8	0.056
Daily basal insulin(U/kg)*	0.50±0.90	0.90±0.10	0.90±0.10	0.10±0.10	0.5±10	0.9±0.10	NS
Daily bolus insulin(U/kg)^{y}	0.6±0.67	0.40±0.87	0.10±0.7	0.01±0.6	0.4±0.7	0.5±0.67	0.090
Bolus of automatic correction(U/kg)^{z}	-	-	-	-	-	-	0.056
Normal BMI(Kg/m^2)	39.6±6.8	0.9±0.5	17.6±6.8	0.20±0.5	39.6±6.8	0.9±0.2	0.001

will probably soon open the door for individualised care. Ultimately, closed-loop therapeutic technology is the best amalgamation of CGM and AI, providing many clinical possibilities and scientific advancements in artificially intelligent biosensors and medicine. It's critical to underline the clinical relevance of these findings. Although we have made some encouraging progress, the application of closed-loop theory in deep learning is still an evolving field. Future research can explore more complex closed-loop control methods and more advanced machine intelligence systems to further improve performance.

REFERENCES

[1] P. Ajay, B. Nagaraj, and R. Huang, *Deep learning techniques for peer-to-peer physical systems based on communication networks*, Journal of Control Science and Engineering, 2022 (2022).

[2] P. Ajay, B. Nagaraj, R. Huang, P. Raj, and P. Ananthi, *Environmental and geographical (eg) image classification using flim and cnn algorithms*, Contrast Media & Molecular Imaging, 2022 (2022).

[3] T. Bobrowski and W. Schuhmann, *Long-term implantable glucose biosensors*, Current Opinion in Electrochemistry, 10 (2018), pp. 112–119.

[4] S. A. Brown, B. P. Kovatchev, D. Raghinaru, J. W. Lum, B. A. Buckingham, Y. C. Kudva, L. M. Laffel, C. J. Levy, J. E. Pinsker, R. P. Wadwa, et al., *Six-month randomized, multicenter trial of closed-loop control in type 1 diabetes*, New England Journal of Medicine, 381 (2019), pp. 1707–1717.

[5] G. Cappon, M. Vettoretti, G. Sparacino, and A. Facchinetti, *Continuous glucose monitoring sensors for diabetes management: a review of technologies and applications*, Diabetes & Metabolism Journal, 43 (2019), pp. 383–397.

[6] O. J. Collyns, R. A. Meier, Z. L. Betts, D. S. Chan, C. Frampton, C. M. Frewen, N. M. Hewapathirana, S. D. Jones, A. Roy, B. Grosman, et al., *Improved glycemic outcomes with medtronic minimed advanced hybrid closed-loop delivery: results from a randomized crossover trial comparing automated insulin delivery with predictive low glucose suspend in people with type 1 diabetes*, Diabetes Care, 44 (2021), pp. 969–975.

[7] P. E. Cryer, *Mechanisms of hypoglycemia-associated autonomic failure in diabetes*, New England Journal of Medicine, 369 (2013), pp. 362–372.

[8] J. Geddes, J. E. Schopman, N. N. Zammitt, and B. M. Frier, *Prevalence of impaired awareness of hypoglycaemia in adults with type 1 diabetes*, Diabetic Medicine, 25 (2008), pp. 501–504.

[9] A. E. Gold, K. M. Macleod, and B. M. Frier, *Frequency of severe hypoglycemia in patients with type i diabetes with impaired awareness of hypoglycemia*, Diabetes Care, 17 (1994), pp. 697–703.

[10] A. Graveling and B. Frier, *Impaired awareness of hypoglycaemia: a review*, Diabetes & Metabolism, 36 (2010), pp. S64–S74.

[11] R. Huang, X. Yang, and P. Ajay, *Consensus mechanism for software-defined blockchain in internet of things*, Internet of Things and Cyber-Physical Systems, 3 (2023), pp. 52–60.

[12] E. Isganaitis, D. Raghinaru, L. Ambler-Osborn, J. E. Pinsker, B. A. Buckingham, R. P. Wadwa, L. Ekhlaspour, Y. C. Kudva, C. J. Levy, G. P. Forlenza, et al., *Closed-loop insulin therapy improves glycemic control in adolescents and young adults: outcomes from the international diabetes closed-loop trial*, Diabetes Technology & Therapeutics, 23 (2021), pp. 342–349.

[13] M. Janssen, F. J. Snoek, and R. J. Heine, *Assessing impaired hypoglycemia awareness in type 1 diabetes: agreement of self-report but not of field study data with the autonomic symptom threshold during experimental hypoglycemia.*, Diabetes Care, 23 (2000), pp. 529–532.

[14] I. L. Jernelv, K. Milenko, S. S. Fuglerud, D. R. Hjelme, R. Ellingsen, and A. Aksnes, *A review of optical methods for continuous glucose monitoring*, Applied Spectroscopy Reviews, 54 (2019), pp. 543–572.

[15] J. Kim, A. S. Campbell, and J. Wang, *Wearable non-invasive epidermal glucose sensors: A review*, Talanta, 177 (2018), pp. 163–170.

[16] S. P. Nichols, A. Koh, W. L. Storm, J. H. Shin, and M. H. Schoenfisch, *Biocompatible materials for continuous glucose monitoring devices*, Chemical Reviews, 113 (2013), pp. 2528–2549.

[17] D. Olczuk and R. Priefer, *A history of continuous glucose monitors (cgms) in self-monitoring of diabetes mellitus*, Diabetes & Metabolic Syndrome: Clinical Research & Reviews, 12 (2018), pp. 181–187.

[18] M. Phillip, R. Nimri, R. M. Bergenstal, K. Barnard-Kelly, T. Danne, R. Hovorka, B. P. Kovatchev, L. H. Messer, C. G. Parkin, L. Ambler-Osborn, et al., *Consensus recommendations for the use of automated insulin delivery technologies in clinical practice*, Endocrine Reviews, 44 (2023), pp. 254–280.

[19] P. P. Ray, *Continuous glucose monitoring: a systematic review of sensor systems and prospects*, Sensor Review, 38 (2018), pp. 420–437.

[20] E. R. Seaquist, J. Anderson, B. Childs, P. Cryer, S. Dagogo-Jack, L. Fish, S. R. Heller, H. Rodriguez, J. Rosenzweig, and R. Vigersky, *Hypoglycemia and diabetes: a report of a workgroup of the american diabetes association and the endocrine society*, The Journal of Clinical Endocrinology & Metabolism, 98 (2013), pp. 1845–1859.

[21] M.-S. Steiner, A. Duerkop, and O. S. Wolfbeis, *Optical methods for sensing glucose*, Chemical Society Reviews, 40 (2011), pp. 4805–4839.

[22] M. Tauschmann, H. Thabit, L. Bally, J. M. Allen, S. Hartnell, M. E. Wilinska, Y. Ruan, J. Sibayan, C. Kollman, P. Cheng, et al., *Closed-loop insulin delivery in suboptimally controlled type 1 diabetes: a multicentre, 12-week randomised trial*, The Lancet, 392 (2018), pp. 1321–1329.

[23] B. J. Van Enter and E. Von Hauff, *Challenges and perspectives in continuous glucose monitoring*, Chemical Communications, 54 (2018), pp. 5032–5045.

[24] Y. Yang and W. Gao, *Wearable and flexible electronics for continuous molecular monitoring*, Chemical Society Reviews, 48 (2019), pp. 1465–1491.

[25] E. Yeoh, P. Choudhary, M. Nwokolo, S. Ayis, and S. A. Amiel, *Interventions that restore awareness of hypoglycemia in adults with type 1 diabetes: a systematic review and meta-analysis*, Diabetes Care, 38 (2015), pp. 1592–1609.

Edited by: B. Nagaraj M.E.
Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing
Received: Oct 18, 2023
Accepted: Nov 25, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1557–1563, DOI 10.12694/scpe.v25i3.2710

DEEP MACHINE LEARNING-BASED ANALYSIS FOR INTELLIGENT PHONETIC LANGUAGE RECOGNITION

YUMEI LIU[*]AND QIANG LUO[†]

Abstract. Modern speech generating systems can produce results that are almost as visually realistic as actual sounds. They still require further production management. This research presents a paradigm for managing prosodic output using explicit, unambiguous, and understandable parameters. We utilize this strategy to emphasize key words and provide a variety of architectural possibilities based on a richness of labelled resources. In an objective voice, we compare the options for producing data with or without labels. We assess them using listening tests that demonstrate our ability to retain the same level of naturalness while effectively attaining regulated concentration over a specific area.

Key words: Prosody management, machine learning, speech analysis, lexical focus.

1. Introduction.

In today's digital age, the rapid development of human-computer interaction and natural language processing technology has led to widespread attention to Automatic Speech Recognition (ASR). Intelligent speech recognition systems have potential applications, ranging from virtual assistants and smart home controls to healthcare and education, providing users with convenient and efficient voice interaction methods. The progress in this field not only provides better experiences for individual users, but also provides more innovation and business opportunities for enterprises and organizations. The most advanced architectures now available offer high-quality results that typically approach or equal what is expected of natural language [7]. Apart from the high quality, these models contain a number of enticing features. They may jointly define multiple waveform properties, allowing correlations between them to be observed . Furthermore, they abandon typical pipeline designs in favour of a loosely connected, unified approach, which is desirable when some pipeline modules (for example, text processing in a foreign language) are difficult to construct. However, they do have some well-known drawbacks, including interpretability issues (it can be difficult to determine which parts of the model are responsible for what functions), controllability issues (it can be challenging to intervene in the model to influence some aspects of the synthesis, which is frequently desired, such as when providing SSML support), and potential instability issues (minor deviations at inference time can become worsened and gen. By adding methods to the architecture that the user can use to modify a particular output property, this work tackles the controllability issue [20]. Although usability factors are not the main emphasis of this project, we support a set of characteristics that will enable these options to be available to the system's end user.

- **Interpreting**: The listener should be able to hear and recognize the impact of a control change (for instance, whether speech is sluggish quicker, better-pitched, or seems more joyful, etc.) [11].
- **Monotonicity**: An aesthetic with sensory impacts that alter inversely when the user changes the knob feels better natural and is simpler to tune [2].
- **Low-dimensionality**: The user shouldn't need to alter many parameters to change the outcome. The model should be able to offer a low-dimensional customizable representation or step in and complete up defaults to take care of the user's work [5].

Disentanglement: While this may be challenging due to the multiple ways in which distinct expression variables interact, controlling the output along relatively separate (perceptual) dimensions is made simpler by a set of controls that are more dissociated from one another [17]. Speed and quantity, for example, may be tweaked independently without having to go back and alter a previously adjusted variable. We analyse the implementation and controllability of constrained lexical emphasis as a case study for those stated previously .

[*]Chongqing City Vocational College, Chongqing, 402160, China (Corresponding Author, YumeiLiu7@126.com)

[†]Chongqing Creation Vocational College, Chongqing, 402160, China (QiangLuo85@163.com)

We want to produce an expressive amount of importance that is distinct from prosodic accentuation by using a "neutral" wide focus. Consider how the phrase that has become a catchphrase responds to the circumstances below. In these instances, the speaker emphasises the focus area by separating the target object from its surroundings more clearly [12]. To produce flexible and suitable synthetic speech, control over the output expression must exist independently of spoken text. Since significant non-textual speech variance is rarely marked, output control must be learned unsupervised. In this research, we thoroughly investigate techniques for statistical speech synthesis unsupervised learning of control. For instance, we demonstrate how some auto encoder models can interpret standard unsupervised training techniques as variational inference. These new probabilistic interpretations' ramifications are examined. It encourages the potential of unsupervised learning to provide output control in speech synthesis in general [8].

Amortized inference-based methods are promising for upcoming applications since they provide comparable performance to existing heuristics, making training and latent-variable inference easier. We can force some latent variables to adopt consistent and understandable purposes by providing partial supervision to some of them, which was previously impossible with completely unsupervised TTS models. With as little as 1% (30 minutes) of care, our model can accurately find and regulate crucial but rarely tagged speech characteristics, such as effect and speaking rate [15].

2. Literature Survey. We can consistently and reliably learn to regulate specified parts of prosody, in contrast to earlier wholly unsupervised techniques. Any latent aspect of speech, continuous or discrete, for which a moderate amount of labelling can be collected, can be used with our method. When precise duration labels are unavailable or sparse in the training data, the proposed model can be trained explicitly with duration labels or unsupervised or semi-supervised using a fine-grained variational auto-encoder . When trained and wholly supervised, the proposed model slightly beats Tacotron 2 on naturalness [13]. The suggested model outperforms Tacotron 2 on naturalness with unsupervised or semi-supervised duration modelling while still much more resilient on over-generation and equivalent on under-generation.

At the time of inference, the duration predictor additionally offers per-phoneme and utterance-wide duration control. This study introduced the Non-Attentive Tacotron, which considerably exceeded Tacotron 2 in terms of robustness as measured by the unaligned duration ratio and word deletion rate while outperforming Tacotron 2 in terms of naturalness. Tacotron 2's attention algorithm was replaced with Gaussian upsampling and an explicit duration predictor to achieve this. We also demonstrated that the duration predictor could be used to change both the utterance's overall pacing and the rate at which individual syllables are spoken [18].

The technique works with both expanded state sequences—each corresponding to a single feature frame—and state sequences with defined durations. We also give a thorough examination blended sample' phonetic composition. The evaluation incorporates phonetically motivated, gradual, and universally applicable phonological processes and input-switch rules, encompassing the dialects' historically divergent phonological evolution against the standard language. We describe an expanded technique that uses a step function for input-switch practices while linearly interpolating phonological processes [10].

Our investigation shows that phonological knowledge of this kind improves dialect speakers' capacity to judge the dialect authenticity of synthesised speech. Our methods can be utilised to alter voice output systems because progressive alterations between kinds are a common occurrence. For state-level interpolation, it locates HSMM-state mappings using DTW. One feature frame is produced for each state using either of two techniques for dealing with state durations: either continuing with the unexpanded states or increasing every instance with a length of N to N states with an interval of 1 [4].

Our findings imply that DTW's linear interpolation of its mapped HSMM states was fair. The machine translation component of a comprehensive interpolation system would also translate a standard variety into dialect. We would get input switch rules for words from this component, which may also produce syntactic modifications. Phonetic criteria must be used to derive rules for phonemes [16].

3. Materials and Methods.

3.1. Design. The model has been enhanced with decoder-to-facilitator components. 25 January 2021 saw the addition of controls and increased stability during decoding. This series-to-sequence model generates an auditory spectroscopy (eds-prosodic model that is then fed to an asynchronously instructed, LPC-Net-based brain vocoder in order to generate high-quality samples in real-time.

- Emphasis anchoring (A) is a basic verification method based on binary indicative features, which is used to derive the emphasis focus in discourse. The basic principle of this method is to determine the position of emphasis focus by analyzing specific features in discourse, namely binary indicative features. This process can be regarded as an equation with the aim of verifying the existence and position of the focal point. In short, emphasis anchoring is a method of analyzing discourse structure by identifying specific binary indicative features to determine the focus of emphasis in discourse. This method helps to gain a more detailed understanding of the role and expression of emphasis in language [19].
- A front-end programming generator (C) that utilizes bidirectional short-term memory (Bi-LSTM) layers and convolutional layers to encode the combined embeddings from (A) and (B).
- To aid with training in a setting with several speakers, an overall phrases-level speaker embedding (D) is distributed throughout the episode.

The Decoder is an autoregressive network that modifies the fundamental architecture's concentration process, self-regressive input, target selection, and instructional costs. These are listed below and were previously covered. An improved two-stage attention system is created by using a technique that promotes monotonicity and unimodality in the alignment matrix following the Tacotron2's material- and GPS-based attention system [3]. This modification is essential for enhancing stability during inference, especially when there are external controls. The model is exposed to both the final ground truth output value and the initial projected value during training using a two-pronged feedback technique (i.e., inference mode and instructor forcing). At the time of inference, the anticipated value is replicated. The model also incorporates the parameters needed to anticipate the 80-dim mel cepstral characteristics from a separately trained LPC-Net neural vo coder. These traits—which we refer to as "LPC features"—include 22-dim vectors with 20 cepstral coefficients, log f0, and f0 correlation for 22kHz signals. Instead of using post-net refinement, the mel task processes the anticipated Lcp elements using two put up-nets (one to enhance the area was found and one to enhance the pitch-related parameters).

$$L = \mathrm{MSE}(\tilde{y}^M t, y^M t) + 0.8\mathrm{MSE}(\hat{y}^L t, y^L t) + 0.4\mathrm{MSE}(\tilde{y}^L t, y^L t) + 0.4MSE(\Delta\tilde{y}^L t, y^L t), \tag{3.1}$$

The modulo operator applies the starting time interval to the sequence, and MSE(,) represents the mean-squared error. To save space, we remove some information from this exposition and direct the reader to [5, 2] for more context and formulas.

3.2. Traditional Monitoring. This architecture depends on a Boolean indicator feature when labelled data is available. The audio signals' ground truth values are applied during training.

3.3. Without Supervision. The structure denoted by the letters A, C, D, as well as E, and so F, G, and H provide a way to increase the system's responsiveness during learning and to control the implementation of the prosodic rhythms at the time of inference using a configurable array of parameters (cf. the integer power of the controlled design) [6].

3.4. Mixed. Although parts D through I result from an unattended strategy, prosodic patterns may still be realized even in the presence of labelled data by working in conjunction with an explicit feature. We examine a "mixed" approach—specified by the entire framework Any–G—that combines controlled learning with technology to handle the circumstance without access to investigate this further [1].

3.5. Tiered Standard Mode Prosodic. The "layered standard pattern prosody" refers to an analytical model of language prosody, which is used to describe the organization of sound rhythm and phonological structure in language. This model typically includes multiple levels or levels to help understand different aspects of language rhythm. Following the motivation for a perceptually-interpretable, low-dimensional prosodic control mechanism discussed. We suggest an ordered collection of four prosodic regulators to condense information about a signal's length and pitch travel through linguistically significant and natural regions of the prosodic hierarchy. The strategy made it possible to manipulate common traits like general tempo. However, more control was needed to achieve the level of departure from long-term trends necessary to produce local emphatic concentration. These regulations apply to both domestic and international properties. They are a development of that strategy. Before we get there, let's define the following statistics [14].

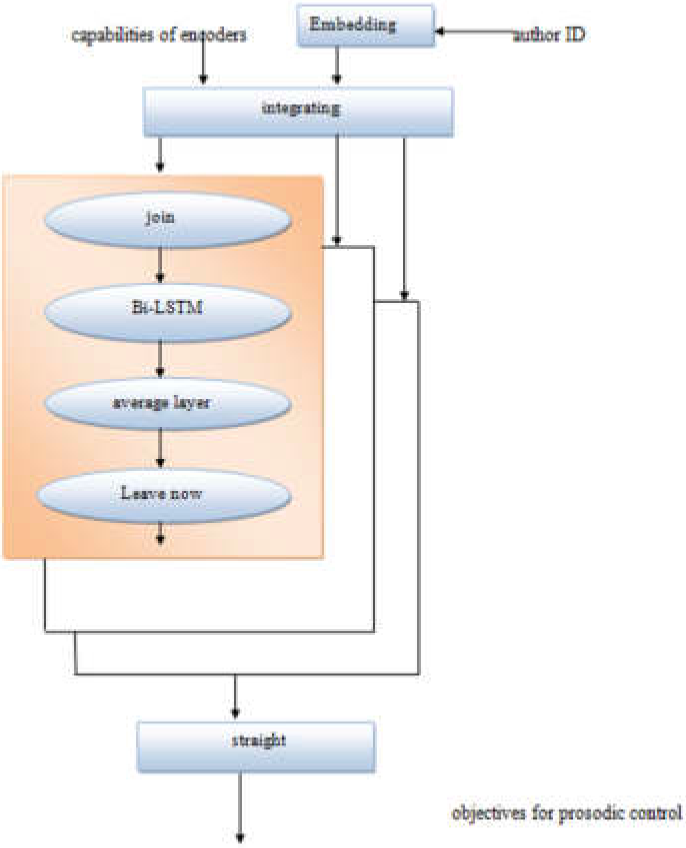


Fig. 3.1: Working flowchart of proposed Bi-LSTM for prosodic control

- S_{dur}: A record of the typical phone durations along a sentence (excluding quiet).
- S_{f0}: A "spread" of log-f0, which is the difference between its 95th and 5th percentiles, along a sentence.
- W_{dur}: A word-by-word log of the typical per-phone durations (discussed above).
- W_{f0}: A log-f0 "spread" along each word (as stated above).

$$PC = Norm_{\sigma}[S_{dur}, S_{f0}, W_{dur} - S_{dur}, W_{f0} - S_{f0}], \tag{3.2}$$

At the moment of inference, the prosodic-control subnet's predictions are adjusted to be roughly constant concerning the divination readings used to train the system. The evaluated systems' forecast distance between (known) phrase and term surround is stabilized using a mean pooling function. The design of the prosodic-control classifier. Models are trained using a speaker-embedding layer, whose output is fed into each tumbled interfere, utilizing a multi-speaker technique. We'll discuss how we create object sizes for each component of this architecture when we discuss its various elements [9]. Figure 3.1 describes the working flowchart of proposed Bi-LSTM for prosodic control.

4. Experimentation and Results. Three datasets from three native US English users who are employed as professionals made up the instructional stuff, which was broken down as follows:

- A canon from a male speaker (M1) with approximately 10.8K sentences.
- A corpus from the same male speaker (M1emp) with about 1K sentences containing multiple words with emphasis.

As part of the corpus M1emp, a speaker was taught to realise an emphatic prominence on the words that should receive the most emphasis inside each penalty. His realisations of prosody depart greatly from broad focus prosody in terms of tempo, comparative diameter, stress height, and disjunction from the pertinent content. The sentences were meant to provoke certain limited-focus situations, such as contrast, disambiguation, etc.

Keep in mind that this sample is much smaller than the fundamental texts, and that only one speaker is included in the tagged data. On average, three emphatic words were present in each sentence in M1emp, and their overall frequency was roughly 23%. We are interested in comparing entirely unsupervised approaches that are feasible within the framework outlined in Section 2 with processes that use labelled data (where available) to determine the relative merits of each method.

To achieve that goal, take into account the following systems:

- Basis (NoEmph): A typical sentence-to-sentence communication system solely using worldwide prosodic limitations. The emphatic data is part of the learning batch (Demp), but no other focus-marking element is used.
- Basis (Sup): A basic system that uses Traditional Monitoring (as described in Section 2), world supervision, Demp training, and an apparent byte attribute expressing to determine the stress area.
- PC-Unsup: A fully unattended system with changeable prosodic control (as said in Section 2), where the prosody forecast and parts have been taught using Dbase.
- PC-Hybrid: a combination of models trained with Demp that gives precise Conditional accent signals and changing prosodic control, comparable to the Standard (Sup) system.

Comprehensive explanations and evaluation of the Base (NoEmph) layout with worldwide controllers may be found in [5]. However, in this instance, it acts as a reliable reference point for overall excellence to ensure that the alternative concepts support the benefits of naturalness provided by this technique [20]. LPC-Net is a vocoder model used in the field of speech synthesis and processing. It combines Linear Predictive Coding (LPC) and neural network technology to generate natural speech synthesis. Speaker training is typically used for voice modeling of vocoder models to simulate the speech features of different speakers. The model was chosen and tuned using the following steps. In order to scan the grid across architectures and track the held-out loss to gauge learning speed, 10% of the prosodic sub-networks training data were first withheld.In both instances, the speaker embedding had a size of 20. The remaining extreme parameters of the various combinations were then perceptually adjusted after this was taken care of.The beneficial boost rates that we choose are consistent with the empirical findings in the M1emp group and our theoretical hypotheses, which show that specialised items have longer speaking durations and more prominent tone accents. We found that the Base (Sup) system boosted our pitch incursions when adjusting the Hybrid systems because it recognised these tempo shifts rather well. After fine-tuning a single set of boosting parameters, we frequently find that it performs brilliantly, spanning a variety of words and dialects. Figure 4.1 defines the examples of the four prosodic controllers' phonetic trajectories for a two-sentence input.

5. Conclusion. We have developed and tested a method that enables more precise word syntax management to direct a comprehension of tight focus in the composite. The system is composed of consumer-driven regulations that follow the ideas we have articulated and supported. They offer a structure that divides different prosodic components (duration and pitch) so they can be altered separately. They are intuitive in the sense that changes to the way control is passed on to visual perception affect the outcome. They portray prosody in a flat manner. We have demonstrated that the method only requires additional data for simple word alignments. Different levels of monitoring can be accommodated with the necessary resources. Overall, deep machine learning has greatly improved the performance of intelligent speech recognition, enabling ASR systems to be widely used in daily life. However, there are still some challenges that need to be addressed, such as improving the robustness and accuracy of high-end end-to-end ASR systems to meet the needs of different application fields. This field is still constantly developing, and there will be more innovation and improvement in the future.

Acknowledgement. Supported by the university-level platform of Digital Creative Design Production-Education Integration Collaborative Innovation Center

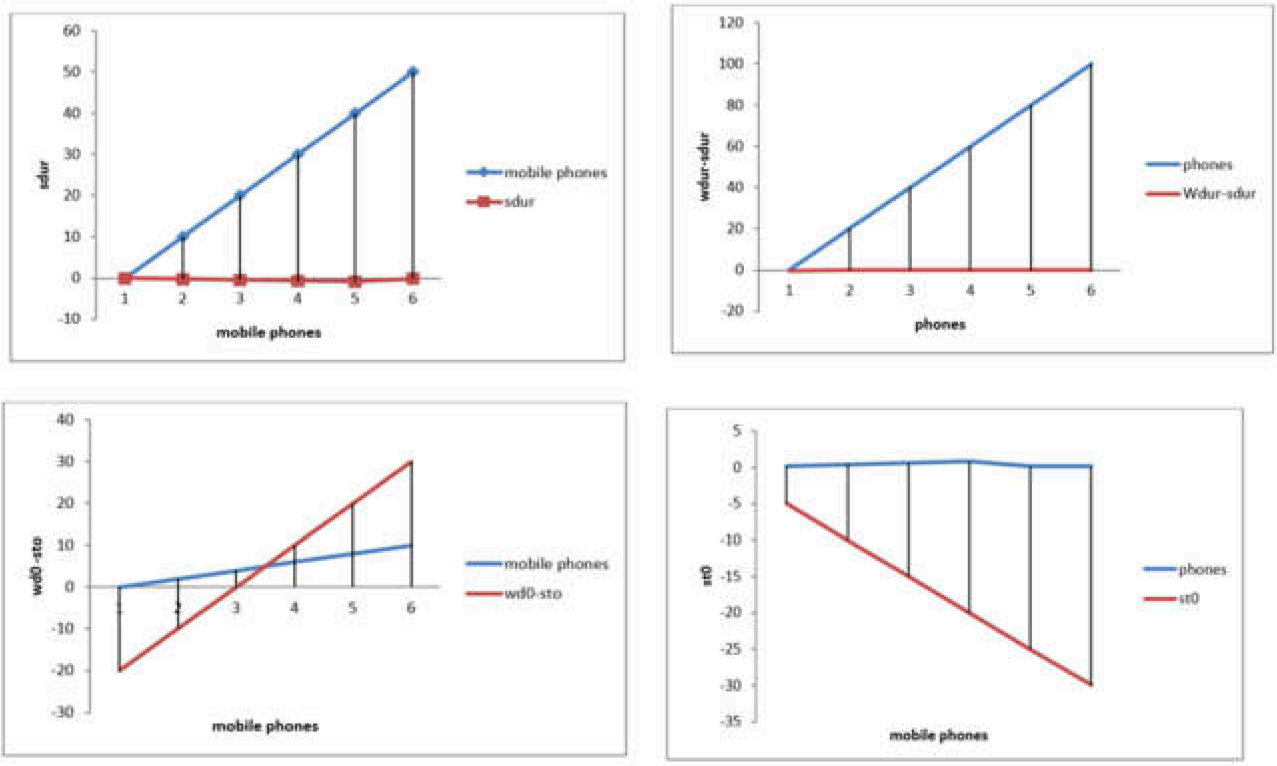


Fig. 4.1: Examples of the four prosodic controllers' phonetic trajectories for a two-sentence input

REFERENCES

[1] P. Ajay, B. Nagaraj, and J. Jaya, *Bi-level energy optimization model in smart integrated engineering systems using wsn*, Energy Reports, 8 (2022), pp. 2490–2495.

[2] P. Ajay, B. Nagaraj, R. A. Kumar, R. Huang, and P. Ananthi, *Unsupervised hyperspectral microscopic image segmentation using deep embedded clustering algorithm*, Scanning, 2022 (2022).

[3] J. L. Ba, J. R. Kiros, and G. E. Hinton, *Layer normalization*, arXiv preprint arXiv:1607.06450, (2016).

[4] R. Fernandez and B. Ramabhadran, *Automatic exploration of corpus-specific properties for expressive text-to-speech: A case study in emphasis*, in Proceedings of the 6th ISCA Workshop on Speech Synthesis, vol. 90, Bonn, Germany, 2007.

[5] W. Hsu, Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang, *Hierarchical generative modeling for controllable speech synthesis*, in Proceedings of the 7th International Conference on Learning Representations, New Orleans, LA, USA, 2019, OpenReview.net.

[6] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, in Proceedings of the 3rd International Conference on Learning Representations, Y. Bengio and Y. LeCun, eds., San Diego, CA, USA, 2015.

[7] V. Klimkov, S. Ronanki, J. Rohnke, and T. Drugman, *Fine-grained robust prosody transfer for single-speaker neural text-to-speech*, arXiv preprint arXiv:1907.02479, (2019).

[8] Y. Lee and T. Kim, *Robust and fine-grained prosody control of end-to-end speech synthesis*, in Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 2019, IEEE, pp. 5911–5915.

[9] Y. Mass, S. Shechtman, M. Mordechay, R. Hoory, O. S. Shalom, G. Lev, and D. Konopnicki, *Word emphasis prediction for expressive text to speech*, in Proceedings of the 19th Annual Conference of the International Speech Communication Association, B. Yegnanarayana, ed., Hyderabad, India, 2018, ISCA, pp. 2868–2872.

[10] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny, *The ibm expressive text-to-speech synthesis system for american english*, IEEE Transactions on Audio, Speech, and Language Processing, 14 (2006), pp. 1099–1108.

[11] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, *Fastspeech 2: Fast and high-quality end-to-end text to speech*, arXiv preprint arXiv:2006.04558, (2020).

[12] A. Sharma, A. Singla, N. Sharma, D. Gowda, et al., *Iot group key management using incremental gaussian mixture model*, in Proceedings of the 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC),

Coimbatore, India, 2022, IEEE, pp. 469–474.

[13] S. Shechtman, C. Rabinovitz, A. Sorin, Z. Kons, and R. Hoory, *Controllable sequence-to-sequence neural tts with lpcnet backend for real-time speech synthesis on cpu*, arXiv preprint arXiv:2002.10708, (2020).

[14] S. Shechtman and A. Sorin, *Sequence to sequence neural speech synthesis with prosody modification capabilities*, CoRR, abs/1909.10302 (2019).

[15] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, et al., *Natural tts synthesis by conditioning wavenet on mel spectrogram predictions*, in Proceedings of the 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), Calgary, AB, Canada, 2018, IEEE, pp. 4779–4783.

[16] V. Strom, A. Nenkova, R. Clark, Y. Vazquez-Alvarez, J. Brenier, S. King, and D. Jurafsky, *Modelling prominence and emphasis improves unit-selection synthesis*, in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 2, 2007, pp. 1282–1285.

[17] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, and Y. Wu, *Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis*, in ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), Barcelona, Spain, 2020, IEEE, pp. 6264–6268.

[18] J.-M. Valin and J. Skoglund, *LPCNet: Improving neural speech synthesis through linear prediction*, in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, IEEE, pp. 5891–5895.

[19] K. Yu, F. Mairesse, and S. Young, *Word-level emphasis modelling in hmm-based speech synthesis*, in Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, Texas, USA, 2010, IEEE, pp. 4238–4241.

[20] Y. Zhang, S. Pan, L. He, and Z. Ling, *Learning latent representations for style control and transfer in end-to-end speech synthesis*, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, United Kingdom, 2019, IEEE, pp. 6945–6949.

Edited by: B. Nagaraj M.E.
Special issue on: Deep Learning-Based Advanced Research Trends in Scalable Computing
Received: Oct 18, 2023
Accepted: Nov 25, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1564–1577, DOI 10.12694/scpe.v25i3.2740

DESIGN AND IMPLEMENTATION OF CHINESE LANGUAGE TEACHING SYSTEM BASED ON VIRTUAL REALITY TECHNOLOGY

TIANYA YANG* AND JIALING WU†

Abstract. Traditional Chinese language teaching methods have always had problems, such as a lack of practical opportunities and interest and an inability to provide sufficient language environment and context. Virtual reality technology provides new possibilities to solve these problems. Through in-depth research and analysis, this article designs and develops a Chinese language teaching system based on virtual reality technology. This system utilizes virtual reality technology to create a simulated 3D Chinese language environment, enabling learners to experience interactive experiences firsthand. The system includes speech recognition, natural language processing, and artificial intelligence, enabling real-time language communication and student interaction. This system s design considers students learning needs and interests, allowing them to practice their Chinese application skills in real-world simulated environments. This system has broad prospects in Chinese language education and provides new research and innovation directions for educational institutions and developers.

Key words: virtual reality technology, Chinese language teaching, system design, teaching innovation

1. Introduction. Chinese is one of the most widely used languages in the world. With China s rise and globalization, more and more people have become interested in learning Chinese. They hope to master Chinese to facilitate communication with Chinese people, conduct business transactions, and gain a deeper understanding of Chinese culture [1]. In this context, Chinese language teaching has become an important topic, and Chinese language education faces a series of problems and challenges. On the one hand, the complexity and uniqueness of the Chinese language bring great learning difficulties to learners. As ideographic characters, learning Chinese characters requires a lot of effort and time, and learning Pinyin systems also require a certain amount of time and skills. In addition, the grammatical system of Chinese is different from that of Indo-European languages, with significant differences in grammatical structure and sentence organization, which is far less simple and intuitive than Japanese and English. Therefore, in response to this problem, more intelligent and efficient teaching methods and resources are needed to help students master Chinese faster and more systematically.

On the other hand, traditional Chinese language education also needs some help. For example, the content of textbooks needs to be more flexible, teaching methods are single, and communication opportunities are limited. To address these issues, it is necessary to carry out educational innovation, provide more practical textbooks and learning tools, and provide students with more communication opportunities and independent learning platforms to meet their diverse learning needs [2]. Correspondingly, opportunities and challenges are also required for Chinese language education. In the current era of rapidly changing information technology, various advanced technologies such as natural language processing, machine translation, virtual reality, and new media can be utilized to promote the development and innovation of Chinese language education [3, 4]. With the rapid development of technology, the application of Virtual Reality (VR) technology in education is receiving increasing attention [5]. Virtual reality technology can create simulated three-dimensional environments [6], allowing users to have immersive interactive experiences [7]. This immersive learning approach has brought new possibilities to education and has been widely applied in various disciplinary fields [8, 9]. In the field of Chinese education, there are some problems in the traditional teaching methods. Learners usually need to learn Chinese characters and grammar knowledge through classroom teaching or written materials, but this learning mode often lacks sufficient practical opportunities, especially for non-native language learners,

*Faculty of Boya, Geely University of China, Chengdu 641423, China
†Faculty of Education, Sichuan Normal University, Chengdu 610066, China (Corresponding author, Wu_Jialing23@outlook.com)

who need to improve their oral fluency and communicative ability, while traditional teaching methods cannot provide adequate language environment and situation [10], Learners can immersivity experience the Chinese language environment and engage in real-time language communication and interaction with virtual characters [11]. Compared with traditional textbooks, virtual reality technology has higher interest and appeal, which can stimulate students learning interest and enthusiasm [12]. In addition, virtual reality technology can also create various scenarios and scenarios [13], such as shopping, tourism, business negotiations, etc., to help students practice Chinese language application skills in real-life simulated environments. Virtual reality technology has been applied to a certain extent in Chinese language education [14, 26]. Some educational institutions and research teams have begun exploring virtual reality technology for Chinese language teaching, designing, and developing Chinese language teaching software and systems based on virtual reality technology, providing learners with an immersive learning experience [16, 17]. These systems typically include speech recognition, natural language processing, and artificial intelligence, enabling students to engage in dialogue, communication, and interaction with virtual characters. The application of virtual reality technology in Chinese language teaching covers multiple aspects, helping students improve their listening and speaking abilities. Students can engage in real-time dialogue and interaction with virtual characters through virtual reality technology, improving their oral expression and listening comprehension abilities [18]. Secondly, virtual reality technology can also create various scenarios and scenarios, enabling students to practice their Chinese application skills in real-life simulated environments, such as shopping in shopping malls and ordering at restaurants. In addition, virtual reality can also provide immediate feedback and personalized guidance, helping students correct pronunciation and grammar errors and improve learning outcomes. However, the application of virtual reality technology in Chinese language education is still in its infancy, and there are some challenges and problems. First, the technical aspects need to choose the appropriate hardware equipment and stable and reliable software platform for development. Second, content creation requires collaboration between educational experts, linguists, technologists, and content creators. Learner interactions need to be designed in ways that are effective and employ speech recognition and natural language processing techniques to provide timely feedback. Finally, the maintenance and updating of the system requires the establishment of a dedicated team and close cooperation with educational institutions and students. Despite these challenges, virtual reality technology still has broad prospects in Chinese language education. With the continuous progress of technology and the accumulation of application cases, virtual reality technology will gradually become an essential component of Chinese language teaching, providing students with a more immersive, personalized, and practical learning experience.

This study aims to design and implement a Chinese language teaching system based on virtual reality technology to improve the effectiveness and fun of Chinese language learning. Through this system, students can use virtual reality devices to enter simulated natural scenes and engage in language communication and interaction with virtual characters, thereby improving their language application and communication skills. In addition, the system will also provide a rich and diverse Chinese language environment and scenarios, helping students flexibly use Chinese in different situations and improving their comprehensive language abilities. The research in this article is expected to provide a new teaching tool and method for teachers and learners in the field of Chinese language education, promoting innovation and development in Chinese language education.

2. Design Scheme of Chinese Language Teaching System Based on Virtual Reality Technology.

2.1. System Designing Objective. The Chinese language teaching system based on VR technology is an innovative solution based on virtual reality technology aimed at helping learners better learn Chinese and understand Chinese culture [19]. Through centralized management of teaching resources, the system provides remote teaching and three-dimensional virtual teaching scenes to meet the needs of modern learners [20, 21]. As shown in Table 2.1, the advantages and disadvantages of mainstream software and a comparison table of application fields are presented. 3ds Max has advantages such as ease of use, rich architectural and renderings, and good interactivity, making it particularly suitable for designing Chinese language teaching systems as a virtual reality technology.When using 3Ds Max for Chinese learning, students can further enhance their learning interest and effectiveness by creating a virtual reality environment. Students can choose courses through the gaze function, learn details, and freely control Settings such as sound. The system sets up 3D models and animations for teachers, and the teachers will react accordingly according to the content in the class, increasing the sense of reality.

Table 2.1: Comparison Table of Advantages, Disadvantages, and Application Fields of Mainstream Software

Software Name	**Advantages**	**Disadvantages**	**Main application areas**
Maya	Strong rendering ability and high modeling accuracy	Difficult to master	Film and television industry
C4D	Good rendering effect and comprehensive functionality	Poor animation effect	TV packaging and advertising
Blender	Open source, comprehensive functionality	Poor FBX support	3D animation
3ds Max	Easy to use, low hardware requirements, and plugins	Slow update and large volume	Rich architecture and renderings
Zbrush	The modelling process can be freely utilized.	Poor topology	Digital carving and painting
Solid works	High modelling accuracy and good model proportion	High design requirements	Mechanical manufacturing, non-standard design

Table 2.2: Comparison Table of VR Engine Characteristics

Name	**Characteristic**
Unity 3D	Strong system compatibility and quick to get started
Unreal Engine 4	Good screen display effect, high hardware requirements, and simple visual programming
Cry ENGINE	Fully functional and challenging to learn

Develop teaching software based on virtual reality technology, and use VR headsets for teaching to provide a more realistic virtual reality experience. The design of the system is an auxiliary teaching tool for school Chinese classroom to help students better understand Chinese knowledge and improve language ability. Through virtual reality technology, students can better understand Chinese pronunciation, intonation, grammar, and other aspects, improving learning effectiveness and fun. Secondly, the system supports remote teaching, allowing students to participate in learning anytime and anywhere without being limited by time and space.

2.2. Virtual Reality Engine Selection. To achieve the goal of experimental teaching, it is necessary to design and develop human-machine interaction functions for virtual scenes, and human-machine interaction needs to be developed and implemented through simulation running platforms. Currently, the commonly used simulation interaction platforms at home and abroad mainly include Unreal Engine 4, Unity 3D, Cry Engine, Cocos 3D, and others. These engines have their characteristics, and Table 2.2 shows a comparison table of the characteristics of each engine.

Considering factors such as one's learning level and system development cycle, Unity 3D was ultimately chosen as the simulation development platform for system functions. At the same time, Unity3D comes with a high-performance lighting system, supports FBX format model files, supports Direct and Open GL low-level rendering, has a concise and easy-to-understand development interface, supports multiple scripting languages, and has realistic particle effects, which can fully meet the development needs of the virtual simulation experiment teaching platform for motors.

2.3. Texture mapping technology. In order to make the 3D model built in 3ds Max closer to natural objects, simple color adjustments are made to the geometric features of the model to present good surface texture details. Therefore, texture mapping technology is selected to optimize the texture surface of the model, which can significantly enhance the realistic visual effect and level of detail of the 3D model [22].

2.3.1. Texture Mapping Principle. The process of texture mapping is to map texture pixels from texture space to screen space. During the entire mapping process, there are mainly two types of mapping relationships: first, mapping from the texture coordinate system to the world coordinate system; The second

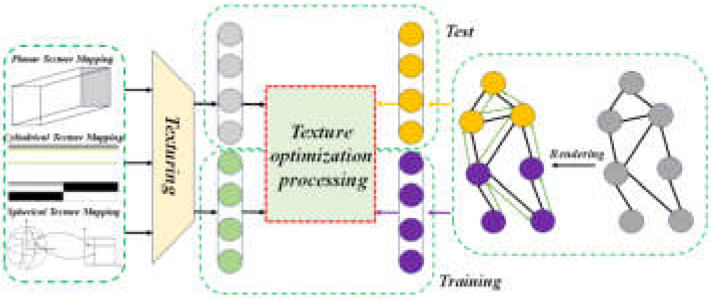

Planar Texture Mapping
Cylindrical Texture Mapping
Spherical Texture Mapping
Texturing
Texture optimization processing
Test
Training
Rendering

Fig. 2.1: Texture Mapping Diagram

is to map from the world coordinate system to the screen coordinate system, and the texture mapping process is shown in Figure 2.1. According to the mapping process, assigning corresponding color information to the vertices on the model surface can reflect the texture information of the object surface.

2.3.2. Common Texture Mapping Methods. A regular lighting model whose texture is only generated when surface properties change. The expression for a regular lighting model is given by formula (2.1):

$$I = I_a K_a + I_p + K_d(N \times L) + I_p K_s(N \times S)^n \tag{2.1}$$

Among them, Ia is the light intensity, Ka is the reflection coefficient, Ip represents the incident light intensity, Kd is the diffuse reflection coefficient, and Ks is the specular reflection coefficient. From this expression, it can be seen that by changing the average vector of the model surface or the diffuse reflection coefficient value, the cooler of the model itself can be changed.There are many mapping methods from texture mapping to image space, which can be divided into forward mapping and reverse mapping according to the mapping relationship between them. At present, plane texture mapping, cylindrical texture mapping and spherical texture mapping are mainly used, and the main application objects are non-parametric models.

1. **Planar texture mapping:** Planar Texture Mapping considers a plane as a basic geometric shape. Then, it projects a texture map from a two-dimensional plane onto it to simulate the appearance and details of the plane. For example, mapping a rectangular ABCD, making $|AB| = |CD| = |X|, |BC| = |AD| = |Y|$, establishing a correspondence between $0 \leq u \leq 1 and - \frac{X}{2} \leq x \leq \frac{X}{2}$ and establishing a correspondence between, $0 \leq v \leq 1 and - \frac{Y}{2} \leq x \leq \frac{Y}{2}$. The mapping function of this mapping method can be obtained from formula (2.2):

$$\begin{cases} u = \frac{1}{X}x + \frac{1}{2}, & -\frac{X}{2} \leq x \leq \frac{X}{2} \\ v = \frac{1}{Y}y + \frac{1}{2}, & -\frac{Y}{2} \leq y \leq \frac{Y}{2} \end{cases} \tag{2.2}$$

2. **Cylindrical texture mapping:** Cylindrical Texture Mapping considers a cylinder as a basic geometric shape. Then, it projects a texture map from a plane onto the surface of the cylinder to simulate its appearance and details. Assuming a cylindrical surface with a radius of r and a height of h, its relationship can be expressed by formula (2.3):

$$(u, v) = (\theta, z), \quad 0 \leq \theta \leq 2\pi \tag{2.3}$$

The parameter equation for the cylindrical surface can be obtained by formula (2.4):

$$\begin{cases} x = r\cos u \sin v \\ y = r\sin u \sin v \\ z = r\cos v \end{cases} \tag{2.4}$$

Mapping to a cylindrical surface through linear transformation as formula (2.5):

$$\begin{cases} u = \frac{2\pi s}{3} 0 \leq s, t \leq 1 \\ v = t \end{cases} \tag{2.5}$$

The parameter equation of the cylindrical surface is expressed as follows formula (2.6):

$$\begin{cases} u = \tan^{-1}(y/x) \\ v = z \end{cases} \tag{2.6}$$

The inverse transformation equation of the final linear transformation is as formula (2.7):

$$\begin{cases} u & = \frac{3u}{2\pi} = \frac{3}{2\pi} \tan^{-1}(y/x) \\ t = v = z \end{cases} \tag{2.7}$$

3. **Spherical texture mapping:** In spherical texture mapping, a sphere is considered as a basic geometric shape, and a texture map on a plane is projected onto the surface of the sphere to simulate its appearance and details. For example, a point u1 on the$u-axis$ of the UV Cartesian coordinate system is vertically mapped onto a meridian with a longitude of θ in the spherical coordinate system; A point v1 on the v axis of the UV Cartesian coordinate system is mapped horizontally onto a latitude line of ϕ in the spherical coordinate system. From the above process, the functional relationship between the texture coordinates (u, v) and the spherical coordinate system can be determined, as shown in formula (2.8):

$$(u, v) = (\theta, \varphi)(0 \leq \theta \leq \pi/2, \pi/4 \leq \varphi \leq \pi/2 \tag{2.8}$$

For a sphere, the parameters on it can be expressed using formulas (2.9):

$$\begin{cases} x = r\cos(\phi)\sin(\theta) \\ y = r\sin(\phi)\sin(\theta) \\ z = r\cos(\phi) \end{cases} \tag{2.9}$$

Map to a sphere through linear transformation, as shown in formula (2.10):

$$\begin{cases} u = \frac{\theta}{2\pi} \\ v = \frac{\varphi}{\pi} \end{cases} \tag{2.10}$$

The resulting mapping relationship is shown in formula (2.11):

$$\begin{cases} u = \frac{\arctan\left(\frac{y}{x}\right)}{2\pi} \\ v = \frac{\arccos(z)}{\pi} \end{cases} \tag{2.11}$$

2.4. Improved Matrix Decomposition Algorithm for Implicit Semantic Model. After analysis, it can be found that the user s mastery of knowledge points is not fixed as assumed by conventional implicit semantic models [23] but gradually decreases over time [24, 25]. Therefore, the traditional matrix decomposition based on the implicit semantic model needs to meet the actual situation of this article. This article improves the traditional matrix decomposition based on the implicit semantic model by combining the scale of knowledge forgetting, using the knowledge point correlation matrix as the algorithm s input. It proposes an improved implicit semantic matrix decomposition method.

1. **Time effect function** People s memory of information will gradually fade over time in daily learning and life. The forgetting curve of memory information proposed by German psychologist Hermann Ebbinghaus indicates that people gradually forget the knowledge they acquire over time, with the later

parts forgetting more slowly. Based on this, this article introduces a time effect function to evaluate the user knowledge point mastery trend over time. However, after directly introducing the forgetting curve, the time effect influencing factors of the two-time points closer to the virtual teaching test interval will be excessively amplified, while the time effect influencing factors of the two-time points more distant from the test interval will be ignored by the base because they are too small. Based on the above reality, combined with the Ebbinghaus forgetting law and the user s mastery of knowledge points, the time effect function is defined as formula (2.12):

$$f(t) = \mu + (1 - \mu) \cdot (1 - e^{-(t-t_0)}) \tag{2.12}$$

Among them, t represents the current time, and 0t represents when the user conducted virtual teaching tests. μ is the time effect influencing factor, representing the degree of time effect on the user, used better to match the user s mastery of knowledge points, and$\mu \in [0, 1]$. If $\mu = 0$, it is considered that forgetting the user s knowledge completely follows the time effect function. Conversely, if µ=1, it is not followed. As time goes by, the content of knowledge points forgotten by users will increase, and the score loss of users on specific knowledge points will gradually increase, reflecting a continuous decline in users' mastery of knowledge points.

2. **Recommendation algorithm incorporating time effects.** According to the implicit semantic model and formula definition, formulas (2.13) and (2.14) are obtained:

$$P = U_{m\times r} \sum_{r\times r} \tag{2.13}$$

$$Q = V_{n\times r} \tag{2.14}$$

The user knowledge point mastery matrix R can be expressed as formula (2.15):

$$R = PQ^T \tag{2.15}$$

Among them, $Pm \times f$ is the implicit semantic matrix of the users classification of knowledge point attributes, $Qn \times f$ is the proportion and weight of each knowledge point in attribute classification. The user s mastery of knowledge points $\widehat{(r_{ui})} = R(u, i)$ can be transformed into formula (2.16):

$$(\widehat{r}_{ui}) = R(u, i) = \sum_{f=1}^{F} p_{uf} q_{if} \tag{2.16}$$

Among them, the implicit class $f \in (1, F], p_{ui} = p(u, i)qif = Q(i, f)$. This article assumes that the difference between the actual results of user's knowledge mastery and the predicted results of knowledge points follows a Gaussian distribution. A knowledge point correlation matrix that integrates multiple factors is constructed as the distribution matrix of existing users' knowledge mastery, and the relevant parameters of the implicit class matrices P and Q are calculated. This article uses root mean square error (RMSE) to evaluate the degree of consistency between the predicted results and the user s actual grasp, and the loss function is expressed as $C(p, q)$ in formula (2.17) and (2.18).

$$\text{RMSE} = \sqrt{\frac{\sum_{(u,i)\in T}(r_{ui} - \widehat{r}_{ui})^2}{T}} \tag{2.17}$$

$$C(p, q) = \sum_{(u,i)\in \text{Train}} (r_{ui} - \hat{r}_{ui})^2 = \sum_{(u,i)\in \text{Train}} (r_{ui} - \sum_{f=1}^{F} p_{uf} q_{if})^2 \tag{2.18}$$

To prevent overfitting during the learning process, it is necessary to add overfitting terms to formula (2.18), then the equation is transformed into formula (2.19):

$$C(p,q) = \sum_{(u,i)\in\text{Train}} \left((r_{ui} - \sum_{f=1}^{F} p_{uf}q_{if})^2 + \gamma\|p_u\|^2 + \gamma\|q_i\|^2 \right) \tag{2.19}$$

Optimize the loss function $C(p,q)$ using the gradient descent method (SGD), and optimize the p_{uf} and q_{if} the partial derivative can be obtained as formula (2.20) and (2.21):

$$\frac{\partial C}{\partial p_{uf}} = -2f(t)\left(r_{ui} - \sum_{f=1}^{F} p_{uf}q_{if} \right) q_{if} + 2\gamma p_{uf} \tag{2.20}$$

$$\frac{\partial C}{\partial p_{uf}} = -2f(t)\left((r_{ui} - \sum_{f=1}^{F} p_{uf}q_{if})q_{if} + 2\gamma q_{if} \right) \tag{2.21}$$

After multiple iterations of optimization, the following results were obtained as formula (2.22) and (2.23):

$$p_{uf} = p_{uf} + \tau\left((r_{ui} - \sum_{f=1}^{F} p_{uf}q_{if})q_{if} + \gamma p_{uf} \right) \tag{2.22}$$

$$p_{if} = p_{if} + \tau\left((r_{ui} - \sum_{f=1}^{F} p_{uf}q_{if})q_{uf} + \gamma p_{if} \right) \tag{2.23}$$

Among them, τ is the learning efficiency, and the larger the value of Among them, τ is the learning efficiency, and the larger the value of τ, the faster the gradient decreases. Its value needs to be obtained through multiple experiments, where $\tau = 0.02$. After multiple iterations above, the parameters $p_u f$ and $q_i f$ can be obtained, and then the user s knowledge point mastery score can be predicted.According to this algorithm, the corresponding implicit semantic matrix P and Q can be obtained. After calculation, a complete user knowledge mastery matrix can be obtained. Based on the predicted value of this matrix, the corresponding weak knowledge points can be recommended for users. The larger the value, the higher the recommendation degree, and vice versa., the faster the gradient decreases. Its value needs to be obtained through multiple experiments, where $\tau = 0.02$. After multiple iterations above, the parameters $p_u f$ and $q_i f$ can be obtained, and then the user s knowledge point mastery score can be predicted.According to this algorithm, the corresponding implicit semantic matrix P and Q can be obtained. After calculation, a complete user knowledge mastery matrix can be obtained. Based on the predicted value of this matrix, the corresponding weak knowledge points can be recommended for users. The larger the value, the higher the recommendation degree, and vice versa.

2.5. Overall System Framework Design. The VR Chinese language teaching system is an innovative solution based on virtual reality technology, which utilizes VR technology to create a virtual classroom with a sense of realism and immersion. Its focus is to provide high-quality Chinese language education resources for users in different geographical locations, making Chinese language learning more convenient and efficient. The development engine of the system adopts 3ds Max, with the 3D model module and human-computer interaction module being the two core modules of the system. The 3D model module is mainly used to construct various models in the scene, while the human-computer interaction module is responsible for achieving interaction between users and the system. In the 3D model module, the 3ds Max software performs 3D modelling of the objects required for teaching. Various technologies are used to construct the required models during the

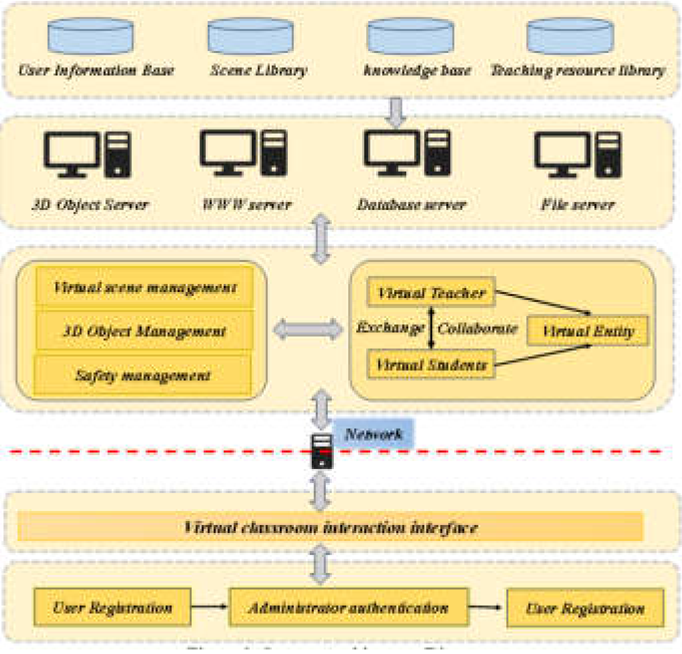


Fig. 2.2: System Architecture Diagram

modelling process quickly. At the same time, in order to improve the realism and quality of the model, it is necessary to render the relevant model. 3ds Max software has a very powerful modeling function, but its disadvantage is that the model data is often large, which is not conducive to scene fusion and network transmission. Therefore, after the model construction is completed, optimizing the established 3D virtual model is necessary to make it more suitable for VR teaching systems. In the human-computer interaction module, it is mainly necessary to consider the interaction between users and the system. It is necessary to design a set of interactive methods to connect users with the system and ensure that users can quickly and conveniently obtain the required information and learning materials [15]. The Unity engine is mainly used during this process, and programming uses the # language. In this way, functions such as interactive operations and data transmission can be easily implemented. The system architecture diagram is shown in Figure 2.2.

3. Construction of Chinese Language Teaching System Based on Virtual Reality Technology.

3.1. Implementation of Chinese Language Teaching System. 3D scene synthesis is a method that utilizes computer technology to transform real or fictional scenes into 3D models and endow them with functionality and interactivity [27]. It is mainly divided into three stages: the data collection stage, scenario modelling stage, and function implementation stage. The data collection stage is to obtain basic information about the scene, including teaching scene maps, authentic images, and model planning maps. These data can help designers determine the structure and style of the scene [28]. The scene modelling stage is to create a three-dimensional model of the scene, including text, graphics, texture mapping processing, 3ds Max modelling, and Vary rendering, to give the scene a realistic appearance and effect. The function implementation stage adds functionality and interactivity to the scene, including gaze function, animation function, interactive Chinese teaching system, and model planning.

The perception layer in virtual teaching testing is the simulation and reproduction of the perception layer in the three-layer structure of the real Internet of Things. This part mainly consists of virtual sensors and intelligent devices, which only exist in the simulation experimental environment but have data parsing and transmission functions, mainly including receiving or sending control instructions and parsing the received control instructions according to a specific protocol format; finally, based on the analysis results, the linkage

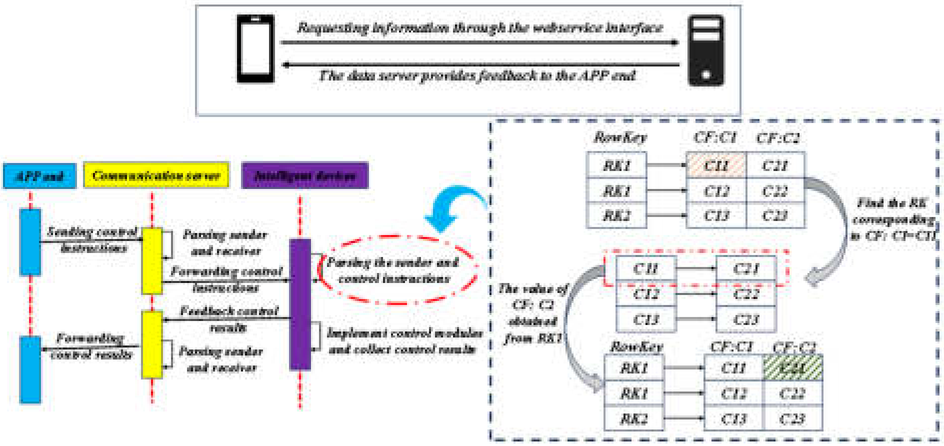

Requesting information through the webservice interface
The data server provides feedback to the APP end
APP end
Communication server
Intelligent devices
Sending control instructions
Parsing sender and receiver
Forwarding control instructions
Parsing the sender and control instructions
Feedback control results
Implement control modules and collect control results
Forwarding control results
Parsing sender and receiver
RowKey
CF:C1
CF:C2
RK1
RK1
RK2
C11
C12
C13
C21
C22
C23
Find the RK corresponding to CF: C1=C11
The value of CF: C2 obtained from RK1

Fig. 3.1: Information transmission process of VR Chinese language teaching system

effect of devices in the virtual scene is achieved [29]. The network layer in the virtual teaching testing of the Internet of Things simulates the communication process between users and intelligent devices in the scene. It can be divided into two parts: the communication server side and the data server side. The network layer plays an indispensable role between users and smart devices. It receives control instructions issued by users through the mobile app, parses the instructions in a specific format, and forwards them to the smart devices in the scene. In turn, it can also transmit the information obtained by the smart devices in the scene after operation to the user s mobile app, completing bidirectional feedback.

Based on the virtual teaching system s actual requirements in this article, establishing a secondary index is used to optimize data retrieval. In order to meet the basic requirements of multi-condition queries, multi-dimensional fields are usually used to combine and assemble Row keys or to find the target data through complete table scanning and filtering. However, this method could be more efficient and meet the system s basic requirements of low latency. Therefore, this article solves this problem by designing a secondary index. The current data table has a Row Key column with RK1 and RK2, including CF: C1 and CF: C2 column families. To establish an index on CF: C1, you only need to establish a mapping relationship between the column value of one of its columns and the row key RowKey. When the user needs to query the value of CF: C2 corresponding to C11, first find the primary key RK1 of the original data table corresponding to C11 through the index table in the diagram, and then query to obtain the value of CF: C2 as C21. The overall implementation process is shown in Figure 3.1. A secondary index utilizes a non-primary key to map the primary key, Row Key, using the non-primary key column names and values as the primary key of the index table. This allows for using the index table s primary key (non-primary key of the original data table) to query and obtain the Row Key when the Row Key of the original data table is not given.

Before using extensive data analysis for prediction and recommendation, it is necessary to migrate the data from the database server (MySQL) to HBase. This article uses the Sqoop tool for data extraction and migration [30, 31]. In addition, this paper designs two data tables in HBase database, which are user information table and knowledge information table. The former stores the user's basic information and assistance behavior, while the latter stores the knowledge point loss information.When conducting data calculations, the two are continuously correlated and queried to obtain basic information about the user s mastery of each knowledge point. Finally, this article adopts the HBase secondary index scheme described earlier, storing the correspondence between primary and non-primary keys in a separate index table to improve query efficiency. The structure of the secondary index table is shown in Table 3.1.

3.2. Implementation of Personalized Teaching Resource Recommendation Module. The system is designed for students of different ages and Chinese proficiency, including beginners, intermediate learners,

Table 3.1: HBase Index Table Structure

Primary key	Knowledge point information	Information value	User Primary Key
(rowkey)	k-rowkey: k-poperty	(value)	(u-rowkey)

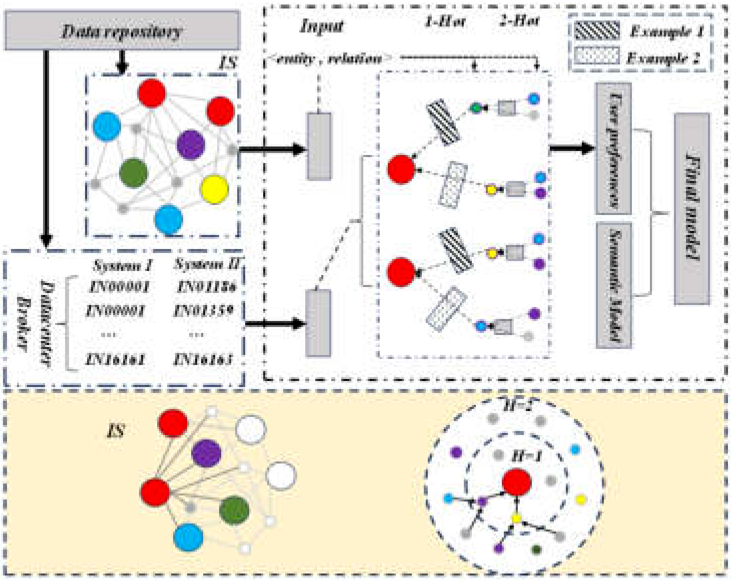


Fig. 3.2: Implementation process of personalized teaching resources

and advanced learners. Some students need to learn Chinese for daily life, while others need to learn Chinese for career development or academic research. The system helps students improve their Chinese skills in listening, speaking, reading, and writing. In addition to the language itself, the system helps students understand the Chinese cultural and social context and promotes cross-cultural communication and understanding.

Implementing a personalized teaching resource recommendation module is based on learners' personal information and many teaching resources provided by the system [32]. Firstly, the system will collect and analyze students' personal information, such as language proficiency, learning goals, learning habits, and learning history. This information can be obtained through learner's registration forms, learning records, and self-assessment [33]. The system will match and filter resources based on learners' personal information and learning needs, combined with the attributes and labels of teaching resources. It uses recommendation algorithms and machine learning techniques to analyze learners' preferences and recommendation history, inferring their potential interests and preferences. Students can obtain teaching resources matching their language proficiency, learning goals, and personal preferences through the personalized teaching resource recommendation module. Such recommendations can help learners learn Chinese more efficiently and improve learning motivation and outcomes. At the same time, the system will continuously optimize personalized recommendation algorithms based on learners learning progress and feedback to provide more accurate, tailored, and diverse teaching resources. The implementation process of personalized teaching resources is shown in Figure 3.2.

MapReduce is an offline batch computing framework in the Hadoop system. The system uses MapReduce to preprocess data and then decompose the matrix of users' lost points in knowledge points with the help of user behaviors.The data preprocessing section mainly cleans and filters the behavior logs of users participating in virtual teaching experiments and assists the data on users score loss in various knowledge points. On the one hand, the distribution matrix of user knowledge point loss is first iterated through the Map stage to generate a user-implicit class matrix. This matrix describes the user s weak knowledge domain; that is, the knowledge

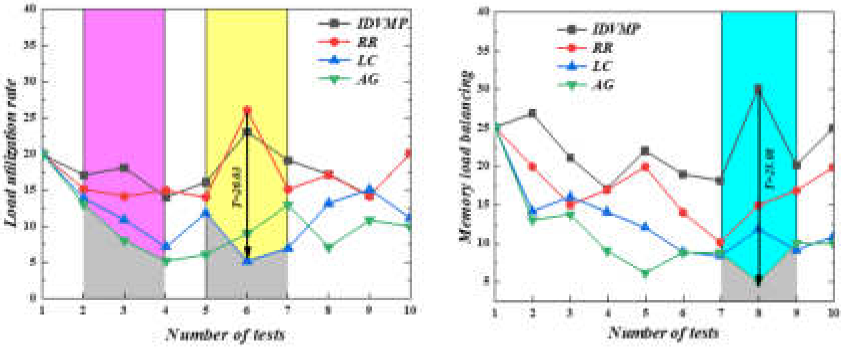


Fig. 4.1: Load balancing tolerance under different algorithms

point loss in that domain is relatively high. Secondly, the data completed by Reduce will be imported into the HBase database and stored as columns in the user information table. Finally, the efficiency of querying relevant information in HBase SQL will be improved by constructing a secondary index table to cope with the subsequent Spark Streaming stream computing user behavior data, updating predictive recommendation models, and completing online recommendations. On the other hand, the user knowledge point loss matrix can also be iteratively generated through the Map stage to generate knowledge point-related implicit class matrices.

4. Analysis of Experimental Results.

4.1. Load Balancing Capability. This section evaluates the load balancing ability, with the primary evaluation criteria including system load utilization. Similarly, the IDVMP algorithm was compared with RR, LC, and AG in experiments, and the experimental results are shown in Figure 4.1. Although the LC algorithm has a slightly lower load balance than the AG algorithm, it still has significant advantages in load utilization compared to IDVMP and RR algorithms, with a maximum load rate of 20.03. LC performed well in terms of load balance, with a maximum difference of 25.08 compared to IDVMP. When the number of tasks is low, there is no difference in execution time among the algorithms. As the number of tasks continues to increase, the execution time of the RR and IDVMP algorithms significantly improves. However, they have significant advantages compared to the LC algorithm, which requires multiple iterations. Therefore, the LC algorithm designed in this article reduces the additional communication overhead generated during the load-balancing process and has better load-balancing capabilities than other algorithms.

4.2. Value of Time Effect Factor α. Considering the impact of time effect on the user s mastery of knowledge points, this paper aims to minimize the solution s root mean square error and obtain the time effect factor µ currently. After conducting several experiments on the time effect influencing factors under different weights, the RMSE under the influence of different time effect factors µ was obtained, as shown in Figure 4.2. The horizontal axis represents the number of implicit classes f under the implicit semantic model, and the vertical axis represents the RMSE under the influence of µ. The experimental results show that when the time effect influence factor $\mu = 0.5$, the matrix decomposition algorithm based on the improved implicit semantic model has the highest accuracy.

4.3. Interactive Analysis of Traditional Teaching and Virtual Teaching Systems. Randomly select 1000 students, use two teaching systems to learn, and obtain the system interaction comparison results shown in Figure 4.3. Through the teaching system designed in Figure 4.4, the number of information submissions is roughly the same as that of information feedback. The degree curve of active and passive interaction fluctuates sharply, indicating that the system interacts frequently and provides real-time active feedback for teaching information to operators using the system. When users ask questions, they promptly provide explanations for the problems. Through comparison, traditional teaching systems have poor interactivity, which affects students' interest in learning. Figure 4.4 shows the distribution of traditional learning methods and virtual

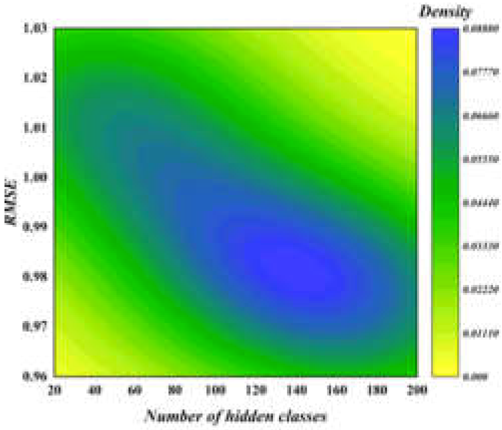


Fig. 4.2: Different Value of time effect factors α

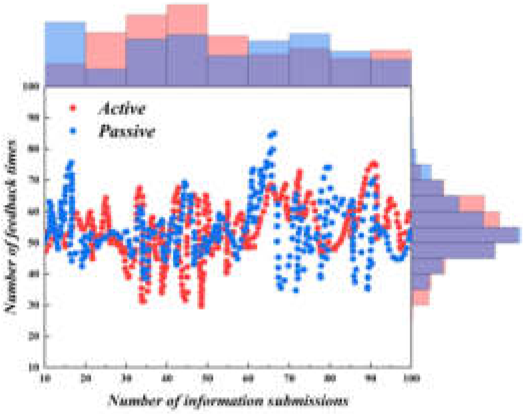


Fig. 4.3: Frequency of Interaction between Active and Passive Learning

reality teaching methods in terms of learning time, homework scores, and test scores. A correlation curve is drawn by collecting the homework and test scores of 1000 students, as shown in Figure 4.4. From the graph, we can see that there is a highly significant positive correlation between learning duration and academic performance (homework and test scores). At the same time, virtual reality teaching technology has a 20% 30% improvement in homework and test scores compared to traditional teaching methods, and students are also more willing to learn. The Spearman correlation analysis results show that traditional teaching methods have R2 values of 0.6953 and 0.0983 for homework and test scores, respectively. Virtual reality teaching methods have R2 values of 0.4944 and 0.3268 for homework and test scores, respectively.

5. Conclusion. This study uses virtual reality technology and has designed and implemented a Chinese language teaching system. Through the development and application practice of this system, the main conclusions are as follows:

1. Virtual reality technology can create realistic virtual environments and provide students with an immersive Chinese learning experience. Students can participate in simulations of various real situations through virtual reality technology, increasing their learning interest compared to traditional teaching methods.
2. Although the LC algorithm has a slightly lower load balance than the AG algorithm, it still has a

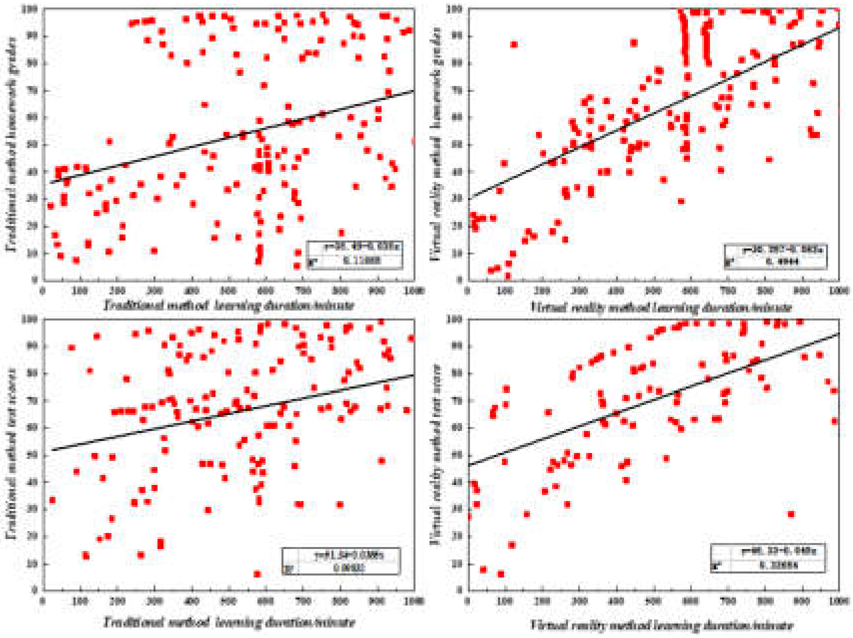


Fig. 4.4: Impact of Traditional and Virtual Reality Teaching on Homework and Test Scores

significant advantage in load utilization compared to IDVMP and RR algorithms, with a maximum load rate of 20.03. Regarding load balance, LC performs well overall, with a maximum difference of 25.08 compared to IDVMP. Therefore, adopting the LC algorithm reduces the additional communication overhead generated during the load-balancing process and has better load-balancing capabilities than other algorithms. For the implicit semantic model, when the time effect influence factor $\mu = 0.5$, the matrix decomposition algorithm based on the improved implicit semantic model has the highest accuracy.

3. Aiming at the low degree of individuation in the teaching system, this paper optimizes the user application layer, recommends personalized teaching methods, builds the knowledge point association matrix integrating multiple factors according to the knowledge point association and the interactive relationship between virtual users, and introduces the rule of time forgetting to decompose the user knowledge point association matrix. According to the decomposition results, personalized teaching guidance for users is realized.

REFERENCES

[1] Learning, C. Exploring Effective Language Learning Strategies. *Taylor And Francis:*. **23** pp. 2023-08 (0)

[2] Angelina, M. Systematic Review of Games for Learning Chinese. *International Journal Of Education (IJE).* **10** pp. 3 (2022)

[3] Xu Y. Online Chinese Learning:, A. Case Study of the Use of YouTube Instructional Videos. *Chinese Language Teaching Methodology And Technology.* **4** pp. 2 (2021)

[4] Strategies, Z. Motivation and Learners Perspectives on Online Multimodal Chinese Learning. *Chinese Language Teaching Methodology And Technology.* **4** pp. 1 (2021)

[5] Wang, P., Wu, P., Wang, J. & Others A critical review of the use of virtual reality in construction engineering education and training. *International Journal Of Environmental Research And Public Health.* **15** pp. 6 (2018)

[6] Elizabeth, C. & W., P. The Power of Presence in Virtual Teaching and Practice Environments. *Nursing Clinics Of North America.* **57** pp. 4 (2022)

[7] Zhang, T., Mc Carthy, Z., Jow, O. & Others Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD. *Australia.* pp. 5628-5635 (2018)

[8] Lee, G., Sharon, F., R, L. & Others Impact of immersive virtual reality simulations for changing knowledge, attitudes, and behaviors. *Nurse Education Today.* **2021** pp. 105 (0)

[9] Yufeng, L., Tongsheng, L. & Qiaoyun, M. Immersive Virtual Reality Teaching in Colleges and Universities Based on Vision Sensors. *Wireless Communications And Mobile Computing.* **2022** pp. 2022 (0)

[10] And, E. and virtual reality in education. myth or reality?. *International Journal Of Emerging Technologies In Learning (IJET).* **14** pp. 03 (2019)

[11] Webb, M., Tracey, M., Harwin, W. & Others Haptic-enabled collaborative learning in virtual reality for schools. *Education And Information Technologies.* **2021** (0)

[12] Mildred, L. Carrillo G J A, Pablo J Á N, et al. *Virtual Reality Vs Traditional Education: Is There Any Advantage In Human Neuroanatomy Teaching? Computers And Electrical Engineering.* **2021** pp. 93 (0)

[13] Zijie, P., Hairong, S. & Li, L. Application of Virtual Reality Technology in the Maintenance of Steam Turbine. *Journal Of Physics: Conference Series.* **1966**, 2021 (0)

[14] S, Y. & P., H. Study of the virtual reality education and digitalization in China. *Journal Of Physics: Conference Series.* **1456** (2020)

[15] Megan, G. & Dimitrios, P. Virtual Didactics Maintain Educational Engagement with Convenience. *Western Journal Of Emergency Medicine: Integrating Emergency Care With Population Health.* **2022** pp. 23 (0)

[16] Farzaneh, S. Integrating diversity in simulation-based education: Potential role of virtual reality. *AEM Education And Training.* **5** pp. 3 (2021)

[17] Vijay, V., Deborah, C., Christian, B. & Others Virtual Reality in Chemical and Biochemical Engineering Education and Training. (Education for Chemical Engineers,2021)

[18] Jeong, J., Oh, W. & And, Y. and implementation of virtual reality educational contents system for smart learning. *Journal Of The Korean Society For Computer Game.* **28** pp. 2 (2015)

[19] Rohit, B., Ram, S., Amandeep, S. & Others Redefining Virtual Teaching Learning Pedagogy. (2023,0)

[20] Mingjie, W., Hengxu, Z. & Tianyu, F. Enhancing the course teaching of power system analysis with virtual simulation platform. *International Journal Of Electrical Engineering & Education.* **60** pp. 3 (2023)

[21] Diane, A., Alex, Z., Chuck, I. & Others Virtual teaching assistants: A survey of a novel teaching technology. *International Journal Of Chinese Education.* **11** pp. 2 (2022)

[22] Lucas, P., Vaca, D., Dominguez, F. & Others Virtual circuits: an augmented reality circuit simulator for engineering students. 2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT). (IEEE,2018)

[23] Jhamb, Y. & Fang, Y. dual-perspective latent factor model for group-aware social event recommendation[J]. *Information Processing & Management.* **53**, 559-576 (2017)

[24] Mongia, A., Jhamb, N., Chouzenoux, E. & Others Deep latent factor model for collaborative filtering. *Signal Processing.* **169**, 6 (2020)

[25] Li, H., Diao, X., Cao, J. & Others Collaborative filtering recommendation based on all-weighted matrix factorization and fast optimization. *Ieee Access.* **6** pp. 25248-25260 (2018)

[26] Megan, G. & Dimitrios, P. Virtual Didactics Maintain Educational Engagement with Convenience. *Western Journal Of Emergency Medicine: Integrating Emergency Care With Population Health.* **23**, 1 (2022)

[27] K, L. & P., M. High-Impact Practices for Virtual Instruction. *Kappa Delta Pi Record.* **58** pp. 4 (2022)

[28] Wohlgenannt, I., Simons, A. & Reality, S. Business & Information Systems Engineering. (2020)

[29] Si, Z., Shan, G. & Haiying, L. Cloud Application in the Construction of English Virtual Teaching Resources Based on Digital Three-Dimensional Technology. (Wireless Communications,2022)

[30] Han, G., Que, W., Jia, G. & Others An efficient virtual machine consolidation scheme for multimedia cloud computing. *Sensors.* **16** pp. 2 (2016)

[31] Liu, L., Zheng, S., Yu, H. & Others Correlation-based virtual machine migration in dynamic cloud environments. *Photonic Network Communications.* **31**, 206-216 (2016)

[32] Samar, M., Samer, R., Roznim, M. & Others Virtual Teaching Assistant for Capturing Facial and Pose Landmarks of the Students in the Classroom Using Deep Learning. *International Journal Of E-Collaboration (IJeC).* **19** pp. 1 (2023)

[33] Yu, Y. Exploration on Teaching Reform of Engineering Mechanics for Application-oriented Undergraduate Based on Information Teaching. *Frontiers In Educational Research.* **5** pp. 21 (2022)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: Oct 27, 2023
Accepted: Dec 29, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1578–1587, DOI 10.12694/scpe.v25i3.2768

A PARAMETER ASSESSMENT OF TEACHING QUALITY INDICATORS BASED ON DATA CLASS MINING FUZZY K-MEAN TYPE CLUSTERING

XINHUA HUANG[*]AND YUZUDI TONG[†]

Abstract. This paper proposes a data-based mining and hesitant fuzzy C-canopy-K mean clustering degree algorithm and uses it in the parameter assessment model of teaching quality indicators. Simulation and training are carried out through data class mining, and information input, followed by combining the hesitant fuzzy K-mean classification assessment method, which involves a hesitant fuzzy type evaluation system, a neural network identification and prediction system, and an application system for module identity verification. The simulation results show that the results of the six simulation conditions are consistent with the actual results, with only slight differences in some amplitudes, and a high degree of consistency in the overall trend, the change rule, and the average peak value. Through the prediction model processing in this paper, the teaching quality index parameter assessment has high accuracy and can reach more than 95.0%, in addition, the development of the law also fits very well. a, b, c, d four kinds of teaching quality parameter assessment of the average calculation of the assessment speed increased by 52.5%. In addition, the assessment test after the integrated design of the module shows that the system can effectively identify the four clustering identification processes that can be seen as excellent, good, medium, and poor; at the same time, the test data show that the system class effectively for teaching quality indicator parameter assessment.

Key words: fuzzy K-clustering; C-anopy-K mean; data class mining; teaching quality; indicator parameter assessment

1. Introduction. With the rapid development of Internet technology, people's lives have become closely related to the Internet. In the context of the rapid popularization of the Internet, the behavioral analysis of Internet users has now become an advantageous means of gaining insight into users' teaching quality and other preferences, learning ability, etc. [1]. The analysis of user Internet behavior provides more diversified choices for intelligent network module authentication, but at the same time, it also puts forward more stringent technical requirements and specifications for intelligent network module authentication. The Internet behavior of network users is monitored by the data platform, while the platform understands the user's intention through data analysis, thus promoting the benign development of the network ecosystem [2, 3]. Currently, servers for certified billing, traffic line monitoring, and other applications are already widely used in the teaching quality management of major universities. These application servers provide management convenience for colleges and universities at the same time but also generate a large amount of log data, which is usually stored in the background database. Analysis shows that the log data contains a large number of user behavioral data on the Internet [4, 5]. Suppose the behavioral data in the logs can be scientifically and efficiently analyzed, and the deep-seated laws hidden in the data can be utilized. In that case, it will greatly improve the speed of network management assessment in universities, build effective support for network management in universities, and provide useful help for the scientific decision-making and management refinement. This paper takes a specific university as an example [6], analyzes the clustering of user online behavior data, mines the intrinsic laws, and helps the smooth implementation of university decision-making [8]. In traditional comprehensive indicators, raw data are usually given in the form of point values (real numbers). However, with the development of society, the evaluation environment is becoming more and more complex, and the evaluator is often affected by some of his own subjective and objective factors, such as knowledge structure [7], judgment level, and personal preference [9-10], and the evaluations made are largely imprecise or fuzzy [9]. In the evaluation of the level of impact of the resumption of production by enterprises on the recovery of the local economy after the epidemic, due to the existence of many uncertain factors, the learning parameters have a hesitant mentality when scoring

[*]School of Materials Science and Engineering, Anhui University of Science and Technology, Huainan, 232001, China (Corresponding author, Xinhua_Huang@outlook.com)

[†]School of Materials Science and Engineering, Anhui University of Science and Technology, Huainan, 232001, China

[10], or multiple learning parameters have multiple different judgment results for them. Zadeh [13] proposed the concept of fuzzy sets in 1965, allowing the degree of affiliation of an element belonging to a set to be taken arbitrarily in the [0, 1]. The theoretical foundation of fuzzy sets was laid down by Torra et al [15, 14], who proposed the definition of hesitant fuzzy sets, where multiple values can be selected as the final degree of affiliation when the decision maker is hesitant. Xu Zeshui's team [16, 17] proposed a distance measure and a similarity degree measure formula between hesitation class fuzzy type sets and gave proof. Xia Meimei proposed a series of hesitant fuzzy distance and similarity degree formulas. [18] Clustering is the process of dividing a series of pairs of images, scenarios, events, etc., into several classes, where the characteristics of the objects in each class have a higher degree of similarity than the other classes. The analysis of clustering is the use of mathematical methods to classify objective things according to defined criteria, the degree of similarity of the samples as the principle of division, so that the selection of the appropriate degree of similarity becomes the key to clustering. Facing different fuzzy environments, various clustering degree algorithms have been proposed to deal with different types of fuzzy data, such as the intuitionistic fuzzy clustering degree algorithm [19], the two-type fuzzy clustering degree algorithm [20], etc. In 2015, Chen Na [21] proposed an algorithm to cluster hesitant fuzzy-type information based on the fuzzy information integration operator and the measure of the distance. In 2008, many scholars and others [22, 23, 24, 25, 26] applied the K-mean algorithm to fuzzy clustering. In the traditional K-mean clustering degree algorithm, due to the initial clustering center being random, sometimes needs to iterate several times to get the final clustering results, to a certain extent, affecting the speed of clustering evaluation. The c-anopy algorithm belongs to a kind of ”coarse” clustering degree algorithm, through a simple, fast distance calculation can be hesitant fuzzy type set into several overlapable clusters. The fuzzy type set can be divided into several overlapping subsets by simple and fast distance calculation [29]. Moreover, compared with the traditional K-mean clustering degree algorithm, it does not need to formulate the number of clusters in advance. Therefore, to simplify the number of iterative approximations in the clustering degree algorithm, this paper proposes a K-mean hesitant fuzzy clustering degree algorithm based on the C-anopy algorithm [27]. Therefore, this paper develops a model of a new algorithm of prediction and assessment based on data class mining hesitant fuzzy K-mean type clustering for parameter assessment of teaching quality indicators, which can greatly improve the speed and accuracy of assessment based on the original model [28, 29, 30, 31].

2. Hesitant Fuzzy Assessment Model and Algorithm. The overall model architecture of this paper is shown in Figure 2.1, which involves a hesitant fuzzy type evaluation system, a neural network identification and prediction system, and an application system for modular identity verification. Through data class mining, information input, after that, combined with hesitant fuzzy K-mean type clustering method for simulation and training. A new algorithmic system for prediction and evaluation is developed by obtaining several predictions. Finally, the assessment of teaching quality and practical application is accomplished through the final component of the clustering degree algorithm. The specific kernel composition and calculation method are described in detail in the following.

2.1. Distance formula for hesitant class fuzzy type sets. In this given set, the occupation ratio, where is the occupation of the element in the set X and satisfies, let M be a hesitant class fuzzy type set defined on the set X. The measured (M, N) of the distance between M and N satisfies the following property.

1. 0 d (M, N) 1.
2. d (M, N) = 0 holds if and only if M = N;
3. d (M, N) = d (N, M)

Under the above-given conditions, the hesitant fuzzy weighted Euclidean degree distance formula is defined as:

$$d_{\mathrm{hw}}(M,N)=\left[\sum_{i=1}^{n} w_i\left(\frac{1}{l_{x_i}}\sum_{j=1}^{l_{x_i}}\left|h_M^{\sigma(j)}(x_i)-h_N^{\sigma(j)}(x_i)\right|^2\right)\right]^{\frac{1}{2}} \tag{2.1}$$

where $h_M^{\sigma}(j)$ and $h_N^{\sigma}(j)$ are the first largest element in the hesitant fuzzy degree number, respectively. To facilitate the calculation, the length of the paste number needs to be the same in each model, so it is necessary to add elements to the set with a short length of the hesitant ambiguity degree. In this paper, it is stipulated

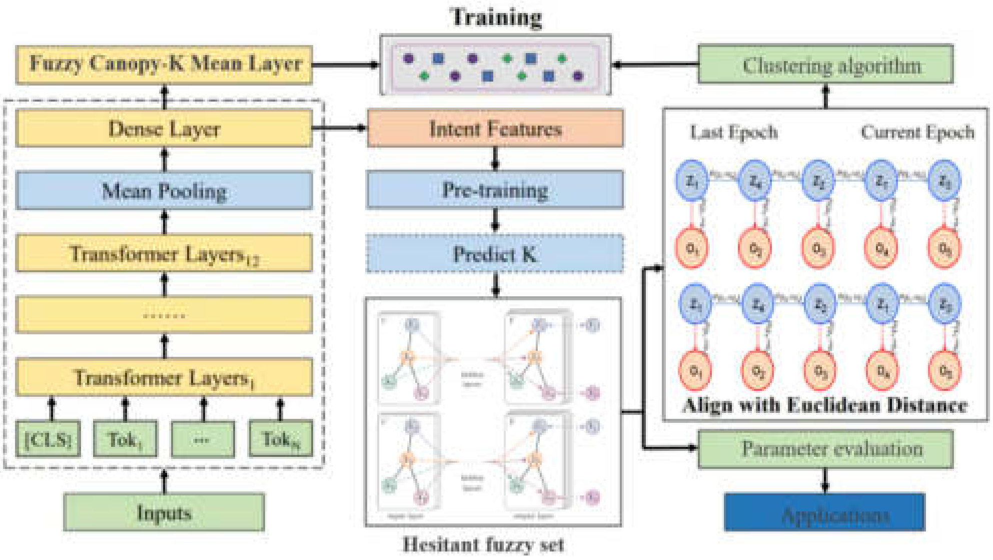


Fig. 2.1: Overall flowchart of the model

that the element with the smallest value in the set is added for the set that needs to be added. Then one of the hesitant fuzzy type sets can be written:

$$M_j = \{(x_i, h_{A_j}(x_i)) \mid x_i \in X\} \quad \text{for } j = 1, 2, \ldots, k \tag{2.2}$$

2.2. Recursive approximation for the assessment of teaching quality parameters. A quantitative recursive analysis method was used to analyze the big data information model for the assessment of comprehensive teaching competence.4 The control objective type function for constructing the predictive estimation of comprehensive teaching competence was:

$$\max_{x_{a,b,d,p}} \sum_{a\in A} \sum_{Ab\in B} \sum_{d\in D} \sum_{p\in P} x_{a,b,d,p} V \tag{2.3}$$

$$\sum_{a} \sum_{b} \sum_{d} x_{a,b,d,p} R_p^{bw} \leq K_b^{bw}(S), \quad b \in B \tag{2.4}$$

Quantitative recursive assessment of the level of comprehensive teaching ability using the gray degree model, assuming that the historical data of the distribution of comprehensive teaching ability is expressed as (x), and the probability density generalization of the predictive estimation of comprehensive teaching ability is obtained as with a certain initial value of the perturbation feature:

$$u_c(t) = K x_c(t) \tag{2.5}$$

In the high-dimensional characteristic type distribution space, the integrated teaching ability prediction estimation statistical model of the continuous function of u: IR→IR, after k-1 iterative approximation, k>1, integrated teaching ability assessment of the grayscale degree sequence to satisfy the N (k) < L, using quantitative type recursive analysis method, to get to the integrated teaching ability assessment of the output of the index distribution of the situation of big data information number of the K nearest-neighbor residue value is:

$$P_{1J} = \sum_{d_i \in \mathrm{KNN}} \mathrm{Sim}(x, d_i) \cdot y(d_i, C_j) \tag{2.6}$$

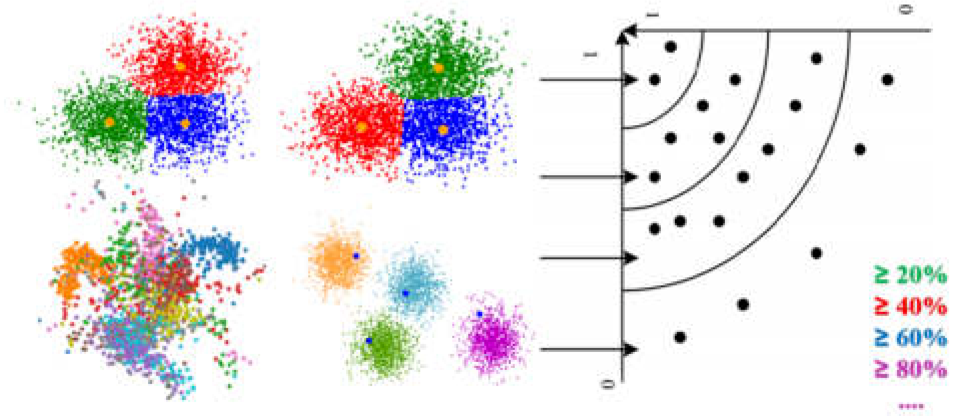


Fig. 3.1: Hesitant fuzzy algorithm clustering process schematic

The sequence of exponential correlation distributions (x) of the comprehensive teaching competence assessment of the large clustering degree data study was quantitatively analyzed and combined with the K-value optimization search method to obtain the quantitative recursive feature extraction results of the teaching competence assessment as:

$$x_n = a_0 + \sum_{i=1}^{M_{\mathrm{AR}}} a_i x_{n-i} + \sum_{j=0}^{M_{\mathrm{MA}}} b_j \eta_{n-j} \tag{2.7}$$

where α_0 is the amplitude of the sampling of the initial comprehensive teaching competency teaching assessment; x_n is the time series of the scalar; b_j is the oscillatory downward decay value of the comprehensive teaching competency assessment.

3. Parameter evaluation simulation program design. The K-mean hesitant fuzzy clustering degree algorithm based on the Canopy algorithm proposed in this paper uses Canopy clustering (Program 1) as a cluster to obtain the K-mean initial class centers, which are subsequently used to obtain the K-mean clusters by K-mean type clustering (Program 2) The final clustering results are obtained. The specific steps are as follows:

Step 1 Suppose k hesitant fuzzy type sets M, M2, ..., Mn

Step 2 Take an arbitrary class M, from which the distance D between the class M, and the remaining k-1 classes is calculated by Equation (1). set the values of the 2-distance kurtosis values T, T2 based on a priori knowledge, where T, > T2.
Here, T is chosen as the mean value after removing the minimum and maximum distances, and if T2<D<T, a weak mark is given to R to indicate that R belongs to that C-anopy, and R is added to it; if D<T2, a strong mark is given to R to indicate that R belongs to that C-anopy and is very close to the center of mass, and R is deleted from the set, and will not be a centroid in the future; if D>T, then R forms a new set of clusters and R is removed from the set.

Step 3 Repeat Step 2 until the elements within each set no longer change, at which point c Canopy (1<c<k) will be formed, each containing one or more hesitant class fuzzy type sets M. In this paper, we will denote each C-anopy as a hesitant class fuzzy type set M, (j=1, 2, c).

Step 4 From Eq. (3) combine M, (j = 1, 2, c) in the hesitant fuzzy type sets M are merged, and the class center of each M is calculated. The process of clustering is shown in Figure 3.1.

Step 5 Obtain the number of categories of clusters c and the initial cluster centers from Procedure 1.

Table 3.1: Indecisive fuzzy assessment information

mould	M1	M2	M3	M4	M5
X1	{0.1,0.4}	{0.4,0.7}	{0.45,0.5,0.6}	{0.3,0.5}	{0.8,0.9,1}
X2	{0.1,0.3}	{0.5,0.6,0.8}	{05}	{1}	{0.9}
X3	{0.4,0.5,0.6}	{0.5,0.6}	{0.1, 0.15, 0.2}	{0.5,0.7}	{0.7.0.8, 0.85}
X4	{0.4}	{0.15, 0.2, 0.35}	{0,0.1,0.2}	{0.4, 0.5, 0.65}	{0.6,0.8}
X5	{0.3,0.4,0.5}	{0.1,0.2,0.3}	{0.2,0.4}	{0.35}	{0.4, 0.5, 0.75}
X6	{0.2}	{0.6,0.7}	{0.5,0.6,0.8}	{0.4}	{0.3,0.35}

Step 6 Calculate the distance between the hesitant class fuzzy type set M and the class center by using Equation (1), and merge M into the class closest to the class center.

Step 7 Calculate the new class center from equation (3).

Step 8 Repeat steps 6 and 7 until the hesitant blur Can-copy-K iterative approximation reaches clarity and stability.

Step 9 Setting up under different working condition characteristics through the joint model, and evaluation application of subsequent equations (4-6), etc.

Considering that different learning parameters may give different assessment values for the attributes of the program, the hesitant fuzzy type set is used to represent the assessment information for the development status of the five teaching qualities. The specific data calculated by preliminary simulation are shown in Table 3.1.

4. Practical test analysis. To better illustrate the effectiveness and stability of the hesitant fuzzy C-anopy-K-mean type clustering degree algorithm proposed in this paper, the specific clustering process of the newly proposed algorithm is first given in combination with example data. Then, it is compared and analyzed with the K-mean type clustering degree algorithm based on hierarchical analysis. Matlab-2020 co-simulation analysis method was used to test the analysis performance under big data of comprehensive teaching ability assessment, statistical type analysis method was used to sample the data of comprehensive teaching ability assessment, the kurtosis value of decision making of teaching ability assessment was taken as D, $= 2$, the correlation parameter of the distribution of comprehensive teaching resources was set as $= 3/5$, $= 2/5$, $= 2/5$, $maxg1(d) = 6/5, maxg2(d) = 3/8$, $maxg3(d) = 1/10$, sampling frequency f=600 Hz, adaptive initial step size p=0.97, and the coefficient of correlation of the distribution of teaching resource characteristics is B=1.14. According to the above parameter settings, the big data reconstruction of the constraint parameters of the comprehensive teaching ability assessment is carried out, and the six time-domain waveforms of the big data distributions of the actual test are obtained as shown in Figure 4.1 shows.

After the fuzzification process through the above model, the waveforms of the six time-domain waveforms directly predicted by the parameters in the simulation model are shown in Fig. 4. It can be found that the results of the six time-domain waveforms predicted by the model in this paper are very similar to those of the original big data, and the peaks are comparable. Among them, the degree of conformity of condition 1 - condition 6 is consistent, only in some amplitude slightly different, which is due to the identification of the sampling frequency interval decided. Still, the overall trend, change rule, and average peak value have a high degree of consistency.

In the paper, four clusters A, B, C, and D are used in the application of K-Mean-s, and each of the four clusters accounts for 10%, 20%, 30%, and 40% of the total sample capacity. The overall parameters in the above model are made as the object of study, data clustering and information fusion processing are carried out and realized to achieve the assessment of teaching ability, and the results are shown in Figure 5. It can be found that through the above model parameters, the classification identification carried out and the prediction results are in good agreement. It shows that after processing through the prediction model in this paper, the assessment of teaching quality index parameters has high accuracy, which can reach more than 95.0%, in addition to the development law is also very suitable.

In addition, our model was compared with the traditional model in terms of computational assessment speed for the assessment of four teaching quality parameters, A, B, C, and D. The results are shown in

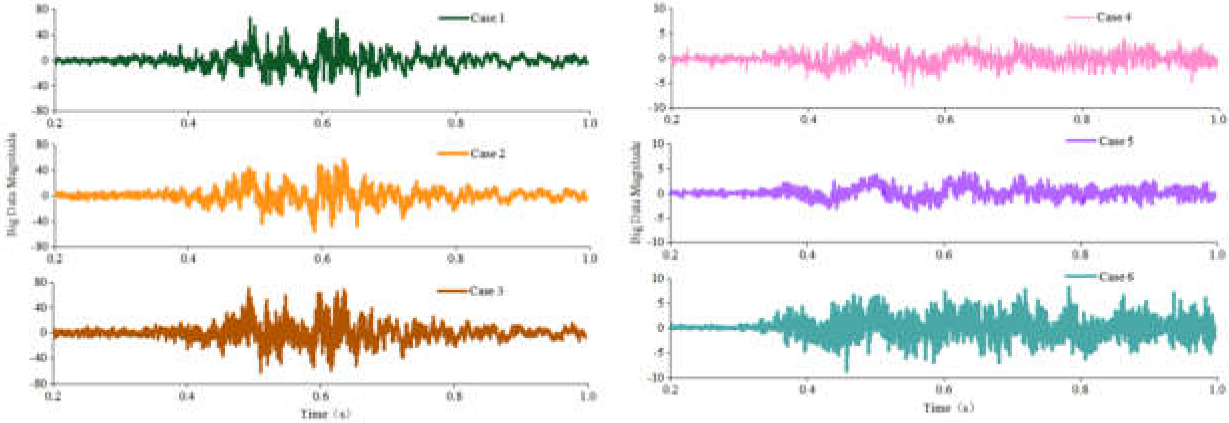


Fig. 4.1: Schematic representation of the six waveforms of big data on teaching quality

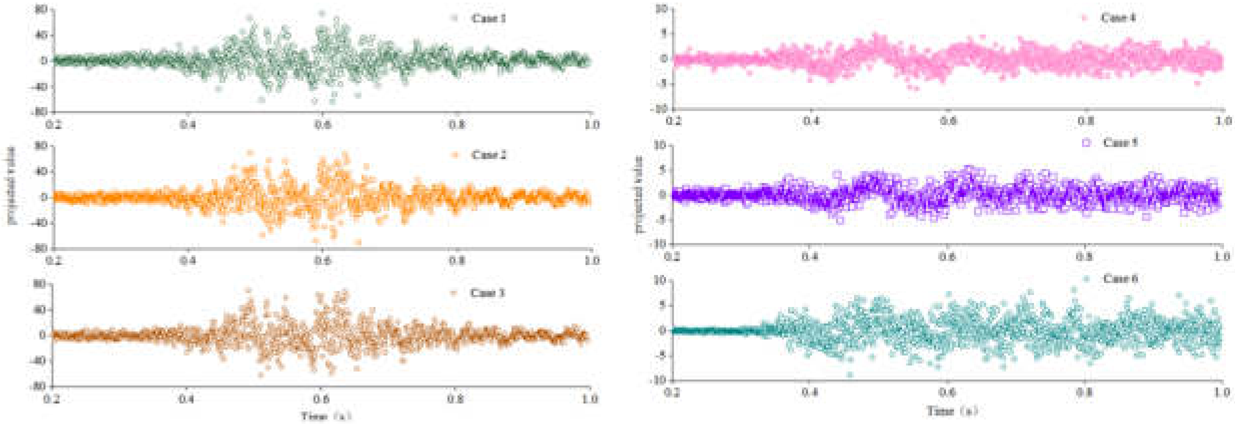


Fig. 4.2: Schematic of the six waveforms of the parameters predicted by the simulation of teaching quality

Figure 4.4. It can be found that all the methods proposed in this paper have high computational assessment speed. Among them, under the working condition A of teaching quality index parameter evaluation, the method of this paper improves by 59.0% compared with the traditional method, which is the highest among the four working conditions. In Case B, this paper's method is 56.2% faster than the traditional method, and in Case C, this paper's method is 49.2% faster than the traditional method. In condition D of the evaluation of teaching quality parameters, the method of this paper has improved by 45.6% compared to the traditional method. Summarizing the centralized conditions, the average speed of calculation and evaluation increased by 52.5%.

5. Integrated module design and evaluation of indicator parameters. Combining the fuzzy evaluation and data class mining methods mentioned above, this paper designs the module identity verification system shown in Fig. 5.1. In this system, there are six sub-functional modules, which are module identity verification module, evaluation basic information module, evaluation program design module, user online evaluation module, evaluation result statistical analysis module and system setting management module. The specific functions of each module are as follows:

1. Module authentication. The system designed in this paper is oriented to four categories of users:

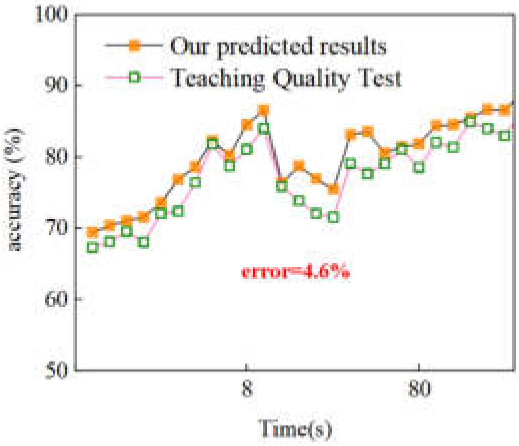


Fig. 4.3: Comparison of Simulation Results with Evaluated Parameters

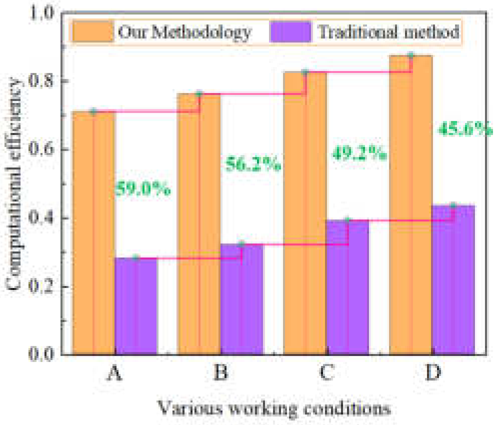


Fig. 4.4: Comparison of computational evaluation speeds for four working conditions A/B/C/D

students, teachers, experts, and system administrators, different categories of users have different user rights; different categories of users have different functional requirements in the system. Therefore, the system needs to distinguish user rights according to different user needs to ensure the smooth progress of the teaching evaluation workflow.

2. Evaluation of basic information settings. In the actual work of teaching evaluation, different users need to combine the specific needs of their information in the system to maintain it, and to determine the validity of the results of each evaluation. For student users, they need to browse the relevant evaluation information; for teachers and expert users, they need to browse and query the relevant evaluation information; for system administrators, they need to maintain all personnel information promptly.
3. Design of the teaching evaluation program. The users of this module are mainly experts who promote the work of teaching evaluation, and these experts, combined with the preliminary research, will add, delete, and distinguish the relevant evaluation indicators in the system's evaluation information setting module to occupy the final generation of the evaluation program.
4. Online evaluation of teachers by users. In this module, combining the different needs of different users,

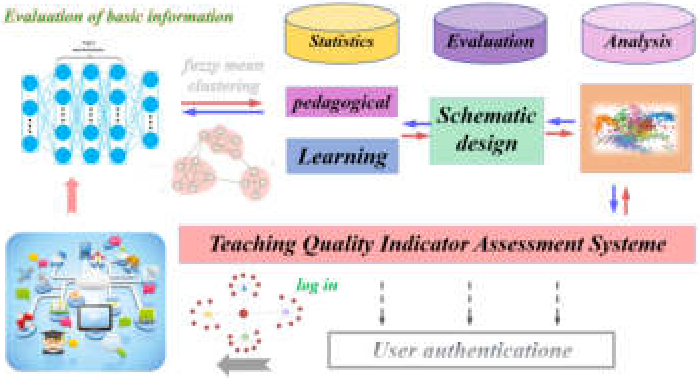


Fig. 5.1: Design and evaluation system for the modular system as a whole

Table 5.1: Indecisive fuzzy assessment information

Mould	excellent	very much	center	differ from
quality assessment	26.0%	13.0%	29.0%	32.0%

online evaluation of teachers' teaching content, teaching level, political level, scientific research level, and other dimensions is carried out, and this module can realize the evaluation of students on teachers, mutual evaluation among teachers, and the inspection and evaluation of teachers by experts.

5. Statistical analysis of evaluation results. This module combines the fuzzy evaluation and data class mining algorithms in the above section to carry out automated evaluation result statistics. Users can view and query the evaluation results with the help of this module, and at the same time, the historical evaluation results of each teacher will be saved in the database for all kinds of assessments.
6. System settings management. The main user of this module is the administrator of the system, which can be used for the allocation of user rights and responsibilities and the viewing of system operation logs.

After the final evaluation system of teaching quality index parameters is realized, according to the situation of the collected valid evaluation data, the actual evaluation test is carried out, and the results are shown in Figure 5.3, which shows that the four clusters of excellent, good, medium and poor are identified. By occupying the allocation, and then combining the K-M algorithm to obtain the parameter evaluation ratio of each teacher's teaching quality indicators, the evaluation results of all teachers can be obtained. Comprehensive pilot test calculations are shown in Table 5.1, which shows that 26.0% of the teachers in the school received excellent, 13.0% good, 29.0% moderate, and 32.0% poor in a teaching evaluation.

6. Conclusion. In this paper, a data class mining and hesitant fuzzy C-canopy-K mean clustering degree algorithm based on data class mining and information input, followed by simulation and training test combined with hesitant fuzzy K-mean classification evaluation method is proposed. Specific results are shown. The results of the six simulation conditions are consistent with the actual results, only slightly different in some amplitude, and the overall trend, change rule, and average peak value are highly consistent. After processing through the predictive model of this paper, the parameters of teaching quality indicators are assessed with high accuracy, which can reach more than 95.0%, in addition to the law of development fits well. The average computational assessment speed of the four teaching quality parameters assessment of A, B, C, and D increased by 52.5%. In addition, the assessment test after the integrated design of the module showed that the system can effectively identify the four clustering identification processes that can be seen as excellent, good, moderate, and poor; at the same time, the test data showed that the system class effectively assesses teaching quality indicator

Excellent,
good
poor
fair
Excellent,
good
poor
fair

Fig. 5.3: Schematic representation of the results of the data clustering assessment

parameters.

REFERENCES

[1] Li, H., Zhaoyun, D., Yan, J. & Others Candidate category search in large-scale hierarchical classification. *Journal Of Computing.* **37**, 41-49 (2014)

[2] Runmu, Z., Zhiyong, L., Shaomiao, C. & Others Parallel optimized sampling clustering K-Mean-s algorithm for big data processing. *Computer Applications.* **36**, 311-315 (2016)

[3] Deng, Z., Cao, L., Jiang, Y. & Others Minimax probability TSK fuzzy system classifier: a more transparent and highly interpretable classification model. *IEEE Transactions On Fuzzy Systems.* **23**, 813-826 (2015)

[4] Zarinbal, M., Zarandi, M. & Turksen, I. Relative entropy fuzzy c-means clustering. *Nformation Sciences.* **260**, 74-97 (2014)

[5] Li, B., Jinsong, W. & Wei, H. A new clustering degree algorithm in the big data environment. *Computer Science.* **42**, 247-250 (2015)

[6] Hu, J., Hu, X. & JX., C. Spark-based hybrid computing model for big data. *Computer System Applications.* **24**, 214-218 (2015)

[7] Torra, V. & Narukawa, Y. On hesitant fuzzy sets and decision[C]//Proceedings of the 18th IEEE International Conference on Fuzzy Systems. *EEE.* pp. 1378-1382 (2009)

[8] Torra, V. Hesitant fuzzy sets. *Nternational Journal Of Intelligent Systems.* **25**, 529-539 (2010)

[9] Xia, M. & Hesitant, X. fuzzy information aggregation in decision-making. *Nternational Journal Of Approximate Reasoning.* **52**, 395-407 (2011)

[10] Zhu, B. & S., X. Some results for dual hesitant fuzzy sets. *Ournal Of Intelligent And Fuzzy Systems.* **26**, 1657-1668 (2014)

[11] And, X. and similarity measures for hesitant fuzzy sets[1]. *Nformation Sciences.* **181**, 2218-2138 (2011)

[12] Wang, Z., Zs, X. & Liu SS., A. netting clustering analysis method under an intuitionistic fuzzy environment. *Pplied Soft Computing.* **11**, 5558-5564 (2011)

[13] Hwang, C. & Clustering, R. interval type-2 fuzzy approach to C-means. *EEE Transactions On Fuzzy Systems.* **15**, 107-120 (2007)

[14] Na., C. Research on decision-making method and clustering degree algorithm in hesitant fuzzy environment. *Nanjing: Southeast University.* pp. 107-118 (2015)

[15] Chunxu, W., Dy, W. & Ning, J. A fuzzy clustering degree algorithm based on information entropy and K-mean iterative approximation model. *China Management Science.* **2008** pp. 152-156 (0)

[16] Song, J. Improvement research on K-Mean-s clustering degree algorithm. (Anhui University,2016)

[17] Meimei, X. Research on fuzzy decision-making information integration approach and measurement. (Southeast University,2012)

[18] Chengde, Y. Research on Evaluation Indicators of County Township Science and Technology Enterprises in the Context of High-Quality Development–Taking Taicang Township Science and Technology Enterprises as an Example. *Shanxi Agricultural Economics.* **2020**, 103-104 (0)

[19] Xiongsheng, Y. & Zhendai, Y. Research on Comprehensive Evaluation Indicator System of Enterprises. *Financial Research.* **1998**, 40-47 (0)

[20] Lujf, T., Tang, Z. & Others Hierarchical initialization approach for K-Mean-s clustering. *Attern Recognition Letters.* **29**, 787-795 (2008)

[21] Weixi, Q. & Bin, L. A study on the progress of multi-objective evolutionary algorithms. *Computer And Digital Engineering.* **36** pp. 16-18 (2008)

[22] Qing, D., Liugen, Z., Aibing, Z. & Others Research on user behavior analysis of campus network based on K-Mean-s clustering degree algorithm. *Microcomputer Applications.* **31** pp. 06 (2010)

[23] Liu, J. Comparison of Kohonen neural network based clustering methods in remote sensing classification. *Computer Simulation.* **26** pp. 1744-1746 (2006)

[24] Center, D. . (Web User Behavior Log Collection [R/OL],2020), http://www.datatang.com/data/43910

[25] Sun, J. & Graphics, Y. Beijing: Tsinghua University Press. (0)

[26] Li, D. and application of data class mining in user behavior analysis. (Beijing University of Posts,2009)

[27] Zongcheng, L., Zhonglin, Z. & Miaofeng, T. Network behavior analysis based on association rules. *Electronic Science And Technology.* **28** pp. 16-18 (2015)

[28] Wang, J. Design and realization of a teaching evaluation system based on fuzzy evaluation model. *Computer And Digital Engineering.* **44**, 1737-1742 (2016)

[29] Li, J. & JJ., C. Research on network-based collaborative teaching mode and its effect. *Electronic Science And Technology.* **26**, 150-153 (2013)

[30] Li, Y. Design and realization of teaching evaluation system for college teachers based on a fuzzy comprehensive evaluation algorithm. (Hubei University of Technology,2018)

[31] Chengcheng, L., Yuan, Y. & Pengrui, S. Teaching and training quality assessment based on relative gray correlation ideal solution. *Computer And Digital Engineering.* **43** pp. 10 (2015)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: Oct 31, 2023
Accepted: Dec 12, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1588–1602, DOI 10.12694/scpe.v25i3.2778

PREDICTION OF ENGLISH TEACHER CAREER DEVELOPMENT BASED ON DATA MINING AND TIME SERIES MODEL

LIPING FAN*

Abstract. With the gradual growth of the teaching profession, the teaching profession is facing new trends in reform and development, and the same dilemma exists for English teachers' career development and planning. To this end, the study first uses a modified K-means clustering method to cluster and analyse the factors affecting English teachers' professional development, forming a system of indicators on English teachers' professional development. The Long Short-Term Memory (LSTM) network employed the time-series features to create a time-series model, and the Support Vector Machine (SVM) was used to forecast the course of English teachers' career development. To assess the current career status of English teachers and their impact on people and organizations, this study proposes a career prediction model for English teachers. This model utilizes data mining and time series modeling to provide accurate predictions. In accordance with the experimental findings, the precision of the upgraded K-Means model was 98.58%, and the error between the projected sample data and the actual sample data for the training of the trend prediction model for English teachers' career growth was 0.032. It was able to accurately predict teachers' career development and explore the specific factors affecting English teachers' career development, so as to solve the problems in teachers' career development.

Key words: Data mining, Time Series, English Teacher, Career development, Predictive models

1. Introduction.

As modern science and technology develop rapidly, the Internet and artificial intelligence and other science and technology are reshaping educational forms such as teacher education development, and the development of network technology also provides data support for English teachers' professional development planning [1]. Based on data mining technology applied to teachers' professional education, through the collection of statistical teaching big data, deep learning and intelligent analysis of teachers' characteristics, it can provide personalized guidance solutions for teachers' quality improvement [2]. The career development stages of English teachers are generally divided into career exploration stage, career establishment stage, mid-career stage and late career stage [3]. Factors influencing English teachers' career development are generally categorized as social, family, personal and organizational [4, 5]. The Department of Teacher Education of the Ministry of Education in 2001 suggested that the main influencing factors for in-service teachers include school environment, life environment, teachers' social status, students, and teachers' peer groups [6]. There are serious issues with the professional development of English teachers, and the number of people participating in team building for teacher development is growing. In addition, teacher professional education is confronting new challenges related to reform and development [7]. The study proposes to use K-Means algorithm as a cluster analysis algorithm for data to predict teachers' career development, I use a time series prediction model and a Support Vector Machine (SVM) to categorize, analyze, and determine the stages of English teachers' promotion opportunities and the direction of their career development. It also explores the specific factors that affect English teachers' career development so as to solve the problems in teachers' career development.

2. Related works.

Addresses the difficulties of choice and career planning faced by English teachers in their professional development. With the rapid development of information technology, English teachers can efficiently and accurately choose the right career path for their development by using technologies such as data mining and artificial intelligence [8]. Tim A et al. addressed data mining for electrochemistry, with a general discussion from information to knowledge, describing the location of the nanochannels themselves by performing species transport on them [9]. Zou C et al. discovered high-strength ductile titanium alloys based on the integration of data mining and machine learning. And the integration allowed for more efficient and

*College of Humanities and Law, Gannan University of Science and Technology, Ganzhou, 341000, China (flp080881@163.com)

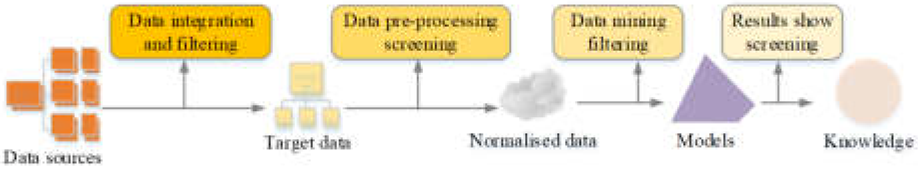


Fig. 3.1: Data mining specific flow chart

cost-effective design of high-strength and ductile titanium alloys [10]. By using data mining approaches, Feng Z et al. suggested identifying additional suspected harmful viruses in pangolins. They found two genomes of the genus Gemykibivirus and nine types of bat-associated circovirus, respectively [11]. With the implementation of five distinct text different classifiers that perform well in tweet categorization, Rahman R presented a real-time Twitter data mining method for inferring user mobile perceptions [12]. He Y et al. used data mining techniques with statistical metrics to analyse strategies for database-based energy-efficient building design, building a database of near-zero energy-consuming buildings and developing a customised data mining [13]. Parashkooh H I proposed a data mining technique based on oil-in-water droplet aggregation to guide the study of molecules and performed a series of molecular dynamics models. The findings exhibit a comprehension of the joint conduct of all species in multiphase systems, which can be applied to numerous emulsion formation fields [14]. Chen H et al. investigated an adaptive recommendation method based on online learning styles that can personalise learning resources according to users' pedagogical needs and personal preferences. Experimental results showed that the model facilitated data mining for learners and that the accuracy of recommendations was higher than other traditional recommendation models [15].

To process this temporal data, many researchers have created both short- and long-term memory artificially learning modules for temporal prediction models [16]. Wang H et al. proposed a more efficient fusion algorithm by combining BP neural networks with genetic algorithms and particle swarm optimization algorithms in data mining techniques [17]. The fusion algorithm was found to consume less energy and run more stably after experimental comparison [18]. To address the problem of predicting the career development of English teachers (CDET), this study proposes a model for predicting the CDET based on data mining and time series models. The model classifies and analyses the stages of teachers' career development and the direction of English teachers' career development based on the time series prediction model and SVM to explore the specific factors affecting English teachers' career development. The model uses an improved K-Means algorithm as a clustering analysis algorithm for data to predict teachers' career development.

3. English Teacher Career Development Analysis and Forecasting Design.

3.1. Construction of a Career Development Indicator System Based on K-means Clustering. The study uses techniques such as data analysis to predict the CDETs in relation to their current situation. Among the factors that affect English teachers' career development are the imbalances in gender, education, age, title and professionalism in the English teaching force, which in turn affect the development of the teaching force. Individual English teachers' professional development is influenced by both personal and organizational factors. Therefore, when studying the CDETs, the data can be aggregated to analyse the factors that influence their career development, and then to predict the CDETs. The study first uses data mining techniques to analyse the data and further optimize the predicted data in English teachers' career development. The specific flow chart is shown in Figure 3.1.

Finding information data or specific correlations between data that fulfill English teachers' objectives of professional growth from a huge amount of information data is the main goal of the data mining technique shown in Figure 1. The main steps are data integration and screening of data sources to form target data, followed by pre-processing screening operations on the target data. After the data has been normalized, the

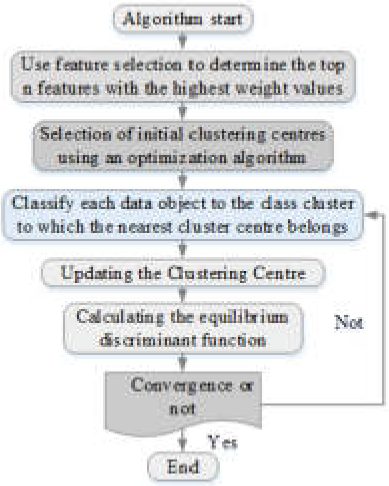

Algorithm start
Use feature selection to determine the top n features with the highest weight values
Selection of initial clustering centres using an optimization algorithm
Classify each data object to the class cluster to which the nearest cluster centre belongs
Updating the Clustering Centre
Calculating the equilibrium discriminant function
Not
Convergence or not
Yes
End

Fig. 3.2: Flow chart of improved K-Means feature selection algorithm

corresponding data model is formed by mining and filtering, and finally the results are presented. The application areas of data mining include clustering analysis, association rule analysis, feature algorithm analysis and classification analysis. The study uses the K-Means algorithm and improves it according to the problem of predicting the CDETs. The dataset chosen for this study originates from English professors in a university. It encompasses teachers' personal information, teaching accomplishments, research achievements, career plans, and other relevant details. The K-means algorithm is applied for cluster analysis due to its simplicity and ease of implementation. It is optimal for handling large-scale datasets. When selecting the basic principle behind K-means algorithm, the research mainly considers the following aspects: The K-means algorithm demonstrates a strong clustering effect. It offers high processing efficiency, enabling fast cluster analysis of large-scale data sets. Moreover, K-means has good interpretability and can directly reflect the distribution of data. Additionally, data mining technology is applicable to various data types. To better adapt K-means algorithm to the needs of English teacher career development prediction in cluster analysis, the initial cluster center is optimized. In the initialization phase of the algorithm, a more effective method for selecting the initial cluster centers is employed to circumvent local optimal solutions. Following this, the number of clusters is adjusted dynamically. During algorithm operation, the number of clusters gets dynamically adjusted based on evaluation index of clustering results for optimal clustering effect. Subsequently, to accurately represent the CDET, teacher's personal information, teaching achievements, scientific research achievements, and other relevant data characteristics are considered. Through the aforementioned enhancements, this study employs the K-means algorithm to extract and group the data. A comprehensive flowchart of the revised method is presented in Figure 3.2.

The detailed steps of the enhanced K-means feature selection algorithm are shown in Figure 3.2, starting with the initialization of the weights for each feature property. Secondly, the weight vector of feature attributes is calculated by the feature selection algorithm and reassigned. Third, the obtained feature attribute weight vector is sorted according to the weight value from largest to smallest, and then the top n elements are selected from all features. Fourth, n clustering centers with optimal contribution are selected among the initial ones. Fifth, the Euclidean distance calculation between data objects is performed, and based on the result of the calculation, the division is made into the class clusters with the smallest Euclidean distance value from the current data object, respectively. Sixth, a reassignment and update is performed at the centre of each class cluster. Seventh, a balanced discriminant function is used to calculate and then determine whether the final value gradually converges to 0. Eighth, if the function value is getting closer to 0, then the algorithm run ends, otherwise go to step 5. In terms of feature selection, the selected dataset D dataset is divided into partitions of $D = D_1, D_2, .., D_n$. The equation for calculating the weights of specific feature attributes is given in equation 3.1.

$$W_t^{i+1} = W_t^i - \sum_{x\epsilon T(c)} \frac{diff(t, D_i, x)}{n * d} + \sum_{x\epsilon S(S_i)} \left[\frac{q_c}{1 - q(D(D_i))} \sum_{x\epsilon G(c))} diff(t, D_i, x) \right] /(n * d) \tag{3.1}$$

In equation 3.1, each partition contains q attributes. where $D = D_{i1}, D_{i2}, \ldots, D_{in}$ and D_i attributes have the category $C_i \in C$, $C = C_1, C_2, \ldots C_k$ is a different set of categories for k, and the centre of mass is D_i. the data set will then be partitioned into categories, and subsequently will then go on to select d data objects from each category of data samples at a distance of D_i, where the updated weight vector for the feature attributes is $W = W_1, W_2, \ldots W_k$. n is the number of times the sample data is extracted. Equation 3.2 is used to calculate the diff(t, S_i, x) function, which indicates the differential function of the data objects.

$$\text{diff}(t, S_i, x) = \left| \frac{D_{it} - D_{jt}}{max_t - min_t} \right| \tag{3.2}$$

Equation 3.2 equalizes D_i distance to be more comparable to the d sample data by using the maximum and lowest values on the particular desired. For the optimal selection of the initialized clustering centers, assume that the data set, $X = x_1, x_2, \ldots, x_n$ has n different data objects, each containing p features, i.e. $x_i = x_{i1}, x_{i2}, \ldots x_{ip}$. Equation 3.3 defines the Euclidean distance between any two data objects, x_i and $1 < i < j < n$.

$$d(x_i, x_j) = \sqrt{\sum_{a=1}^{p} (x_{ia} - x_{ja})^2} \tag{3.3}$$

Equation 3.3 defines the distance density function for a sample data set X that corresponds to a data object named $x_i(1 < i < n)$ in equation 3.4.

$$\text{density}(x_i) = \sum_{j=1}^{n} \frac{d(x_i, x_j)}{\sum_{i=1}^{n} d(x_i, x_j)} \tag{3.4}$$

In equation 3.4, it is assumed that the neighbourhood radius R_i of data object $x_i(1 \leq i \leq n)$ in the dataset can be defined in equation 3.5, where $cR(0 \leq cR \leq 1)$ is used as a moderating factor for the neighbourhood radius. The clustering effect is expected to be more favorable if the value of cR is assumed to be 0.1, based on prior experience. The details are shown in equation 3.5.

$$R_i = n^{cR} * \frac{1}{n} \sum_{i=1}^{n} e^{-\textbf{density}(x_i)} \tag{3.5}$$

The data set X's data object is supposed to be $X_i(1 \leq i \leq n)$, with R_i centre and radius neighbourhood radius of the spherical region contains many data pairs, i.e. corresponding point density X_i , denoted as $S(X_i)$. In the spherical region where this data object is placed, there are more points per unit area the greater the value of $S(X_i)$, as shown in equation 3.6.

$$S(x_i) == |p|d(x_i, p) \leq R_i, p \in X| \tag{3.6}$$

Assume that $MS(x)$ is the average density of data objects in dataset x , see equation 3.7.

$$MS(x) = \frac{1}{n} \sum x \in XS(x) \tag{3.7}$$

The citation criterion function, also known as the objective function, is the criteria specified to decide data clustering. The objective function is used to calculate whether the similarity of the data objects in the same category meets the requirements and whether the differences between the different categories are close

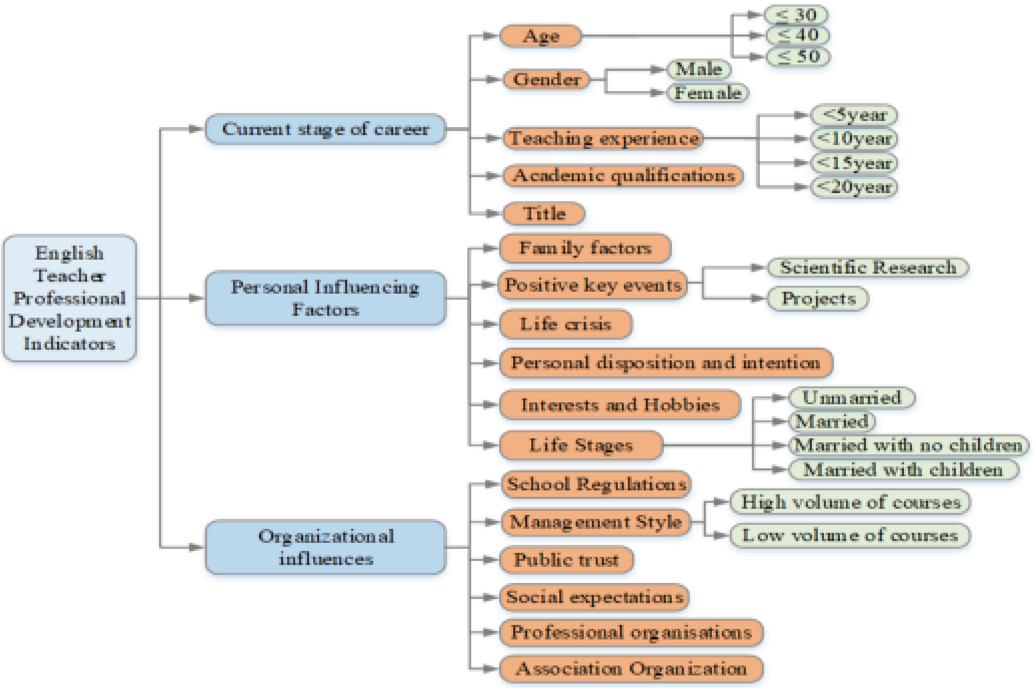


Fig. 3.3: Model diagram of English teacher career development indicator system

to the maximum. Based on this discriminative clustering, the optimal number of clusters can be obtained. The squared distance between the cluster's data objects and its center must be determined when intra-cluster variance is used to assess how compactly the clusters inside a cluster are organized, as shown in equation 3.8.

$$w(c)=\sum_{i=1}^{k} w(c_i)=\sum_{i=1}^{k}\sum_{x\in C_i} d(x_i,c_i)^2 \tag{3.8}$$

The computing the Euclidean distance from the cluster centers and subsequently the difference between the clusters, the difference between the clusters can be determined. Assuming that c_i and c_j are the centers of the ith and jth class clusters respectively, then the difference between the two class clusters $b(c)$. The detail is shown in equation 3.9.

$$b(c)=\sum_{1\leq j\leq k} d(c_j,c_i)^2 \tag{3.9}$$

The equilibrium discriminant function is then introduced in equation 3.10, where k is the number of clusters, and the differences between class clusters $b(c)$ and within class clusters $w(c)$ need to be first normalized.

$$W(c,k)=\frac{1}{1+e^{b(c)-w(c)}} \tag{3.10}$$

The factors that affect English teacher's professional development were analyzed through data clustering and aggregation. This process led to the identification of different factors that influence their development and resulted in the creation of an indicator system. The specific English teacher career development indicator system model is shown in Figure 3.3.

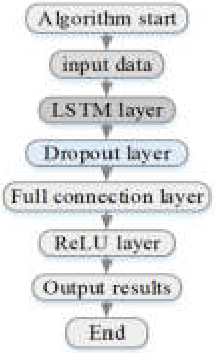


Fig. 3.4: LSTM model structure diagram

3.2. Time Series Model-based Design for Predicting English Teachers' Career Development. The theory of teacher professional development is divided into two areas: stages of professional development and influencing factors. Data mining techniques are used to determine and analyse the influencing factors in teacher professional development, while English teacher professional development requires a time series model of the stages of teacher professional development. The three main categories, from point to point and then comprehensive, are usually period theory, stage theory and cycle theory. Due to recent advancements in the realm of data science, LSTM has emerged as a highly efficacious solution for all time series prediction issues. As a recurrent neural network architecture, it can operate on varied interval values, rendering it a perfect fit for classifying, processing, and prognosticating time series with unknown durations or lags. The specific LSTM model structure is illustrated in Figure 3.4.

In Figure 3.4, a hierarchical analysis of the specific LSTM model structure is presented. In order to build the time-series features, the data was first pre-processed and then the factors influencing English instructors' professional development were identified. In constructing time-series features, objective information is melded to unify data traces from multiple behaviors in a time series. English teachers are then structurally entered into the LSTM network chronologically, based on age and teaching experience. Subsequently, an LSTM with a Dense function is assembled, holding an input layer with two fully connected layers. Finally 50 more features were extracted separately as time-series features to represent the dynamic changes of the above indicators. To further improve the interpretability of the behavioral features. The weights of each linear indicator for each teacher are calculated in equation 3.11.

$$\begin{cases} (N-\text{Rank}(x_n))/N, \text{Corr}(X_k)>0 \\ \text{Rank}(x_n)/N, \text{Corr}(X_k)<0 \end{cases} \tag{3.11}$$

In equation 3.11, N denotes N English teachers and K features extracted. $\text{Corr}(X_k)$ is the Pearson correlation coefficient between the kth feature X_k and the individual influences on English teachers' professional development, where $k \leq K$. $\text{Rank}(x_n)$ denotes the ranking of the Nth English teacher's professional development feature (denoted as $u_n \text{and} n \leq N$) among all English teachers' professional development factors. For example, there are four English teachers (u_1, u_2, u_3, u_4), if their kth characteristic (e.g., English teachers under thirty) is (0.8, 0.5, 0.7, 0.6), then $\text{Score}_{k1}=0, \text{Score}_{k2}=0.65, \text{Score}_{k3}=0.25 and \text{Score}_{k4}=0.15\text{Corr}(X_k)>0$. The weighted average of the characteristics is obtained by further substituting the following equation.

$$\sum_{k=1}^{k} = (|\text{Corr}(X_k)|^{*}\text{Score}_k^n) \tag{3.12}$$

Specifically, the average value of impacts on English teacher professional development is determined by a weighted average of all the influences with the respective weights defined by correlation coefficients. Following

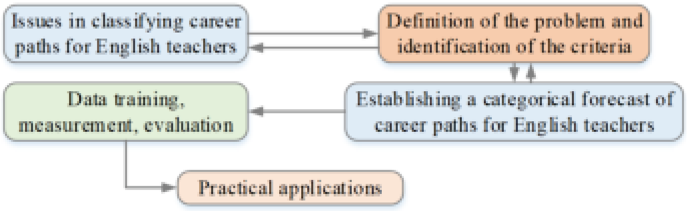

Issues in classifying career paths for English teachers
Definition of the problem and identification of the criteria
Data training, measurement, evaluation
Establishing a categorical forecast of career paths for English teachers
Practical applications

Fig. 3.5: English Teacher Professional Development Forecast Implementation Flow Chart

the above steps, a weighted average of English teacher career development and a weighted average of all characteristics can be calculated into the weighted average of English teacher career development factors and all characteristics, respectively. The data is then normalized. The specific normalization equation is shown below.

$$x_{scale} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.13}$$

In equation 3.13, x denotes the normalized variable object, x_{min} and x_{max} denote the minimum and maximum values of the variable object respectively, and x_{scale} is the normalized value, specifically ranging from 0 to 1. The practical linear classifier is then constructed, and the regression algorithm is later implemented by SVM for outlier detection to do the classification work. The specific English teacher career development prediction implementation process is shown in Figure 3.5. The SVM model was applied to the directional classification problem of English teacher career development. The goal of the model is to maximize the classification interval. The separation distance between two hyperplanes is the classification interval of the SVM, and the hyperplane should be oriented in such a way that it is as far away from the nearest data point of each class as possible. A decision boundary with a larger interval means that the model has a smaller generalization error. Whereas a smaller interval means that the model may overfit, the training samples that are closest to the hyperplane interval are the support vectors. With w serving as the weight vector and serving as the bias, the ideal hyperplane can be defined as $wx^T + b = 0\dot{w}$ and b will meet the requirements in the equation for each component of the training set.

$$\begin{cases} wx_i^T + b \geq 1, \text{if} = 1 \\ wx_i^T + b \geq -1, \text{if} = \text{-1} \end{cases} \tag{3.14}$$

Identifying w and b is the key to training the SVM model so that the hyperplane is as distant from the various classes of points as is feasible. Figure 3.6 depicts the SVM model's underlying principles.

500 linear features, 120 non-linear features, 60 LSTM-based features, and 30 basic features were among the 780 distinct types of characteristics that were retrieved for the study (including gender, age and teaching age). Firstly, pre-processing was carried out to zero-fill the behavioral features that were missing. Also, to eliminate order-of-magnitude differences between features, all features were scaled to between 0 and 1 using the maximum-minimum normalization method. The strategy for dealing with data gaps in the sample was to select the feature to be omitted when more than half of the gaps were present, and its gaps were zero-filled.

$$x' = \frac{x - min}{max - min} \tag{3.15}$$

Next, in the selection of features, the SelectKBest function in the scikit-learn library was used as the feature selection function and f-classif was used as the evaluation index in the python 3.8 environment to screen the set of features that have a more significant impact on the professional development of English teachers. Indeed, data usage and privacy protection are crucial ethical concerns for data-driven research. In this study pertaining to the prediction model for CDET, the following ethical implications must be addressed:

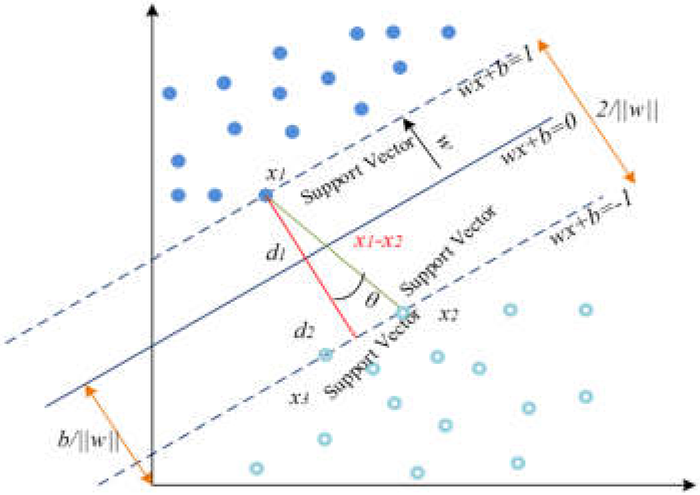


Fig. 3.6: Classification and prediction principle of SVM model

1. Data Use and Privacy protection: When collecting and utilizing teacher data, strict adherence to data protection principles and privacy policies is imperative. The collected data must be utilized exclusively for research purposes while ensuring necessary security measures are taken to safeguard against unauthorized access and use.
2. Transparency and research purposes: The utilization of teacher data in research should be transparent and overt. It is incumbent on researchers to explicitly articulate to faculty and other stakeholders the employment and objectives of the data while ensuring ethical standards are upheld.
3. Respect the wishes of teachers: Teachers' will and rights must be respected when collecting and using their data. Researchers should respect teachers' decisions not to participate in research or provide personal information and ensure that their choice does not hinder their professional growth.
4. Fairness and harmlessness: Research utilizing teacher data must abide by the principles of impartiality and non-harmfulness. The study must guarantee that it does not have a negative impact on the educators' professional growth, and that the complete variety of potential outcomes and impacts are adequately considered.
5. Sustainability and accountability: Research utilizing teacher data should take into account sustainability and accountability. Researchers must ensure that the data is sustainable and accessible, and implement measures to protect against unauthorized access and usage. In conclusion, the ethical considerations surrounding the use of teacher data for research are crucial. It is imperative that researchers abide by strict ethical guidelines, laws and regulations to safeguard against privacy breaches while ensuring transparency, respect for faculty wishes, fairness, harmlessness, sustainability, and accountability.

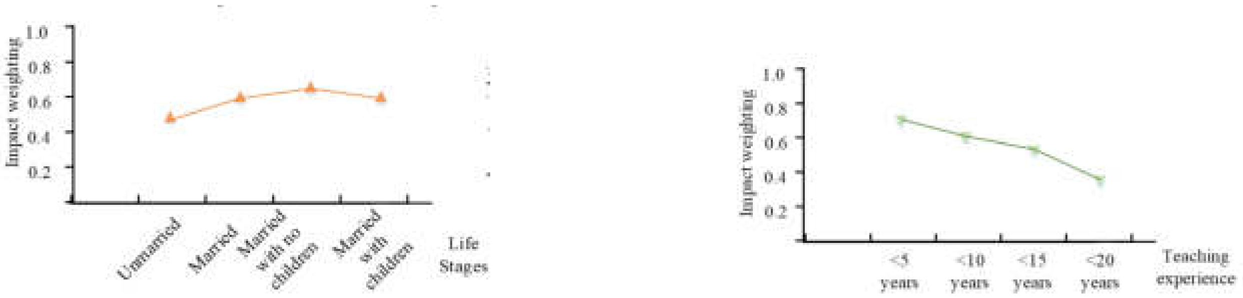
4. Analysis of the Effects of Predicted Professional Development of English Teachers. The system software environment based on the experiment was a Windows 10 system, an Intel(R) Core(TM) i7-6700 processor with 4G installed memory. The existing English teachers' career development database was imported into the prediction software and predicted online, and the existing data samples were compared and analysed according to the prediction results of the prediction model. A total of 1,000 English teacher career development data were processed to obtain 880 valid data, of which the results of the weighting of influencing factors in some English teacher career development predictions are shown in Table 4.1. The research findings are based on a substantial amount of experimental data and repeated tests. Senior university faculty were employed to analyze the test results, providing a high level of credibility for the study.

Table 4.1: Forecast results table

Tier 1 indicators	Weighting	Secondary indicators	Weighting
Career Status	0.332	Age	0.312
		Gender	0.251
		Teaching experience	0.354
		Title	0.345
Personal Influences	0.335	Family factors	0.325
		Positive key events	0.321
		Life crisis	0.254
		Personal disposition and intention	0.326
		Interests or hobbies	0.214
		Life Stages	0.462
Organizational influences	0.333	School Regulations	0.123
		Management Style	0.231
		Public Trust	0.341
		Social Expectations	0.325
		Professional Organization	0.252
		Association Organization	0.125

As the English teacher career development predictors are positive indicators. After calculating the weighting of the indicators in the table above, the results show that the most influential factor in the CDETs is life stage, with a weighting ratio of 0.462, which is one of the more important indicators of personal influence. This further indicates that each English teacher's career development focus varies at certain stages of life, either by improving their theoretical knowledge of English or pursuing further studies. Alternatively, they may choose to evaluate their titles and pursue additional education or concentrate on teaching and prioritize their families. The next factor is the number of years teaching English, which is weighted at 0.354, with different goals for English teachers in their career development. For new teachers, the immediate goal is to continue to hone their teaching and gain experience in teaching English. For teachers with some years of experience, their career development may be focused on refining their professionalism. The less influential factor in English teachers' career development is the organizational factor of school regulations, with a specific weighting of 0.123. The main reason for this is that school regulations are generally humane and therefore less influential in English teachers' career development. Life stage refers to the life cycle of teachers and is categorized into different periods: exploration, establishment, stability, maintenance, plateau, and retirement. The practical implications of life stage and school regulations are as follows. The impact of life stage varies during different stages of career development and presents English teachers with varying tasks and challenges. For instance, during the exploratory stage, teachers must adjust to the teaching setting, acquire teaching expertise, and establish their own educational concepts. In the stable stage, they face the challenge of career plateau and must sustain their enthusiasm and teaching drive. Hence, differences in life stages exert a significant influence on teacher professional growth. The Effects of School Regulations: School regulations are the norms of teaching and management that English teachers must adhere to. Reasonable regulations can safeguard the rights and interests of teachers while providing an optimal teaching and development environment. However, unreasonable regulations may impede teachers' instructional autonomy and professional growth. Consequently, school regulations are crucial factors that impact the CDET. Based on the analysis, it is evident that numerous factors impact the career growth of English teachers, with life stage and school regulations being the most significant. Thus, when developing the career plan for English teachers, one should consider the full impact of these factors. This consideration will provide better career advancement prospects and support for teachers. A visualization was used to show the variability of life stage and teaching age in English teachers' career development, as shown in Figure 4.2.

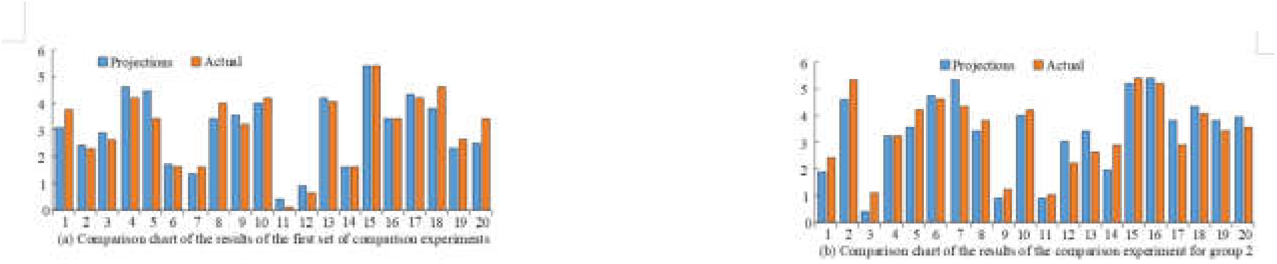
In the visual comparison in Figure 4.2, life stage is a personal influence on the CDETs, with a high weighting in the influence indicators. The English teacher career development model was completed and the data set was



(a) The impact of life stages on the professional development of English teachers

(b) The impact of teaching age on the career development of English teachers

Fig. 4.2: The impact of different life stages and years of teaching on the professional development of English teachers



(a) Comparison chart of the results of the first set of comparison experiments

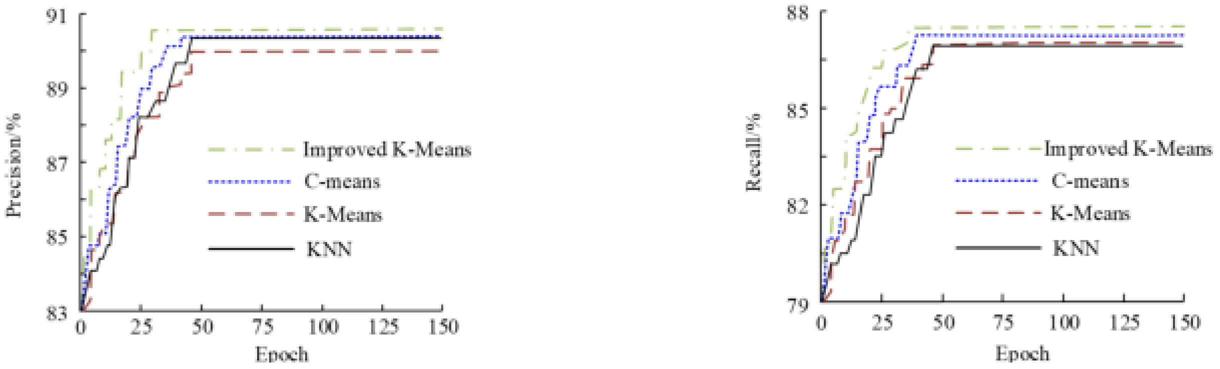
(b) Comparison chart of the results of the comparison experiment for group 2

Fig. 4.4: Validation of the graph of the correctness of the boys' data set with change

analyzed using data from the previous research. Experiments were also conducted on the relevant data to verify the performance and accuracy of the English teacher career development prediction model. The 2 sets of experimental results are shown in Figure 4.4.

Figure 4.4 demonstrates that the accuracy of the CDET prediction model following the aforementioned training model has met the accuracy criterion between the absolute error limit of 0.1. The two sets of sample data were compared, where the error between the actual sample data and the predicted sample data was 0.032. The results indicated that the trend prediction model for English teachers' career development had a high accuracy. The performance of the constructed indicators to evaluate the improved K-Means algorithm was tested to achieve the matching of the actual influencing factors of English teachers' career development. The data were divided into two datasets, one of which contained 480 records, 33 attributes and 323 entities. The other dataset contained 520 records, 33 attributes and 432 entities. The improved K-Means algorithm is compared with the three current state-of-the-art models. The Precision and Recall of the above three models are shown in Figure 4.6.

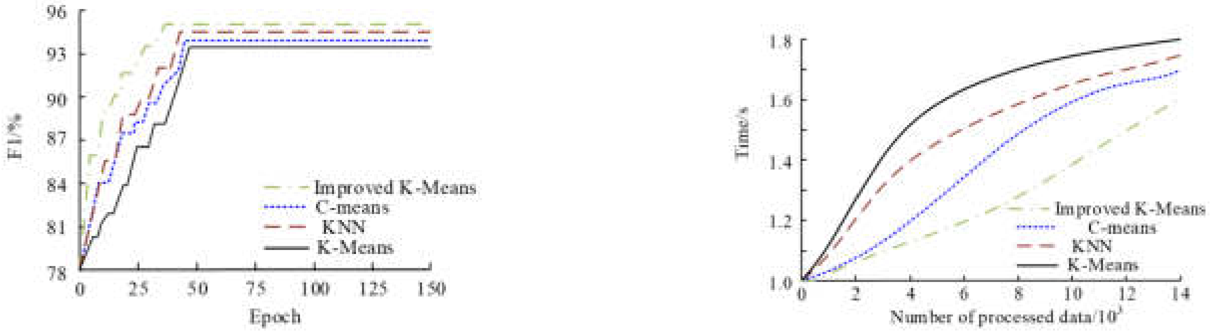
In Figure 4.5a, the highest Precision of the study model is achieved at 32 iterations, which is 10 and 16 times less than the C-means model and the KNN model respectively. At this stage, the precision of the improved K-Means model stands at 98.58%, surpassing the C-means model, KNN model, and traditional K-Means model by 0.12%, 0.36%, and 28%, respectively. In Figure 4.5b, the improved K-Means model attains the maximum recall of 44 iterations, which is significantly lower than the C-means model, KNN model, and traditional K-



(a) Precision

(b) Recall

Fig. 4.6: Precision and Recall of the model



(a) Epoch

(b) Time

Fig. 4.8: F1 and running time of the model

Means model, by 2, 5, and 7 times, respectively. The recall rate for the improved K-Means model was 86.62%, surpassing the C-means, KNN, and traditional K-Means models by 0.13%, 0.29%, and 0.32%, respectively. The above results demonstrated that the performance of the English teacher career development model constructed by the study based on the improved K-Means was superior. F1 and running time of the C-means model, KNN model and traditional K-Means model are shown in Figure 4.8.

In Figure 4.7a, the F1 of the C-means model, KNN model and traditional K-Means model are roughly positively correlated with the number of iterations. However, after a certain number of iterations, F1 stops growing and stabilizes. The F1 score for the improved K-Means model peaked at 31 iterations and then stabilized at 95.67%. In comparison, the C-means model only achieved its maximum F1 score after 46 iterations and remained stable thereafter. The F1 score for the traditional K-Means model eventually stabilized at 93.12%, which is 2.55% lower than that of the improved K-Means. The final F1 for the KNN model stabilized at 94.20%, which is 1.27% lower than that of the improved K-Means. When the sample size reaches 14000, the improved K-Means model requires 1.51 seconds in Figure 4.7b, which is 0.13 seconds, 0.19 seconds, and 0.21 seconds less than the C-means model, KNN model, and traditional K-Means model, respectively. It displays better performance than the other three models and proves to be superior. The improved K-Means model proposed in the study has a much better performance in entity matching. The performance of the four algorithms in

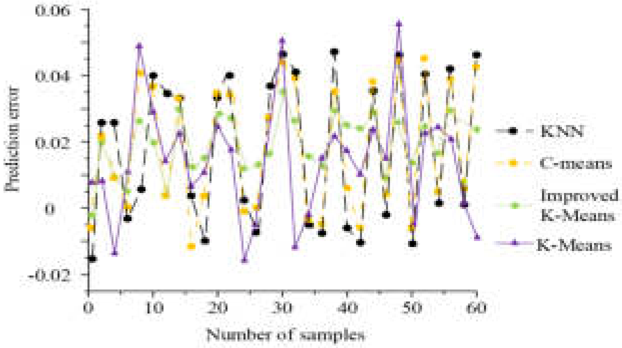


Fig. 4.9: Evaluation error results of different algorithms in training samples

Table 4.2: Comparison of the prediction results of several models under different data sets

Project	Data set 1		Data set 2		Data set 3	
	Forecast accuracy (%)	Prediction time (s)	Forecast accuracy (%)	Prediction time (s)	Forecast accuracy (%)	Prediction time (s)
Model 1	94.81	1.25	93.66	1.08	94.05	1.11
Model 2	90.44	1.88	90.13	1.96	90.28	1.97
Model 3	85.12	2.53	86.07	2.52	85.49	2.50
Model 4	89.47	2.15	89.18	2.22	89.72	2.10
Model 5	80.74	3.00	80.28	2.97	80.15	2.91

terms of evaluation error in the training sample is shown in Figure 4.9. As can be observed from Figure 4.9, all four algorithms are able to evaluate the given training samples, but the Improved K-Means algorithm is much better than the other three models. The prediction error of the improved K-Means algorithm can be controlled to within 0.002-0.032, possessing a better prediction accuracy compared to the remaining three algorithms. The traditional K-Means algorithm has a maximum prediction error of 0.055, with the worst prediction effect. kNN has a prediction error within -0.016-0.046, with a maximum prediction error value of 0.046. the C-means algorithm is second only to KNN, with a prediction error that can be controlled within -0.016-0.045, with a maximum prediction error value of 0.045. To assess the predictive capability of the model (Model 1) built by the Institute, Model 1 was compared with the more advanced career development prediction models in existing studies in three different data sets. Comparison models consist of a career development prediction model based on graph convolutional network (Model 2), a career development prediction model based on data mining (Model 3), a career development prediction model based on multi-attribute important weighted K-nearest neighbor algorithm (Model 4), and a career development prediction model based on extreme learning machine (Model 5). Table 4.2 displays the comparison results.

As Table 4.2 illustrates, model 1 shows an average prediction accuracy of 94.17% and an average prediction time of 1.15 seconds. In comparison to model 2, the average accuracy of its predictions is improved by 3.89%, 8.61% compared to model 3, 4.51% compared to model 4, and 13.78% compared to model 5. Furthermore, the average prediction time is reduced by 0.79 seconds compared to model 2, 1.37 seconds compared to model 3, 1.01 seconds compared to model 4, and 1.81 seconds compared to model 5. The model constructed by the research institute exhibits excellent predictive performance and achieves efficient and accurate career development predictions.

Precision, recall, and F1 values are widely employed in machine learning to assess classification model performance. These measures are interdependent. The accuracy rate represents the fraction of correctly predicted positive samples in the total predicted positive sample. The recall rate is the percentage of true positive cases correctly identified as positive. F1 score represents the balance between precision rate and recall rate, providing a comprehensive evaluation of the model's performance. In machine learning, it is crucial to consider both accuracy and recall for optimal classification performance. Focusing on one of these metrics exclusively may result in underperformance of the model in certain scenarios. For instance, if emphasis is solely placed on accuracy, it can lead to overlooking several true positive samples. Similarly, if only recall rates are given importance, there may be incorrect reporting of multiple negative samples. Thus, it is imperative to weigh the accuracy rate and recall rate to obtain an improved F1 value. Based on an analysis of various dimensions, the results demonstrate that the model significantly enhances the matching of entities in the K-Means algorithm. Compared to the other three models, the enhanced K-Means model exhibits superior performance in accuracy rate, recall rate, F1 value, and running time. Therefore, this model demonstrates stronger stability and higher efficiency when handling large-scale datasets. Comparative analysis reveals that the traditional K-Means algorithm has the least satisfactory performance in entity matching. The traditional K-Means algorithm is often impacted by noisy data and outliers, leading to subpar clustering outcomes, particularly when working with large data sets. By incorporating the weight mechanism, the enhanced K-Means algorithm accounts for the similarity among samples, leading to improved accuracy and stability of clustering. Moreover, despite the satisfactory accuracy and recall rates demonstrated by the KNN algorithm, its F1 value remains relatively low. This suggests that the KNN model may suffer from overfitting or underfitting issues when dealing with large-scale data sets. The C-means algorithm generally performs well in terms of accuracy and recall rates, but the F1 value is relatively high. The C-means model demonstrates stability and accuracy when handling large-scale data sets. The improved K-Means algorithm offers significant advantages in entity matching. In comparison to the other three models, the improved K-Means algorithm exhibits superior performance in accuracy rate, recall rate, F1 value, and running time. Therefore, in practical applications, the enhanced K-Means algorithm could enhance entity matching and improve the efficiency and accuracy of data processing.

5. Conclusion. In the field of English language teacher career development, this paper presents a data mining and time series model-based predictive approach for English teacher career growth. This study aims to thoroughly examine the current career status of English teachers, including individual and organizational factors that influence it. The K-means algorithm has been enhanced and utilized to implement cluster analysis. Next, time series features are constructed and inputted into an LSTM to develop a time series model. Finally, SVM is employed to predict the career progression direction of English teachers. The study revealed that life stage and teaching age were the most important factors influencing English teachers' career development, and their weight ratios were 0.462 and 0.354, respectively. These results showed that English teachers' career development had different characteristics and needs at different stages of their life cycle. Organizational factors like school rules and regulations had little influence on the CDET, with a weight ratio of 0.123. This could be attributed to the fact that the policies and regulations implemented in most schools tended to be more humane and had relatively little impact on the CDET. During the training of the prediction model, the error between the actual sample data and the predicted sample data was 0.032. When compared to the conventional K-means model, the enhanced K-means model demonstrated a 28% surge in precision, achieving 98.58%. This suggested that the model had a high degree of accuracy. These results indicated that the enhanced K-means model performs better in forecasting the career progression of English teachers. The study presents a fresh view and technique for promoting the career development of English instructors. It aids in a thorough comprehension of the requirements and traits for career progression and offers tailored policies and proposals for educational institutions and departments. Nevertheless, limitations exist, including scarce data sources and insufficient consideration of other factors that may influence English teacher career development. Future research can expand data sources and consider additional influencing factors to enhance the model's accuracy and applicability.

Although the research model demonstrates high predictive ability for the CDET, certain limitations persist. These limitations are as follows:

1. Limitation of data sources: The study data was obtained from a specific English teacher database, which may not represent all English teachers. Additionally, this study did not consider certain individual and organizational factors that could have a more significant effect on English teacher career development.
2. Limitations of model application: Although the model constructed in the study performs well in training sets, it may encounter uncertainties in forecasting novel situations or instructors. This is due to the

fact that models are trained using existing data and may lack sufficient information to make predictions about new scenarios or instructors.

3. Limitations of the improved K-Means algorithm: Although the enhanced K-Means algorithm performs well in cluster analysis, it exhibits weak outlier processing capabilities. The presence of outliers can significantly impact the clustering outcomes, ultimately affecting the model's overall performance.
4. Failure to consider individual differences of teachers: The study did not give full consideration to the individual differences among teachers, including their personal background, educational philosophy, and teaching style. These differences could potentially impact the career development of teachers in important ways.
5. Lack of long-term observation: The brief observation period of this study limits its ability to capture the long-term changes and trends in English teacher career development. A longer observation period could reveal more about the rules and factors influencing teacher career development. To address these limitations, future research can refine data collection methods and broaden data sources to enhance the model's accuracy in capturing the real-world conditions of diverse English teachers. Additionally, alternative clustering algorithms or ensemble learning techniques may be employed to upgrade the overall efficacy of the model. In addition, the influence of individual teacher differences on career development should be further explored in order to provide teachers with more targeted career development suggestions and planning.

Fundings. The research is supported by: Provincial Project, Teaching Design and Practice of Ideological and Political Case English Writing Based on POA Concept, (NO., JXJG-22-36-2).

REFERENCES

[1] Dey, L. & Mukhopadhyay, A. Biclustering-based association rule mining approach for predicting cancer-associated protein interactions. *IET Systems Biology.* **13**, 234-242 (2019)

[2] Eriya, K., Nugrahani, F. & Ghosh, A. Recommendation system using hybrid collaborative filtering methods for community searching. *Journal Of Physics: Conference Series.* **17**, 27-35 (2019)

[3] Jiang, W., Liu, P. & Wen, F. Speech Magnitude Spectrum Reconstruction from MFCCs Using Deep Neural Network. *Chinese Journal Of Electronics.* **3**, 42-47 (2018)

[4] Hai-Tao, L. & Yuan, S. Corrosion prediction of marine engineering materials based on genetic algorithm and BP neural network. *Marine Sciences.* **44**, 33-38 (2021)

[5] Zhou, K., Lin, W., Sun, J., Zhang, J., Zhang, D. & Feng, X. Prediction model of end-point phosphorus content for BOF based on monotone-constrained BP neural network. *Journal Of Iron And Steel Research International.* **29**, 751-760 (2022)

[6] Liu, M., Yao, D., Guo, J. & Chen, J. An Optimized Neural Network Prediction Model for Reservoir Porosity Based on Improved Shuffled Frog Leaping Algorithm. *International Journal Of Computational Intelligence Systems.* **15**, 11-19 (2022)

[7] Solodovnik, D., Tatonova, Y., Urabe, M., Besprozvannykh, V. & Inoue, K. Three species of Exorchis Kobayashi, 1921 (Digenea: Cryptogonimidae) in the East-Asian region: Morphological and molecular data. *Parasitology.* **148**, 1578-1587 (2021)

[8] Liu, M., Zhang, B., Li, X., Tang, W. & GQ., Z. An Optimized k-means Algorithm Based on Information Entropy. *The Computer Journal.* **64**, 1130-1143 (2021)

[9] Tim, A., Cao, X., Chen, D., Manuel, C., Edwards, M., Andrew, E., Stefano, F., Justin, G., Luke, G. & Mining, A. from information to knowledge: general discussion. *Faraday Discussions.* **233** pp. 58-76 (2022)

[10] Zou, C., Li, J., Wang, W., Zhang, Y. & Xu, D. Integrating data mining and machine learning to discover high-strength ductile titanium alloys. *Acta Materialia.* **202** pp. 211-221 (2021)

[11] Feng, Z., Dai, Z., Zhao, C., Jin, K., Shen, Q., Sun, R., Zhang, W., Yang, S., Wang, X. & Ning, S. Novel putative pathogenic viruses identified in pangolins by mining metagenomic data. *Journal Of Medical Virology.* **94**, 2500-2509 (2022)

[12] Rahman, R., Shabab, K., Roy, K., Zaki, M. & Hasan, S. Real-Time Twitter Data Mining Approach to Infer User Perception Toward Active Mobility. *Transportation Research Record.* **2675**, 947-960 (2021)

[13] He, Y., Chu, Y., Song, Y., Liu, M., Shi, S. & Chen, X. Analysis of design strategy of energy efficient buildings based on databases by using data mining and statistical metrics approach. *Energy And Buildings.* **258**, 1-11181 (2022)

[14] Parashkooh, H. & Jian, C. Data Mining Guided Molecular Investigations on the Coalescence of Water-in-Oil Droplets. 2022. (0)

[15] Chen, H., Yin, C., Li, R., Rong, W., Xiong, Z. & David, B. Enhanced learning resource recommendation based on online learning style model. *Tsinghua Science And Technology.* **25**, 348-356 (2020)

[16] Ta, X., Liu, Z., Hu, X., Yu, L., Sun, L. & Du, B. Adaptive Spatio-temporal Graph Neural Network for traffic forecasting. *Knowledge-based Systems.* **3**, 242-251 (2022)

[17] Wang, H., Song, L., Liu, J. & Xiang, T. An efficient intelligent data fusion algorithm for wireless sensor network. *Procedia Computer Science.* **183**, 418-424 (2021)

[18] Bollé, D. & Blanco, J. The Blume-Emery-Griffiths neural network with synchronous updating and variable dilution. *European Physical Journal, B.* **47**, 281-290 (2021)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: Nov 1, 2023
Accepted: Dec 12, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1603–1615, DOI 10.12694/scpe.v25i3.2781

RESEARCH ON THE CONSTRUCTION OF A HIGHER EDUCATION KNOWLEDGE MANAGEMENT MODEL BASED ON THE INTEGRATION OF SHADOW TEAMS

TIAN XIA[*]

Abstract. With the advancement of technology and the development of globalization, higher education institutions are facing increasingly complex knowledge management challenges. Shadow team usually refers to an informal and flexible team, where different departments or fields work together to promote knowledge exchange and integration, and stimulate innovative thinking. In order to enhance the knowledge management capabilities and competitiveness of higher education knowledge management institutions, this study will integrate shadow teams, construct a knowledge management model for higher education training enterprises, and use Analytic Hierarchy Process and Fuzzy Comprehensive Evaluation Method to construct an evaluation model for knowledge management capabilities of higher education training enterprises. The research results indicate that compared with traditional knowledge management models, the knowledge management model of higher education training enterprises integrating shadow teams collects 253 and 164 pieces of intelligence information at the 60th second, respectively; When the intelligence information is 300 pieces, the effectiveness of the analysis of the two models is 87.2% and 39.5%, respectively. An empirical analysis was conducted on a certain postgraduate entrance examination institution, and it was found that compared to students who reviewed independently, most of the students who participated in the training had significantly higher grades than those who did not receive training. The higher education knowledge management model that integrates shadow teams has stronger knowledge management capabilities and higher competitiveness.

Key words: Shadow Teams; Higher Education; Knowledge Management; Competitive Intelligence; Competency Evaluation; Training Companies

1. Introduction. With the development of information technology and the advent of the knowledge economy, competition among modern enterprises has become increasingly fierce, the competitive environment has become more complex, and the means of competition have become more diverse. Therefore, in order for enterprises to survive and develop in competition, they need to continuously improve their competitiveness. Competitive intelligence and knowledge management can promote enterprises to enhance their competitiveness. As a new technology, it is widely used in modern enterprise management and has become an important tool for enterprise management. Currently, higher education institutions are facing increasingly complex knowledge management challenges. With the rapid development of information technology, the amount and types of educational data are exploding, which poses higher requirements for knowledge management systems [3]. However, existing systems are often limited to simple organization and classification of collected data, and have not fully utilized this data to promote the improvement of education quality and optimization of management decisions [4]. Shadow team is a multidisciplinary and collaborative working group dedicated to solving problems in parallel informal environments and collaborating with formal teams. Shadow teams can bring new insights and solutions to organizations, thereby driving innovation and development [5]. The aim of this study is to explore and validate a new knowledge management model that integrates the concept of shadow teams, with the aim of improving the efficiency and innovation capabilities of knowledge management in higher education institutions through this integration. A systematic analysis and evaluation of it will help to better understand the current situation and potential of its knowledge management. Therefore, in order to promote the core competitiveness of higher education and training enterprises such as postgraduate entrance examination institutions, integrate shadow teams, effectively integrate knowledge management with competitive intelligence, construct a higher education knowledge management model, and construct an evaluation model for the knowledge management ability of higher education and training enterprises, to evaluate their knowledge management ability.

[*] Tian Xia, Faculty of Applied Technology, Huaiyin Institute of Technology, Huai'an, Jiangsu, 223001, China (xiatian150432@163.com)

2. Related Works. With the development of our society, knowledge is becoming more and more important, and the importance of knowledge in turn makes it increasingly important for enterprises to carry out knowledge management. For modern enterprises to enhance their knowledge management capabilities and to use knowledge as an intangible asset to create knowledge value in a sustainable manner, thereby enhancing the company's market competitiveness, numerous scholars have launched research on knowledge management.

Enis et al. used a qualitative approach to analyses the relationship between higher education partners based on a knowledge management perspective in order to promote the development of higher education in view of the important role of knowledge management in assisting partnerships to synergies knowledge and strengthen market competitiveness. The findings of the study showed that nurturing is the key to effective knowledge management and that effective knowledge management can facilitate collaboration in the higher education sector [6]. Horban used knowledge management to improve the management model of higher education institutions in order to enhance the quality of higher education, the results of the study showed that adding small additional projects can enhance the quality in higher education management [7]. Jarrahi et al. carried out personal knowledge management and knowledge construction in response to the emphasis of personal knowledge management on individual knowledge workers acquiring knowledge in the organizational environment. practice to analyses personal knowledge management activities, the results of the study showed that shadow information technology can facilitate consultancy management while supporting the construction and practice of personal knowledge base and knowledge management and help organizations to promote a balance between knowledge management strategies and personal knowledge goals [8]. Kranz et al. In order to study the impact information generated by shadow environmental management information systems on the environment in a chemical product in order to study the implementation of shadow strategies in a chemical products company, the results of the study showed that shadow EMIS can contribute to ecological sustainability [9]. Altmay et al. evaluated higher education practices in different countries in order to study the implementation of knowledge management in higher education institutions in distance education, using Nvivo qualitative data to analyses the results of the evaluation, the results of the study showed that knowledge management and sharing, the role of teachers and digital competencies play an important role in distance education [10]. The above studies show that knowledge management plays an important role in educational institutions. This is particularly true for higher education training companies, which are typically knowledge-intensive enterprises. It is also possible to identify from the above studies that shadow strategies also have an important impact on knowledge such as information, so this study will integrate shadow teams, construct a model of knowledge management in higher education and evaluate its knowledge management capabilities.

3. Building a higher education knowledge management model based on fused shadow teams.

3.1. Construction of the shadow team. Shadow teams originate from consulting firms, some of which provide customers with a basis for decision making and obtain information from competitors about strategy choices, thus forming a small team to track, analyze and forecast the situation of competitors required by target customers. The shadow team integrates the existing knowledge resources of the enterprise with the existing knowledge and findings of the enterprise, and organically combines the enterprise's knowledge management system with the competitive intelligence system, thus providing a strong support for the enterprise's strategic decision-making. In terms of key features, shadow teams are part of the competitive intelligence system, but their functions are different from those of the competitive intelligence system. Shadow teams can effectively integrate an enterprise's knowledge and intellectual capital, and can effectively analyze an enterprise's competitive situation, which is also an important tool for determining core competencies. By integrating the company's intellectual capital, the shadow team analyses the environment in which the company operates, identifies the company's core competencies and ensures that the company has an advantageous position in the competition.

Shadow teams are people selected from various functions in an organization who form a small, highly competitive team that tracks competitors' actions and mimics their next strategy to inform senior decision makers. The composition of a shadow team is shown in Figure 3.1.

Members of the shadow team track competitor divisions and implement monitoring, modelling and forecasting in order to provide dynamic decision support to decision makers, and their number can vary depending on the job content and the competitor. It is particularly important to note that the liaison person has the special feature of being either an intelligence specialist, a team leader or a member of the shadow team, acting

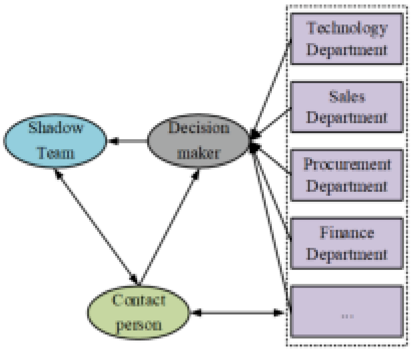


Fig. 3.1: Composition of shadow teams

as a bridge between senior decision makers and junior staff and the shadow team [11]. Usually, the shadow team is relatively small and the number of people is sufficient as long as the task can be accomplished.

This study advocates that the determination of shadow team members should be based on objectives. If the objective is simply to get to know a competitor, only one team is needed. In this case, a team of 3-5 people is basically adequate. If the objective is to achieve an important strategy or to contribute to the long-term development of the company, a larger team is required to achieve the objective. In this case, there are more than ten or even twenty people in a team, who can be divided into several groups and act together.

When setting up a shadow team, it is also important to choose a liaison person or two to help them achieve their goals. The liaison person is an important member of the management team and handles communication with other departments or teams, in addition to communication within the team and at the leadership decision level. The number of liaisons can be determined by the size of the team. When selecting a liaison person, it is best to choose someone who has sufficient time and is able to communicate with the company's decision makers as well as the ground floor staff.

3.2. Model building for the integration of knowledge management and competitive intelligence in higher education. For competitive intelligence and knowledge management to perform their functions simultaneously, they must be supported by the appropriate organization. The integration of knowledge management and competitive intelligence within an enterprise is a key step in achieving this within the enterprise. The principles of organizational integration include the demand-driven principle, the principle of communication sensitivity, the principle of information sharing, and the principle of flexibility of organizational structure [12]. If an enterprise is to gain its competitive advantage in a fiercely competitive market, its organizational structure must have a high degree of flexibility and market adaptability. Based on the principles of organizational integration analyzed earlier, an organizational structure that integrates an enterprise's knowledge management with competitive intelligence is established, as shown in Figure 3.2.

As shown in Figure 3.2, based on the concept of shadow teams, adhering to the principles of driving, communication sensitivity, information sharing, and organizational flexibility, external information sources of the enterprise are formed through enterprise intelligence strategic alliances, industry research companies, experts, or consulting companies. Transferring it to the internal organization, the CEO manages the knowledge and intelligence department, which is mainly composed of the competitive intelligence division, knowledge management division, and shadow team. And connect with functional department bridging personnel to form the organizational structure of the enterprise knowledge management and competitive intelligence integration model [13]. The integrated knowledge management and competitive intelligence departments will be unified as the knowledge intelligence department, thus providing new ideas and methods for the knowledge management and competitive intelligence work of enterprises. In the knowledge intelligence department, a knowledge intelligence leader will be established to provide unified guidance to the work of the enterprise's knowledge management sub-department, competitive intelligence sub-department and shadow team, so as to effectively integrate

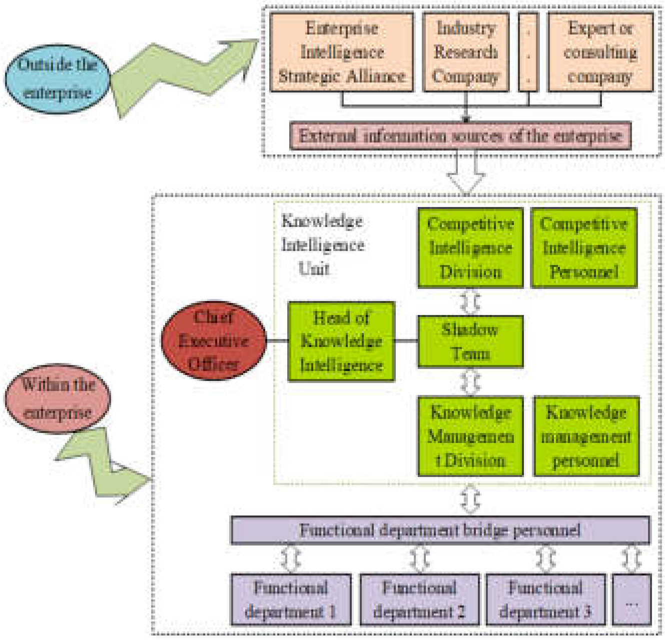

Outside the enterprise
Enterprise Intelligence Strategic Alliance
Industry Research Company
Expert or consulting company
External information sources of the enterprise
Knowledge Intelligence Unit
Competitive Intelligence Division
Competitive Intelligence Personnel
Chief Executive Officer
Head of Knowledge Intelligence
Shadow Team
Within the enterprise
Knowledge Managemen t Division
Knowledge management personnel
Functional department bridge personnel
Functional department 1
Functional department 2
Functional department 3
...

Fig. 3.2: Organizational Structure of Integrating Enterprise Knowledge Management and Competitive Intelligence

competitive intelligence and knowledge management with the enterprise's strategic planning and competitive strategies, thereby improving the enterprise's competitiveness. The shadow team is an intelligence team jointly formed by personnel from different functional departments, integrating the organization's internal knowledge assets and external information, and conducting analytical and research-based research on them with the organization's strategy and decision-making in mind, thus enabling the transformation of the enterprise's knowledge into active intelligence and promoting the integration of knowledge management and competitive intelligence. Functional bridging staff refers to full or part-time knowledge management and competitive intelligence staff in various functional departments [14].

An analysis of the processes of enterprise knowledge management and competitive intelligence activities reveals that the processes of the two have an inter-integration relationship. In this inter-integration relationship, the enterprise knowledge and intelligence information sharing platform plays a key role as a bridge. Therefore, this research constructs a process integration model of enterprise knowledge management and competitive intelligence based on the enterprise information sharing platform for the purpose of enterprise decision-making application, as shown in Figure 3.3.

Figure 3.3 shows that the ideas, methods and technologies of knowledge management are introduced into the process of competitive intelligence, reconstructing the competitive intelligence workflow of the enterprise. Through this model, the process of combining knowledge management and competitive intelligence can be effectively promoted, and advanced knowledge management ideas and technologies are applied to improve the methods of competitive intelligence and enhance the competitive intelligence capability of enterprises.

3.3. Knowledge management capability evaluation model construction for higher education training enterprises. The assessment index system of knowledge management capability is a complex and dynamic system, and when establishing the evaluation index system, it should start from various aspects. To ensure the accuracy of the assessment results, the principles of scientific and feasibility, comprehensiveness and systematicity, and a combination of objectivity and adaptability should be followed in designing the indicator system [15].

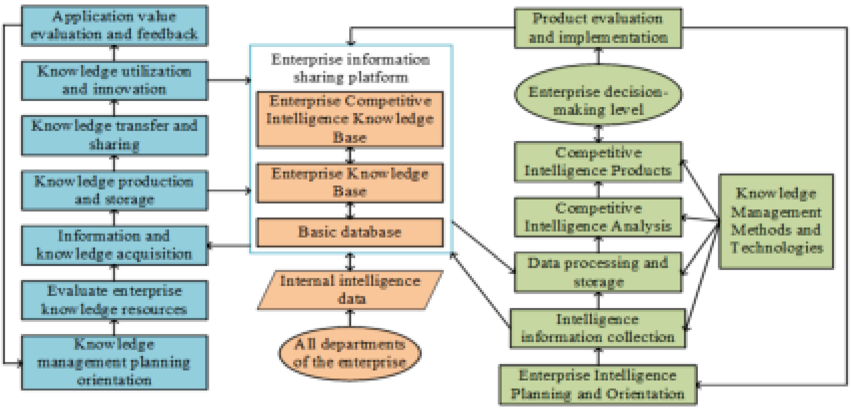

Application value evaluation and feedback
Knowledge utilization and innovation
Knowledge transfer and sharing
Knowledge production and storage
Information and knowledge acquisition
Evaluate enterprise knowledge resources
Knowledge management planning orientation
Enterprise information sharing platform
Enterprise Competitive Intelligence Knowledge Base
Enterprise Knowledge Base
Basic database
Internal intelligence data
All departments of the enterprise
Product evaluation and implementation
Enterprise decision-making level
Competitive Intelligence Products
Competitive Intelligence Analysis
Data processing and storage
Intelligence information collection
Enterprise Intelligence Planning and Orientation
Knowledge Management Methods and Technologies

Fig. 3.3: Knowledge Management and Competitive Intelligence Process Integration Model

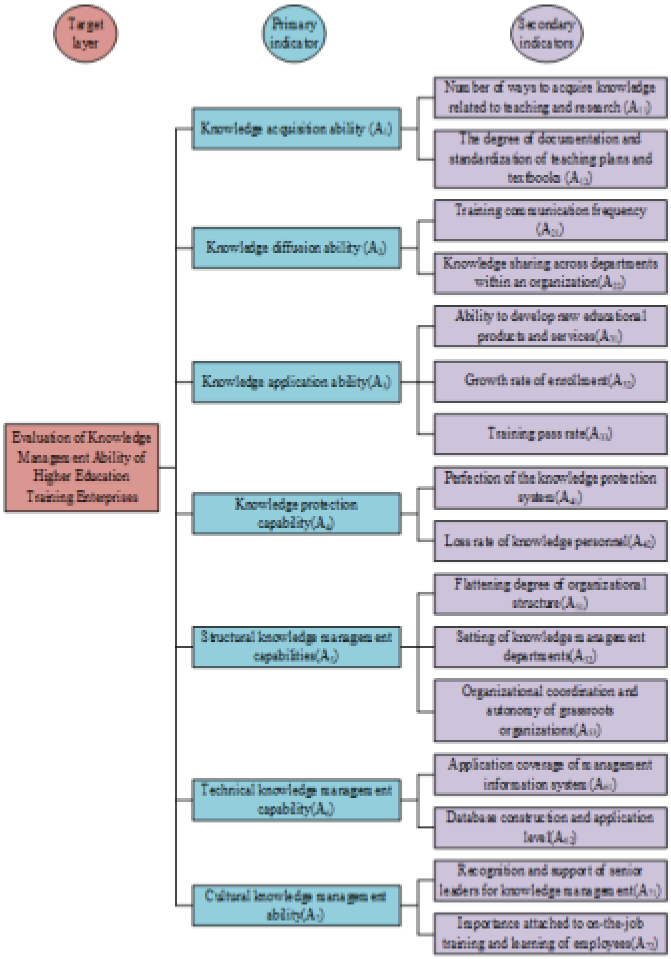

Target layer
Primary indicator
Secondary indicators
Evaluation of Knowledge Management Ability of Higher Education Training Enterprises
Knowledge acquisition ability (A_1)
Number of ways to acquire knowledge related to teaching and research (A_{11})
The degree of documentation and standardization of teaching plans and textbooks (A_{12})
Knowledge diffusion ability (A_2)
Training communication frequency (A_{21})
Knowledge sharing across departments within an organization(A_{22})
Knowledge application ability(A_3)
Ability to develop new educational products and services(A_{31})
Growth rate of enrollment(A_{32})
Training pass rate(A_{33})
Knowledge protection capability(A_4)
Perfection of the knowledge protection system(A_{41})
Loss rate of knowledge personnel(A_{42})
Structural knowledge management capabilities(A_5)
Flattening degree of organizational structure(A_{51})
Setting of knowledge management departments(A_{52})
Organizational coordination and autonomy of grassroots organizations(A_{53})
Technical knowledge management capability(A_6)
Application coverage of management information system (A_{61})
Database construction and application level(A_{62})
Cultural knowledge management ability(A_7)
Recognition and support of senior leaders for knowledge management(A_{71})
Importance attached to on-the-job training and learning of employees(A_{72})

Fig. 3.4: Evaluation Index System for Knowledge Management Ability of Higher Education Training Enterprises

Through discussions and consultations with industry experts, scholars, and corporate executives, the importance of evaluation indicators is determined, and factors crucial to the success of enterprise knowledge management are considered as the main indicators. Through this hierarchical and segmented approach, enterprises can more accurately identify their strengths and areas for improvement in knowledge management, and based on this, develop targeted improvement strategies. The index system for assessing the knowledge management capability of higher education training enterprises adopts a three-tier structure, of which the first tier is the target tier, that is, the ultimate purpose of the assessment is to comprehensively assess the knowledge management capability of higher education training enterprises; the second tier is the primary index tier, which defines the knowledge management capability of enterprises from seven levels; the third tier, which is further subdivided from the secondary index tier downwards The third level is the second level indicator layer, which is further subdivided from the second level indicator layer. Finally, the assessment index system of knowledge management capability of higher education training enterprises, which consists of 7 primary indicators and 16 secondary indicators, was formed, and the results are shown in Figure 3.4.

3.4. Determination of the Weight of Knowledge Management Capability Evaluation Indicators for Higher Education and Training Enterprises. In order to determine the weights of each indicator in a more comprehensive, reasonable and effective manner, both qualitative and quantitative approaches were adopted to determine the weights of each indicator in conjunction with the characteristics of the knowledge management capabilities of higher education training enterprises. After an in-depth analysis, the hierarchical analysis method was also used to determine the weights of each indicator for the knowledge management capability of higher education training enterprises. A is the judgment matrix constructed on the basis of the recursive hierarchy constructed. After constructing the judgement matrix, a mathematical process is applied to sort it hierarchically to obtain the relative weights of the level in relation to the previous level [16]. λ_{max} In other words, the maximum characteristic root of the judgement matrix and its corresponding eigenvector W are calculated and then normalized to W . This study utilizes the sum-product method, which is relatively simple to compute, by first regularizing each item of the judgement matrix, with the expression shown in equation 3.1.

$$(\bar{a}_{1j}) = \frac{a_{ij}}{\sum_{k=1}^{n} a_{kj}} \quad \text{for } i,j = 1,2,\ldots,n \tag{3.1}$$

The subsequent rows are summed to obtain the sum vector and the expression is shown in equation 3.2.

$$(\bar{W_1}) = \sum_{j=1}^{n} \bar{a_i}j \quad \text{for } i,j = 1,2,\ldots,n \tag{3.2}$$

The resulting sum vector is regularized to be able to obtain the feature vector, the expression of which is shown in equation 3.3.

$$(\bar{W_1}) = \bar{W_1}/\sum_{j=1}^{n} \bar{W_1} \quad \text{for } i,j = 1,2,\ldots,n \tag{3.3}$$

Finally, the maximum eigenvalue is calculated and the expression is shown in equation 3.4.

$$\lambda_{max} = \sum_{i=1}^{n} \frac{(AW)_i}{nW_i} = \frac{\sum_{i=1}^{n} \frac{(AW)_i}{W_i}}{n} \quad \text{for } i,j = 1,2,\ldots,n \tag{3.4}$$

After that it can be tested for consistency, it is necessary to calculate the Consistency indicators (CI) first, when the numerator is larger, it means that the consistency of the judgment matrix is worse; when CI=0, it means that the judgment matrix has full consistency [17]. The expression of CI is shown in equation 3.5

$$CI = \frac{\lambda_{\max} - n}{n-1} \tag{3.5}$$

And the corresponding average random consistency index RI can be determined through the index system, which in turn leads to the consistency ratio of the judgment matrix, with the expression shown in equation 3.6.

$$CR = \frac{CI}{RI} \tag{3.6}$$

If the consistency ratio is $CR < 0.1$, it means that the consistency of the judgment matrix is an accepted state; when $CR > 0.1$, it means that the judgment matrix does not meet the requirements of consistency and this judgment matrix needs to be revised again. The total hierarchical ranking refers to the relative weight of each factor in each target level, arranged in a hierarchical order from top to bottom. The relative weight of a tier of indicators to all indicators is its single ranking result, as shown in expression 3.7.

$$W_A = (W_{A1}, W_{A2}, W_{A3}, \ldots, W_{An}) \tag{3.7}$$

The expression for the relative weighting of secondary indicators to primary indicators is shown in equation 3.8.

$$W_{A_j} = (W_{A_j1}, W_{A_j2}, W_{A_j3}, \ldots, W_{A_jn}) \tag{3.8}$$

Combining equation 3.7 and equation 3.8, equation 3.9 can be obtained.

$$W_{A \to A_{ij}} = W_{A \to A_i} * W_{A_i \to A_{ij}} \tag{3.9}$$

Consistency tests for total and single sorting were tested and the results showed that the consistency of the judgement matrix was satisfactory for a consistency ratio of. In the process of knowledge management activities of higher education training enterprises, many factors that are difficult to quantify and the fuzziness of human subjective judgment are involved. The fuzzy comprehensive evaluation method is a comprehensive evaluation method based on fuzzy mathematics, and based on the affiliation theory of fuzzy mathematics, it can achieve the purpose of transforming qualitative evaluation into quantitative evaluation, which can well solve the fuzzy and difficult to quantify difficulties in the current evaluation of the enterprise's knowledge management capability [18]. Firstly, the set of indicators needs to be determined, and the expression is shown in equation 3.10.

$$A = \{A_1, A_2, A_3, A_4, A_5, A_6, A_7\} \tag{3.10}$$

In equation 3.10, A denotes the first level of the knowledge management capability evaluation index system of higher education training enterprises, which is also the target level. Meanwhile, the study divided the evaluation levels in the model into five levels, namely: excellent, good, fair, poor and poor, and the expressions are shown in equation 3.11.

$$V = \{V_1, V_2, V_3, V_4, V_5\} \tag{3.11}$$

In equation 3.11, V indicates the evaluation level, V_1, V_2 , V_3, V_4 and V_5 indicate the excellent, good, fair, poor and poor levels of the rating respectively. According to determine the number of people who belong to a certain level of an indicator, the proportion of the total number of participants in the questionnaire r_{ij} , and finally get the vector of affiliation value of the indicator, the calculation formula is shown in equation 3.12.

$$r_{ij} = \frac{m_{ij}}{m} \tag{3.12}$$

In equation 3.12,m_{ij} represents the number of people who classified the indicator as level and is used to describe the total number of participants in the questionnaire. The single-factor judgement yields the affiliation vector $r_i = (r_{i1}, r_{i2}), r_{i3}, r_{i4}, r_{i5}$, which in turn enables the affiliation matrix to be obtained, the expression of which is shown in equation 3.13.

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} & r_{15} \\ r_{21} & r_{22} & r_{23} & r_{24} & r_{25} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & r_{n3} & r_{n4} & r_{n5} \end{bmatrix} \tag{3.13}$$

Table 4.1: System parameter

Number	Project	Size	Unit
(1)	Operating system	UNIX	/
(2)	Programming tools	Python	/
(3)	Working voltage	220	V
(4)	Memory	1020	Mb

After calculating the set of weights and the affiliation matrix of the indicator system, the comprehensive judgment vector is solved and the expression is shown in equation 3.14.

$$B = W \cdot R = (b_1, b_2, \ldots, b_5) \tag{3.14}$$

In equation3.14, $b_1, b_2, \ldots, b_5$ indicates the comprehensive evaluation results of the evaluation indicators, reflecting the subordination relationship between the comprehensive evaluation indicators of each level and each tier. The comprehensive evaluation is then carried out in accordance with the principle of maximum affiliation, and the conclusion of the comprehensive evaluation is obtained.

If the weight of the model changes, it may significantly affect the evaluation results and final decision, and sensitivity analysis is required. Conducting sensitivity analysis is an important step in evaluating the effectiveness and reliability of a model. To ensure the reliability and applicability of the model, it is recommended to conduct sensitivity analysis by adjusting different weight combinations and observing changes in the results after developing the initial model. This can ensure that the adopted model can still function even in different contexts and make wise adjustments and decisions.

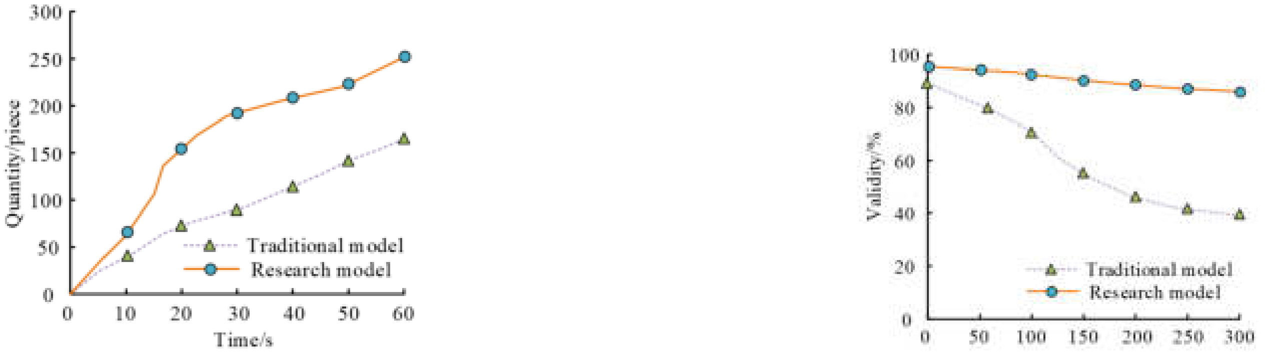
The steps of sensitivity analysis are as follows: determining variables, adjusting variables, observing results, analyzing trends in changes, and explaining and applying them. Firstly, identify the key variables in the model that may affect the results, such as the weights of each evaluation indicator. Systematically changing the values of these key variables (weights) can be a single variable change or adjusting multiple variables simultaneously. Record the impact of these adjustments on the final evaluation results, such as the ranking or score of the company's knowledge management capabilities. Analyze the trend of changes in the results with key variables and understand which variables have the greatest impact on the results. Based on the results of sensitivity analysis, explain and make necessary adjustments to the model or evaluation system.

4. Higher Education Training Enterprise Knowledge Management Capability Evaluation Model and Capability Analysis.

4.1. Analysis of a model for integrating knowledge management and competitive intelligence in higher education training enterprises. A knowledge management model for higher education training enterprises incorporating shadow teams was analyzed and compared with a traditional knowledge management model to verify the validity of the model constructed by the study. The two models were conducted under the same experimental conditions and with the same parameter settings in the experiments, where the parameters of the models are shown in Table 4.1.

According to the competitive intelligence needs of higher education training enterprises, plan their competitive intelligence objectives and implementation priorities, and determine the direction and scope of intelligence collection, and compare the intelligence information collection and competitive intelligence analysis results of the two models, the results are shown in Figure 4.2.

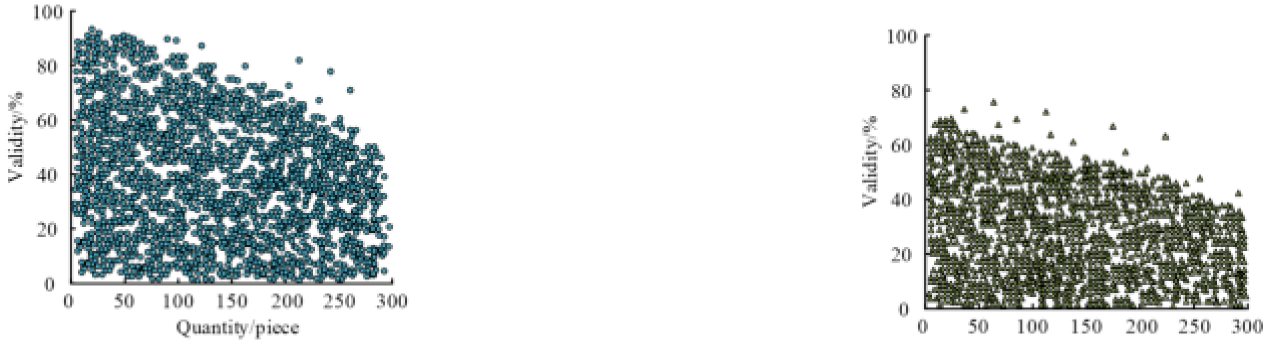
As can be seen from Figure 4.2, the number of intelligence information collected by the two models increases as time increases, but the number of intelligence information collected by the model constructed by the research method is significantly more than that of the traditional model; at 60s, the number of intelligence information collected by the two models is 253 and 164 respectively. In the competitive intelligence analysis, the effectiveness of the analysis decreases as the number of intelligence increases, but the rate of decline of the model constructed by the research method is significantly lower than that of the traditional model; when the intelligence information is 300, the effectiveness of the analysis of the two models is 87.2% and 39.5%



(a) Intelligence Information Collection

(b) Competitive Intelligence Analysis

Fig. 4.2: Intelligence information collection and competitive intelligence analysis results of two models



(a) Research Model

(b) Traditional Model

Fig. 4.4: Two Models for Analyzing Valuable Intelligence Information Results

respectively. The effectiveness of competitive intelligence utilisation is compared and analysed through the competitive intelligence evaluation and feedback mechanism in the knowledge management and competitive intelligence integration model of knowledge higher education training enterprises. Firstly, the results of the two models' analysis of tacit knowledge are shown in Figure 4.4.

Figures 4.3a and 4.3b show the analysis of tacit knowledge by the model constructed by the research and the traditional model respectively. It can be seen from Figure 6 that the model constructed by the research method has a significantly better efficiency profile for the analysis of tacit knowledge than the traditional model. The two models were then analyzed for valuable intelligence information and the results are shown in Figure 4.6.

Figure 4.5a and Figure 4.5bshow the analysis of valuable intelligence information by the model constructed by the research and the traditional model respectively. It can be seen from Figure 7 that the efficiency of the models constructed by the research method for the analysis of valuable intelligence information is also significantly better than that of the traditional models, and the efficiency of the two models for the analysis of valuable intelligence information is better than that of the analysis of tacit knowledge.

4.2. Empirical analysis of knowledge management capabilities of higher education training enterprises. Taking an examination and research institution as an example, an empirical analysis was conducted on its knowledge management capability. The information on the assessment of the knowledge management

Validity/%
100
80
60
40
20
0
0 50 100 150 200 250 300
Quantity/piece

(a) Research Model

(b) Traditional Model

Fig. 4.6: Two Models for Analyzing Valuable Intelligence Information Results

Table 4.2: Evaluation Index System for Knowledge Management Ability of Education and Training Institutions

Target layer	Primary indicator	Secondary indicators	Weight of each indicator relative to the target layer
Evaluation of Knowledge Management Ability of Postgraduate Examination Institutions	A1	A11	0.0557
		A12	0.0186
	A2	A21	0.1367
		A22	0.0273
	A3	A31	0.2321
		A32	0.0774
	A4	A33	0.0774
		A41	0.1757
		A42	0.0586
	A5	A51	0.0052
		A52	0.0171
		A53	0.0094
	A6	A61	0.0143
		A62	0.0429
	A7	A71	0.0431
		A72	0.0086

capability of the examination and research institution was obtained by means of visits and surveys; at the same time, the fuzzy comprehensive evaluation data was obtained from the questionnaire of the staff of the institution, and the questions contained in the questionnaire were all from the 16 secondary indicators of the evaluation index system. To ensure the validity and feasibility of the questionnaire, it was randomly distributed to the staff of the five functional units of the organization, with a representative number of employees in each area and with a certain understanding of the overall situation of the company. The results of the determination of the indicator system weights are shown in Table 4.2.

Combining the constructed evaluation model of knowledge management capability of higher education training and the determined weights of the evaluation index system of the examination and research institutions, the fuzzy comprehensive evaluation method was used in order to verify the validity of the evaluation model of knowledge management capability of higher education training enterprises. On this basis, the affiliation matrix of each indicator was constructed based on the 30-questionnaire data, combined with the indicator weights determined in the previous section, and the results are shown in Table 4.3.

From this, we can obtain the affiliation matrix, which can then be evaluated to obtain the knowledge

Table 4.3: Membership degree of each indicator of knowledge management ability of postgraduate entrance examination institutions

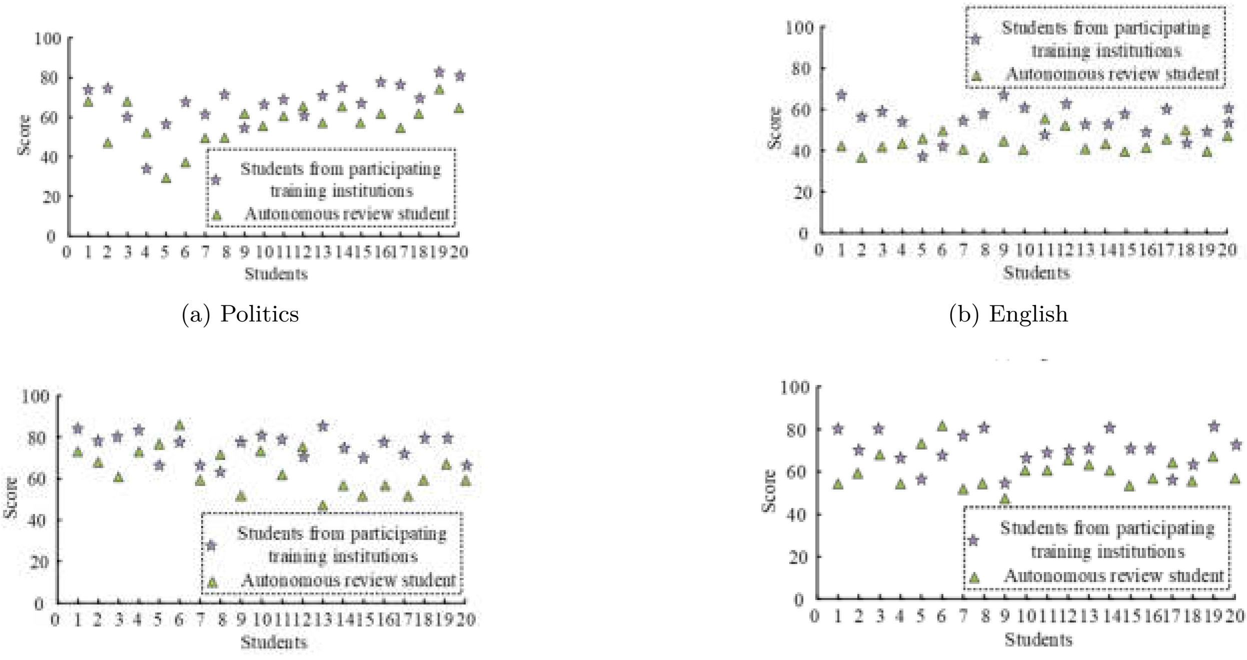
Primary indicator	Secondary indicators	Weight	Fuzzy evaluation of indicators				
			Excellent	Good	Common	Range	Bad
A1	A11	0.0557	0.37	0.43	0.17	0.03	0.00
	A12	0.0186	0.23	0.30	0.37	0.10	0.00
A2		0.1367	0.30	0.43	0.17	0.07	0.03
	A22	0.0273	0.07	0.40	0.40	0.10	0.03
A3		0.2321	0.07	0.30	0.53	0.10	0.00
	A32	0.0774	0.07	0.63	0.30	0.00	0.00
A4		0.0774	0.07	0.67	0.27	0.00	0.00
		0.1757	0.03	0.43	0.40	0.07	0.07
	A42	0.0586	0.00	0.00	0.23	0.67	0.10
A5		0.0052	0.00	0.10	0.43	0.37	0.10
		0.0171	0.20	0.30	0.43	0.07	0.00
	A53	0.0094	0.10	0.23	0.47	0.17	0.03
A6		0.0143	0.20	0.43	0.30	0.07	0.00
	A62	0.0429	0.03	0.50	0.43	0.03	0.00
A7		0.0431	0.17	0.40	0.37	0.07	0.00
	A72	0.0086	0.40	0.43	0.17	0.00	0.00

acquisition ability, knowledge diffusion ability and the fuzzy evaluation vector of the examination and research institution. Finally, according to the principle of maximum affiliation, it can be seen that the knowledge management ability of this education and training institution belongs to the five grades of ”good”, which means that the knowledge management ability of this examination and research training institution is in the middle to upper level. Forty students with the intention of taking the examinations and with an average score of 80±5 in the final examinations at the top and bottom of their junior year were selected and divided into two groups, one group participating in the training of the examination institution and one group studying on their own, and their examination results were compared to verify the knowledge management ability of the training institution. The results are shown in Figure 4.8.

In Figure 4.8, the test scores have been standardized by converting all scores to a total of 100. Figure 4.8 shows that the results of the students who participated in the training were, for the most part, significantly higher than those of the untrained students. This indicates that the institution has good knowledge management skills and has good first-hand knowledge to provide students with more informative training to help them achieve better exam results. To further evaluate the institution, the participating students were asked to rate the knowledge management capabilities of the institution and the results are shown in Figure 4.9.

As can be seen in Figure 4.9, the students who participated in the training all rated the knowledge management competencies of this examiner higher, with all rating values above 4.2. This further indicates that the higher education knowledge management model incorporating the shadow team has better capabilities and in doing so can enhance its own competitiveness.

5. Conclusion. With the development of our society, knowledge is becoming increasingly important, and the importance of knowledge makes it increasingly important for enterprises to carry out knowledge management. Shadow teams can effectively integrate knowledge management with competitive intelligence, enhancing the core competitiveness of enterprises. The education and training industry, as a typical modern knowledge intensive industry, needs to strengthen its own knowledge management and enhance its competitiveness. In order to enhance the knowledge management capabilities and competitiveness of higher education knowledge management institutions, this study integrates the advantages of shadow teams, constructs a knowledge management model for higher education training enterprises, and evaluates their knowledge management capabilities. The results show that the overall performance of the knowledge management model for higher education training enterprises integrating shadow teams is superior to traditional knowledge management models, and the effec-



(a) Politics

(b) English

(c) Professional Course 1

(d) Professional Course 2

Fig. 4.8: Main Figure

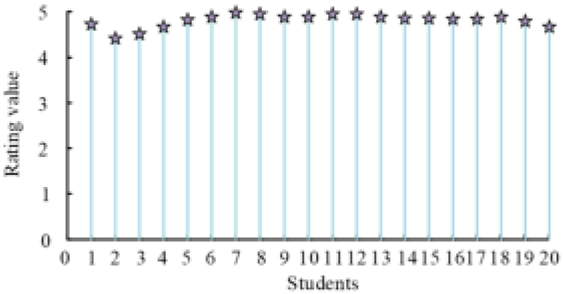


Fig. 4.9: Students' evaluation of knowledge management ability of postgraduate entrance examination institutions

tiveness of the analysis of the two models is 87.2% and 39.5% respectively when the intelligence information is 300 pieces. The former is significantly more efficient in analyzing valuable intelligence information than traditional models, and both models are more efficient in analyzing valuable intelligence information than in analyzing implicit knowledge. The students who participated in the training rated the knowledge management ability of the postgraduate entrance examination institution relatively high, with scores higher than 4.2 points. Further illustrate that the higher education knowledge management model that integrates shadow teams has better capabilities and can enhance its competitiveness. The uniqueness of this study lies in the integration of shadow teams into the process of constructing knowledge management models for higher education and training enterprises. There are still some shortcomings in this study, such as the incomplete indicator system. In the future, we will enrich and improve the evaluation index system based on the development of the industry, and construct a comprehensive and scientific evaluation index system.

REFERENCES

[1] Haryani, C. & Suryasari, S. Critical Success Factors of Knowledge Management in Higher Education Institution. *International Journal Of New Media Technology.* **7**, 111-118 (2020)

[2] Mukhtar, M., Sudarmi, S., Wahyudi, M. & Burmansah, B. The Information System Development Based on Knowledge Management in Higher Education Institution. *International Journal Of Higher Education.* **9**, 98-108 (2020)

[3] Hou, Y., Song, J., Leadership, L. & 'Green', T. Knowledge Management In Higher Education-Mediating Effect Of Trust. *Journal Of Environmental Protection And Ecology.* **21**, 2381-2388 (2020)

[4] Fayda-Kinik, F. The Role of Organizational Commitment in Knowledge Sharing Amongst Academics: An Insight into The Critical Perspectives for International Journal of Educational Management. (2022)

[5] Abdullah, H., Aldahhan, I. & Hameed, J. Building a Knowledge Management Strategies Model in Private Higher Education: An Analytical Research in a Group Xi'an Jianzhu Keji Daxue Xuebao/Journal of Xi'an University of Architecture & Technology, 2020. (2020)

[6] Enis, E. & Christopher, B. Experiential examination of higher education partnerships in the UK: a knowledge management perspective. journal of knowledge management. (2022)

[7] Horban, O. Knowledge Management as The Basis of Quality of Higher Education. *Ecological Discourse.* **33**, 45-59 (2021)

[8] Jarrahi, M., Reynolds, R. & Eshraghi, A. Personal knowledge management and enactment of personal knowledge infrastructures as shadow IT. information and Learning Sciences. (2021)

[9] Kranz, J., Fiedler, M., Seidler, A., Strunk, K. & Ixmeier, A. Unexpected Benefits from a Shadow Environmental Management Information System. mis quarterly executive. (2021)

[10] Altmay, F., Altinay, M., Dagli, G. & Altinary, Z. study of knowledge management systems processes and technology in open and distance education institutions in higher education. *Campus Wide Information Systems.* **36**, 314-321 (2019)

[11] Halberstadt, J., Timm, J., Kraus, S. & Others Skills and knowledge management in higher education: how service learning can contribute to social Journal of Knowledge Management. (2019)

[12] Burch, H., Moore, C., Burditt, J. & Patterson, M. Measuring the Nutrition Knowledge of Weight Management and Diabetes Risk in a Low Socioeconomic Population. *Topics In Clinical Nutrition.* **34**, 47-56 (2019)

[13] Information, D. & Management, K. in Higher Education Institutions: The Polish Case. *Online Information Review.* **43**, 1209-1227 (2019)

[14] Hazarika, T. & Doley, P. Perception of Library Professionals towards Knowledge Management Practices in Higher Education Institutions in Assam. *IASLIC Bulletin.* **65**, 179-189 (2020)

[15] Bangotra, P. Chahal B P S. *Knowledge Management In Indian Higher Education -a Critical Analysis.* **5**, 2190-2198 (2020)

[16] Nawaz, N., Durst, S., Hariharasudan, A. & Shamugia, Z. Knowledge Management Practices in Higher Education Institutions -A Comparative Study. *Polish Journal Of Management Studies.* **22**, 291-308 (2020)

[17] Gachino, G. & Worku, G. Learning in higher education: towards knowledge, skills and competency acquisition. *International Journal Of Educational Management.* **33**, 1746-1770 (2019)

[18] Binyamin, S., Rutter, M. & Smith, S. The moderating effect of gender and age on the students' acceptance of learning management systems in Saudi higher Knowledge Management and E-Learning. (2020)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: Nov 1, 2023
Accepted: Dec 12, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1616–1630, DOI 10.12694/scpe.v25i3.2788

INTELLIGENT CLASSROOM NOTE-TAKING APPLICATION SOFTWARE WITH HIGHER PERFORMANCE

LI ZHUANG[*]

Abstract. In the process of multimedia teaching, it is common for learners to miss out on important notes and fail to summarize and organize the course content in a timely manner, resulting in a lower learning efficiency. Aim: This paper designs an intelligent classroom note-taking application software that combines traditional note-taking with the internet, based on the needs analysis of both teachers and learners. This software utilizes the single shot multibox detector and MobileNet to build a network platform, and establishes a MySQL-based database. It has been deeply developed using various intelligent algorithms and technologies, and includes modules for learning notes, searching, recognition, and recording. Through Testin and usage testing by learners and teachers, the proposed software has been proven to effectively recognize and record learning content, ensuring the recording and expansion of teacher's knowledge points, ultimately improving students' learning efficiency. Based on the current situation of classroom note-taking, this paper explains the level of learners' awareness of classroom note-taking and the problems of note-taking in the classroom, and summarizes the design ideas and basic requirements of a classroom note-taking application. The design process of an intelligent classroom note-taking application is proposed, and the design and development of the software is further completed.

Key words: Classroom, note-taking, application, software

1. Introduction. Since the 21st century, the rapid development of 4G (the 4th generation mobile communication technology) mobile Internet technology has changed every aspect of life, and also deeply affected the way of education, teaching and learning [1, 2, 3, 4]. Universities have delayed the start of school due to the epidemic, and learners have adopted a home-based learning model, such as using the tools of smart mobile terminals for independent, personalized and inquiry-based learning. In traditional classroom learning, learners often use pencil tools to collect and organize scattered knowledge in class, and then go through these materials and notes after class to further deepen their learning [5, 6, 7, 8]. However, in terms of the change of teaching mode caused by the epidemic, once video courses and live classes, which are previously used as an aid, become the main teaching method, learners cannot quickly adapt to the new teaching method leading to a decrease in their learning conscientiousness. Moreover, due to the special nature of video courses, it is difficult for learners to summarize and summarize the courses in time and to understand and remember the relevant knowledge points in time, thus creating certain difficulties in learning. Therefore, it is necessary to design applications for class notes. How to ensure the students' consciousness and quality of learning and how to summarize after class is set as the target of the research. By analyzing the needs of teachers and students, the direction of the product needed by users is clarified. This is conducive to the full utilization of students' independent learning time and the improvement of learning efficiency.

The learning software for students on the market today is mainly based on video courses [24, 10, 12, 13]. In [9], An experience with learning software for DaimlerChrysler is reported. It aims to study experience reuse and apply insights and collected experiences to ware process improvement. To address the difficulty of meeting requirements in authoring environments in real projects, especially in large development projects, [11] proposes an unsupervised approach. This approach recognizes templates from the requirements themselves by extracting the common syntactic structure of the requirements. The effective recognition of standard and non-standard templates strongly proves the effectiveness of the proposed method for learning software. In [12], a web application called QualiTeam is developed to enhance the quality of the teaching and learning process. The program provides concrete examples while helping students to cope with challenges. This motivates students

[*]School of Electronics and Computer Engineering, Southeast University Chengxian College, Nanjing 210088, China (zhuanglizhuang@outlook.com)

to practice and clarify classroom topics in a timely manner. Their advantage is the high degree of freedom and the wide range of options available to students. But their disadvantages are also very obvious. It is difficult for students to organize and summarize the content of the video courses in time. The memory time for the knowledge points is short, so it is troublesome to try to go to the relevant videos to find the knowledge points. In addition, for the knowledge points that the teacher did not show on the lesson, students have difficulty to remember and understand them effectively. To sum up, learners need a tool to summarize what they have learned. They can recall the key points through the notes in the future review, and can expand on their existing knowledge to improve themselves.

By reviewing the relevant literature, the similar software is feasible in terms of development cost and time cycle. Whether in daily learning life or in video courses, students should summarize and summarize what they have learned in a timely manner. Therefore, this paper designs a smart classroom note-taking application that combines traditional notes with the Internet to assist daily learning. This prompts the use of students' independent learning time and enhances the recording and expansion of the teacher's classroom knowledge points.

Aiming at the problem that it is difficult to effectively combine traditional notes with video course learning, this paper designs a high-performance intelligent classroom note-taking application software, based on the theories of cognitive psychology, learning strategies and personal knowledge management. This software integrates a variety of intelligent algorithms and platforms such as TensorFlow to realize a deep development. The organic integration of classroom notes and video learning effectively improves the learning efficiency.

The innovation of the classroom note-taking application software designed in this paper is:

1. Intelligence of functions and operations. By intelligently identifying the knowledge points of documents and video courses as well as expanded knowledge, and accurately recording all the knowledge points taught by the teacher and organizing and summarizing them according to the time line, it helps learners consolidate their understanding of certain fragmented and blurred memory knowledge points, saves learners' time in watching video courses, and improves the efficiency of recording notes. Meanwhile, this software comes with a keyword search function, which has the ability to easily locate the content you want to find from many notebooks, helping learners to solve the problems of blurred memory of knowledge points and final summarization review, etc.
2. Implementation of cloud sharing function. By sharing learners' notes to the cloud, it provides a platform for other learners to learn from.
3. The diversity of usage scenarios. The designed software can be used in daily meetings, large lectures and many other scenarios. The notes and semantic-to-text recognition mode can quickly and accurately summarize and sort out the content of meetings, saving time in organizing meeting minutes later. At the same time, the note-taking and sharing modes prompt more people to learn about the meeting content and work requirements. All these effectively ensure the universality of this software in different scenarios. The developed intelligent classroom note taking application has significant advantages in terms of responsiveness and usability over the existing applications. It operates with high effectiveness and low memory utilization.

2. Basic theory. There are few national and international studies devoted to classroom note-taking theory, but the role, principles, and strategies of classroom note-taking have been addressed in psychological and learning-related theoretical studies. This paper presents the theoretical foundations of classroom note-taking in terms of cognitive psychology, learning strategies, and personal knowledge management.

2.1. Cognitive psychology. Cognitive psychology incorporates the essence of information theory, cybernetics, systems theory, and computer science, among others, and revolves around the core view of information processing [13, 14, 15, 16]. It is an approach that advocates an experimental approach and an information processing perspective to the study of human mental processes. It believes that human memory is an information processing system and that the memory structure consists of three subsystems: sensory, short term and long term memory. These different types of memory differ in terms of the retention time and capacity of information. Experiments have proved that the picture representation form of knowledge is more beneficial to learners' memory than the text form, especially the vivid pictures are more effective for memory.

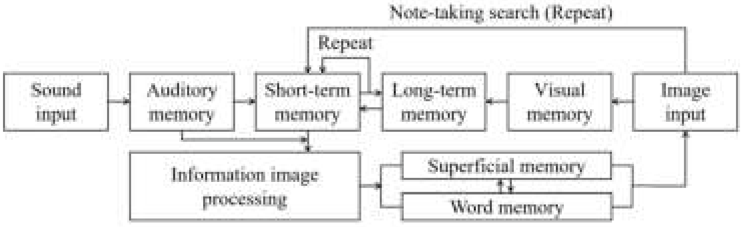


Fig. 2.1: A cognitive process model for classroom note-taking

This paper attempts to construct a complete and specific model for the cognitive processes in classroom note-taking, which includes input, information processing, comprehension, and memory of classroom information (Fig 2.1. Learners discern what the speaker says and writes on the board, Microsoft Office PowerPoint (PPT), etc. through auditory perception and visual memory. After processing the picture and sound information into meaningful short-term memory, this classroom information then entered long-term memory in the form of different representations of episodic and semantic memory. After the class, listeners subjectively and actively retrieve and retell the valid content that remains in the long term memory based on the relevant cues recorded in their notes. Such a process is similar to a kind of recording process of English listening notes.

At the same time, from the perspective of cognitive psychology and cognitive processes, learners taking notes in class not only improve their learning efficiency, but also facilitate their attention to be passively focused on the collection and capture of key information in vivid pictures. The resulting orderly cognitive chain of key content facilitates the management of future reviews. The way learners take notes and the instantaneous recording of information also directly affect the knowledge acquisition degree.

2.2. Learning strategy. Since the concept of learning strategies was introduced in 1956, a more complete concept has gradually emerged [19, 20, 21, 22]. It is a series of processes or steps that facilitate the acquisition and storage of knowledge and the use of information. Some scholars argue that it is essentially an intellectual activity or step of thinking, mainly in the extraction, analysis and encoding of information [23, 24, 25]. In this paper, the learning strategy is understood as a mode of thinking in which learners select, use and regulate learning procedures, methods, techniques, resources, etc. during the learning task. For example, the way they choose how to take notes on what they learn in class, how to review their notes from the learning task after class, etc.

Various definitions and classifications of learning strategies exist. However, the majority of scholars agree that it consists of a variety of methods and techniques that facilitate learning to occur and enhance learning outcomes [26, 27, 28]. During the classroom note taking process, learners are advised to choose the appropriate learning tools and more effective learning strategies for the situation at hand.

2.3. Personal knowledge management. Personal knowledge management is a branch of knowledge management [29, 30, 31, 32]. The main ways to divide knowledge management tools are:

1. The tools formed based on the knowledge process division are knowledge acquisition and classification, knowledge storage and management, knowledge retrieval, knowledge analysis and mining, and knowledge sharing.
2. The tools formed based on the knowledge type division are literature management tools, patent depth analysis tools, mind maps, cloud-based notes and web depth information management tools, and social networking tools. The impact of personal knowledge management on classroom note-taking will be described from three aspects: 1) the ability to classify and manage personal note knowledge in an orderly manner, especially the classification of key contents; 2) the increase of more ways to learn and process information; 3) the ease of finding and retrieving relevant knowledge efficiently and quickly.

3. Core technology. There is a wide variety of foreign note-taking tools, and many of them are powerful with the support of hardware. Take eClass, E-notes and myBase as examples, they can realize various forms

Table 3.1: Comparisons of existing note-taking tools

Tool	**Basic function**	**Photo function**	**Circle function**
eClass	Video and audio recording, communication and sharing within the local area network	Yes	No
E-notes	Lecture notes with extensive teaching contents	No	
myBase	Information access, editing, viewing, indexing, searching and sharing	No	
OneNote	Support built-in search function for picture and audio repositories, and the ability to record notes during the recording process	Yes	
Evernote	A note editing platform that can save complete web pages and store text and photos	Yes	
Wiz	Knowledge manager, sync tool, editor, viewer, web capture tool and document import tool	No	
Youdao Cloud Note	Text, photo, album and handwritten notes	Yes	
DaubNote	Multi-platform login, photo recognition text, social sharing and book sweep	Yes	

of text input, voice recording, video and audio recording, screenshot, document upload and download, sharing resources and other functions [33, 34, 35, 36, 37]. However, their specific operations and applications are difficult to meet the efficacy of note taking, and there is no software that can go through the cell phone mobile terminal to record notes with the camera and edit photo notes at any time. Although domestic scholars are late to study classroom notes, there are a considerable number of note-taking software, such as Wiz, Youdao Cloud Note, and DaubNote [38, 39, 40]. They are mostly used for data storage and online socialization.

3.1. Tool comparison. To design a tool suitable for taking notes in the classroom, the functions of the mainstream existing tools are compared (Table 3.1).

The functions of these software in Table 1 vary and are well referenced. Most of the software meets the basic needs of students to record, manage and share their notes. However, there are few course notes software that are simple to use, highly operational, and can be used on mobile terminals. The following details the technology of the classroom note-taking application software designed in this paper.

3.2. Technique details. Five core technologies are used in this software.

1. TensorFlow is created by developers and engineers on the Google Brain team. It has been used for research in machine learning and deep neural networks and has been used in numerous fields [41].
2. OpenCV is a cross-platform computer vision and machine learning software library that runs on most operating systems. It provides interfaces to most languages such as Python, Ruby, MATLAB, etc., and implements many common algorithms in image processing and computer vision [42].
3. OCR (Optical character recognition) is a process of acquiring text and layout information. Baidu's OCR service relies on excellent deep learning algorithms and massive quality data, with an accuracy rate of 99% for keyword segments [43].
4. Ali real-time speech recognition has the ability to do real-time recognition of audio streams of unlimited duration. It has built-in intelligent sentence breaking, which provides the start and end time of each sentence. It can also be used in scenarios such as real-time live video captioning, real-time conference recording, real-time courtroom trial recording, and intelligent voice assistant [44].
5. Luban is an image compression tool for Android. It first sends nearly 100 images of different resolutions, and then compares the original and compressed images. The resulting reverse-derived compression algorithm has become the core algorithm of Luban [45].

4. System design. First, the theoretical basis for the development of the intelligent classroom note-taking application is laid out from the theories related to classroom note-taking. By analyzing the requirements of the classroom note-taking application software, we find the basis for software development and design. Then, we

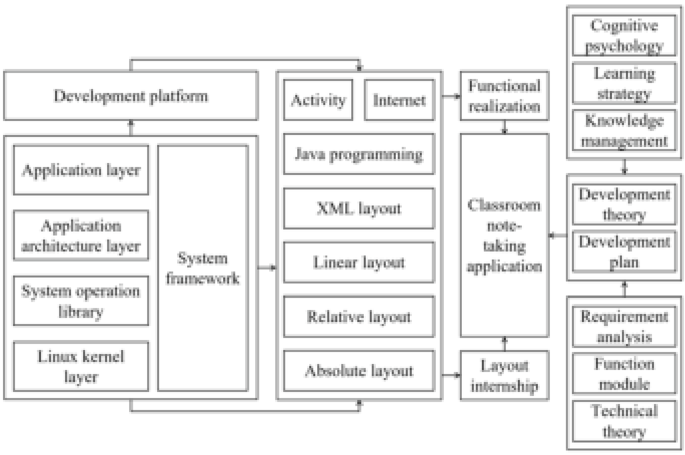


Fig. 4.1: System structure of the designed software

analyze the technology required to implement the layout and functionality of this software, and describe the design and development process of this software. Finally, testing and modification are performed.

4.1. System structure. Figure 4.1 illustrates the system structure of the designed application software.

The implementation steps of the proposed system are as follows. Step 1: When the software is opened, the screen displays modules such as My Home, Course Library, Recognize, Search, and Study Notes. Step 2: By clicking My Home, users can see My Uploads, Personal Center, Settings, About Us and other functional modules. All notes that have been uploaded can be seen by clicking on My Uploads. Step 3: Click Recognize to create a new note to compose a note. It records the class in real time and supports voice to text conversion. Once the note is built, this system also recognizes the title of the note, core concepts, and related knowledge points. Step 4: Click on Course Library to select notes from different courses for further sharing of notes and learning documents. Step 5: Click on Study Notes to view various categories of study notes for learning. During the learning process, a share anytime anywhere is implemented for the better notes. Step 6: Learners can use Search to make inquiries and possibly share via keywords.

4.2. Core algorithm.

4.2.1. Target detection. SSD (Single shot multiBox detector) is a model that directly generates class probabilities and location coordinates of objects without generating candidate regions [46, 47]. Using SSD, the final detection result can be obtained after only a single detection. MobileNet is a network structure used in the SSD process for feature extraction. Combining SSD with MobileNet not only preserves the network structure of the original SSD model, but also ensures the accuracy of the model. Based on this, this paper transforms the original large number of redundant parameters in the model into small parameters. This not only reduces the amount of network computation, but also reduces the hardware resource consumption and effectively improves the model performance.

4.2.2. FAST feature point extraction. FAST (Features from accelerated segment test) is a simple and fast feature extraction algorithm, which is widely used in real-time detection because of its fast computing speed and good feature extraction effect [48, 49, 50]. The fixed threshold value in traditional FAST leads to poor robustness. Meanwhile, it is difficult for FAST to achieve better detection results for images with uneven illumination and different local contrasts. Adaptive thresholding is introduced so that the threshold value changes with the local contrast of the image to improve the FAST-based feature point extraction effects.

4.2.3. BRIEF feature description. BRIEF (Binary robust independent elementary features) is an algorithm for computing feature descriptors, which serves to alias the feature points [51, 52]. So, feature points

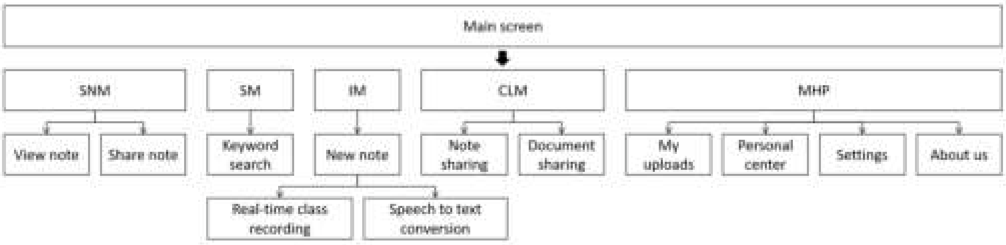


Fig. 4.2: System modules and functions

need to be extracted beforehand. Here, the image feature points are extracted beforehand using FAST, and then the binary descriptors of the features are created in the feature point neighborhood using BRIEF. Since the operation of BRIEF is based on the specific pixel values of the image, it is more sensitive to noise and does not carry rotation and scale invariance. ORB (ORiented Brief) is an algorithm after improving the BRIEF algorithm. It does not use direct comparison of pixel points, but chooses a small patch centered on a pixel as the comparison object, obviously improving the noise immunity.

4.2.4. Laplace operator. The Laplace operator is one of the simplest isotropic differential operators with rotational invariance and is particularly suitable for highlighting isolated points, isolated lines or line endpoints in an image [53]. Like the gradient operator, the Laplace operator also enhances the noise in the image. When using this operator for edge detection, the image needs to be smoothed first.

4.3. Database design. The server-side database uses MySQL, which is the most popular relational database management system (RDBMS) [54]. Also, it is one of the best RDBMS applications in terms of WEB applications. And the Android side database uses SQLite, which is the most widely deployed SQL database engine in the world. It has the advantages of miniaturization and full functionality, making it the top solution for Android as a built-in database.

4.3.1. User table. User represents the user table, which is used to store user data, while id is a auto-increment primary key that identifies each user. Mobile and email represent the user's cell phone number and email address, respectively, while token is used to store the user's identity token.

4.3.2. Note table. Note represents the note table, which is used to store each note. Also, type represents the type of the note, its possible values are speech, ppt, knowledge, etc. After capturing the image, filename represents the image file name.

4.4. System module. As shown in Figure 4.2, this software contains a study note module (SNM), a search module (SM), an identification module (IM), a course library module (CLM), and a my home page (MHP).

4.4.1. SNM details. The knowledge points in IM are included in SNM, where users can view the content of previously recorded notes as well as the key points. SNM supports functions such as viewing large images and marking key points to reduce the burden of recording notes for students. Students can also share their notes, so that more students can participate in the discussion of the notes. Fig 4.3 illustrates the flow chart of SNM.

The section called My Notes is located on the first page of the software when it is opened. The first image in each notebook is set as a preview. Users can find the notebook they want to view by previewing when the title was created. The viewed notes display the PPT and the corresponding notes. Clicking on a notebook in My Notes will take you to the module called Notes View Editor, which is used to browse and edit the notes. Everything in the notebook expands on the timeline. Clicking on 'Show' allows you to fully replicate the class and automatically scroll through the timeline while listening to the instructor. Repeated playback of voice clips is allowed. Also, the ability to view larger images is provided for PPT. A separated structure is used

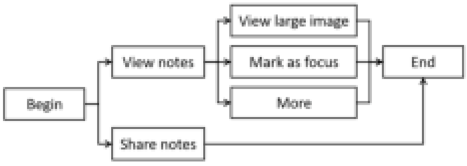


Fig. 4.3: A flow chart of SNM.

to design the editor's interface. Specifically, above and below the PPT are the recording function and notes, respectively. Finally, each note can be annotated on both the recording and viewing editing pages to ensure that no sentiment is missed. Tapping on the annotation function pops up a separate dialog box for entering the annotation content. The annotations are displayed at the bottom of the PPT.

4.4.2. SM details. In SM, the identified content in the notebook can be retrieved by typing in keywords. Clicking on an image will take you to that note to find the content. This is a convenient and quick way to easily retrieve existing knowledge and deepen students' understanding and organization of existing knowledge.

The search interface is simple and clear. After entering the keywords of knowledge points in the text box, the system matches the text in the annotations and displays the PPT and the corresponding notes. Relevant results are displayed below the text box. At the same time, the searched matching content is marked in red.

4.4.3. IM details. After entering IM, the user first has to set the permissions for camera, microphone, storage space and hover window. After the permission is turned on, the camera starts to recognize and shoot the course. Place the PPT or board in the designated recognition area, and the proposed software will recognize the content and give the analysis. The recognized content will be automatically recorded in the notebook. When switching the PPT, the content will be automatically recorded to prevent confusion of knowledge points during later review. Clicking on images to view text, annotate, highlight, and delete knowledge points is allowed. The search function is used to quickly retrieve knowledge points and locate notes. At the same time, the real-time lecture recording function is embedded with a speech-to-text function, which can record the lecture content and convert it into text for users' reference. IM is useful not only for identifying and analyzing knowledge points, but also for summarizing and summarizing what was said in class in chronological order. If students look at a single knowledge point, they may not know how that knowledge point is derived. However, with the class content recorded by IM, it is possible to find the required knowledge points in the timeline. The content before and after is connected by this recording method, which deepens the students' understanding of the content and knowledge.

Using the recording module in IM, the newly captured content will be displayed in the interface and users can go through it freely. Clicking the blinking rec button prompts a quick setup. In this interface, a quick press of the volume button twice triggers a forced capture. During the recording process, users can perform operations such as converting text, annotating, marking as key, deleting, etc. on the notes. IM uses an interface distribution with PPT on the left and annotation on the right.

4.4.4. CLM and MHP details. Users can share and discuss their study notes and documents in CLM. Each study note has been added with a comment area for mutual communication and sharing of review materials. CLM supports a variety of document formats such as announcements, final papers and exams, and supports local downloads. MHP contains My Uploads, Personal Center, Settings, About Us, etc. The proposed software supports third party login. Once logged in, you can manage your personal information in MHP. The interface of CLM implements the functions of PPT previewing and provides a library of notes and discussions for easy communication with each other.

5. Results and analyses. This section first introduces the system environment, followed by qualitative and quantitative testing of the proposed software, and finally further testing based on specific test examples.

5.1. System environment. Table 5.1 records the system environment information of this software.

Table 5.1: Information about the system environment

Type	Information	Type	Information
Operating system	Ubuntu 18.04 LTS	Server software	Nginx 1.16.1
Development tool	Android Studio v3.3.2	MySQL version	5.5.62
Development language	Java, Html, JavaScript (Client), Python (Server)	Redis version	5.0.5
Cloud server platform	Alibaba Cloud's lightweight application server	Python version	3.7
Server operating environment	Ubuntu 18.04.2 LTS	Client running platform	Android 5.1 and above

Table 5.2: Results of performance tests

Test items	Specific contents	Test results
Compatibility	Install	Pass
	Start	Pass
	Traversal	Pass
	Monkey	Pass
	Script	–
	Uninstall	Pass
Performance	Installation time (s)	3.149
	Start-up time (s)	0.704
	Traversal time (min)	1.017
	Monkey (min)	1.016
	Script run (min)	–
	Unloading time (s)	0.547
	AV of OR (%)	7.17
	PV of OR (%)	35.7
	AV of MU (MB)	185.21
	PV of MU (MB)	256.01
	Total flow consumption (B)	1.1501797E7
	Uplink flow (MB)	0.94
	Downlink flow (MB)	10.03
	Maximum FPS	89.0
	AV of FPS	82.93
	Minimum FPS	44.0
	Average battery temperature (℃)	27.0

5.2. Performance test. The designed application software is tested using the testing tool Emmagee (Table 5.2) [55]. The average and peak values, and frame per second are defined as average value (AV), peak value (PV), and FPS respectively. The occupation ratio (OR) of the central processing unit (CPU) of less than 8%, and the memory usage (MU) of less than 260 MB, strongly indicate the competitive performance of this software.

Based on a live classroom, the functionality of the software is tested. The main task is to test the operation of the main functions. The relevant results are shown in Table 5.3.

A comparison with the mainstream tools in Table 3.1 is shown in Table 5.4, for a same test task.

Table 5.4 shows that the proposed software has the lowest AV of OR (7.02%) and PV of OR (36.21%), and the highest AV of FPS (83.35Hz). This software's AV of OR is 8.25% lower than the worst-performing E-notes (15.27%). This software's PV of OR is 21.9% lower than the worst-performing E-notes (58.11%). This software's AV of FPS is 38.45Hz higher than the worst-performing eClass (44.9Hz). The proposed software shows competitive advantages in both OR and FPS.

Table 5.3: Test results of main functions

Test modules	Test contents	Test results
Photo-taking function	Can normally call the photo, and can quickly focus and then take pictures	Normal use
Circle function	Can circle on the instant photo to emphasize the corresponding content	
Graffiti function	Can doodle on instant photos to complete custom graphic circles	
Straight line function	Can draw straight lines on instant photos	
Recording function	Can achieve instant recording	
Study digest	Can mark learning experiences and recorded contents	

Table 5.4: Comparisons between multiple note-taking tools

Tool	AV of OR (%)	PV of OR (%)	AV of FPS (Hz)
eClass	11.3	53.02	44.9
E-notes	15.27	58.11	52.28
Wiz	8.46	40.37	69.95
Proposed software	7.02	36.21	83.35

5.3. Case-based test. Based on the system and module testing done earlier, it is necessary to further design test cases to test various functions of the software. First, the proposed system is tested without turning on the permissions. Results show that the system is unbootable. This ensures the controllability of the system. Then, test cases are set up in major dimensions such as system login, recording video and identification to test the system in all aspects. Relevant settings are shown in Table 5.5.

In Table 5.5, this system is able to fulfill multiple devices logging into the same account. A number of functions, such as creating new notes, writing notes, real-time classroom recording, converting voice to text, recognizing knowledge points of notes, viewing notes, sharing notes, etc., run normally. Video recording and document recognition meet the use of classroom notes and have a certain accuracy. On the basis of the above test results, this system is applicable to a wide range of scenarios of teacher teaching and student learning.

5.4. Validation and Testing. The purpose of this section is to evaluate the usability of the proposed software. The functional architecture, interaction design, navigation design, visual interface design, iconography and other elements of the software are investigated and evaluated for user satisfaction. Also, improvements are made to the problems found.

Users use the designed software to complete nine tasks (Table 5.6) through the mobile phone interface. The purpose and content of the test are indicated to the users before the formal test, and the users' usage behavior is observed and problems encountered in the operation are recorded during the test. After the test is completed, users are interviewed in detail and asked to fill out an evaluation questionnaire. The questionnaire uses a 7-point Likert scale, with 7 levels from very dissatisfied to very satisfied.

A total of 8 users are selected for this test, including 5 male and 3 female (4 from both Arts and Science categories). All of them are aged 20-26 years old, and are experienced in mobile internet and mobile app usage.

The software performs well in terms of usability. In Table 5.7, the software has the highest level of satisfaction in terms of functional design (5.85) and learning and mastery (5.95). This indicates that the design is reasonable and simple, and is in line with the user's perception. An additional ease of learning and mastery indicates that users recognize the usefulness and usability of the software. However, the satisfaction levels for interaction, usage performance, visual design and icon design are relatively low. This is due to the focus of the software on functional design and architecture and less on visualization. Overall user satisfaction is better (5.825).

6. Discussion. This study designs and develops a software that meets the needs of college students for classroom note-taking, which has distinct advantages over similar products. The shortcomings are:

1. The application effect of the proposed software has not been tested extensively. It is planned to promote

Table 5.5: Case test results

No.	**Test items**	**Test contents**	**Prerequisites**	**Test data**	**Test steps**	**Expected output**	Actual results	Test
1	No permissions on	Application for camera, microphone , storage space, hover window, etc. permissions	Normal page display and network connection	No permissions on	1) Click to record 2) Permission Request 3) No Permission on	Unable to access the recording screen and reapply for permission	Failure to enter the recording interface and re-accept permission requests	Pass
2	Recording video and recognition	Recognition accuracy in case of clear recording contents	Normal network connection and all permission requests have been granted	Example PPT files	Aim the sample PPT at the camera hover window to ensure a clear shot	Display a clear picture in sample PPT, and the recognized content is consistent with the content shown in the picture	Clear picture display in sample PPT, with functions such as annotation mark deletion of pictures and keeping the content consistent	
3	Video recording accuracy test	Recognition effect in case of blurred recording contents	Normal network connection and all permission requests have been granted	Example PPT files	Misalign the sample PPT with the camera hover window and take a shaky shot with the device to create a blurring effect	Display ‘Blurred picture’	Display ‘Blurred picture’	
4	Recognition based on different PPT	Recognize different PPT to test automatic recording	Normal network connection and all permission requests have been granted	Different example PPT files	1) Click record to enter the recording interface 2) Create a new notebook 3) Recognize the difference PPT	Switch different PPT files, accompanied by automatic recording	Clear display of sample PPT pictures, consistent contents, and PPT files are automatically recorded when switching	
5	Same account login for multiple devices	Test whether data can be saved properly when multiple devices log into the same account	Normal network connection and all permission requests have been granted	Different devices and same account	1) Click ‘My’ with a different device to enter the login screen 2) Login to the same account and use the recording function	Different devices can properly save data changes under the same account	The login status of all devices can be saved normally, and the software can cope well with simultaneous login of multiple devices	
6	Audio test	Click on an audio while playing another audio	Normal page display and network connection	Audio	1) Click on the notebook 2) Click on the audio play button under a text to play that audio 3) Click on the audio play button under another text to play that audio	The previous audio playback is paused, and the next audio file is played, without overlapping between audio playback	The previous audio playback is paused, and the next audio file is played, without overlapping between audio playback	

Table 5.6: Description of tasks

Type	**Description**
Software guidance	Navigating the software guide page
Software login	Navigating the software guide page
New notes	Add/remove notes as needed
View notes	View notes already taken
Take down notes	Add new note content (try multiple forms of adding)
Search notes	Search notes by keywords
Share notes	Share notes with classmates or course library
Real-time video recording	Turn on the live video recording function and test the effect at different times
Speech to text	Turn on voice recognition and test the effect of different voices

Table 5.7: Satisfaction results

No.	**Form of question**	**Survey purpose**	**Average satisfaction value**
1	Whether the software features are useful or not	Functional design satisfaction	5.85
2	Whether the software is easy to use	Interaction satisfaction	5.4
3	Software utilization error rate	Usage performance satisfaction	5.525
4	Whether the software is easy to learn and master	Learning and mastery satisfaction	5.95
5	Whether the software interface is aesthetically pleasing	Visual design satisfaction	4.875
6	Whether the software icon is easily recognizable	Icon design satisfaction	5.5
7	Overall software experience	Overall satisfaction	5.825

this software among the college student group in order to further improve the software by collecting feedback [56, 57].

2. The functions of the software are not comprehensive enough, such as the lack of sharing, classroom video recording and other personalized functions. We hope to add more functions to keep up with the times [58, 59, 60].
3. The aesthetics of the software needs to be improved. In addition, possible theoretical and methodological deficiencies need to be addressed [61, 62].
4. Privacy protection and data security need to be strengthened [63, 64, 65, 66].
   (a) Strengthening laws and regulations and privacy protection policies: Develop software use systems based on laws and regulations and privacy protection policies established by the government and relevant organizations. These systems should cover regulations for data collection, use, sharing and storage to ensure that personal privacy is adequately protected;
   (b) Anonymization and encryption: For sensitive personal data, the use of anonymization and encryption is an effective means of protection. By removing personally identifiable information or encrypting personal data, the risk of privacy leakage can be reduced.
   (c) Data access and usage authority control: the artificial intelligence system should implement a strict access and usage authority control mechanism. Only authorized personnel can access and use specific data to ensure data security and privacy.
   (d) Security auditing and monitoring: A security auditing and monitoring mechanism is established to monitor and record data access and operations in the artificial intelligence system in real time, so that potential security risks can be discovered and responded to in a timely manner.
   (e) Transparency and Interpretability: The increased transparency and interpretability of artificial intelligence algorithms and models prompts review and validation to minimize the risk of data misuse or system attack.

Data show that in January 2023, the number of active people of note products on the whole network was 43.199 million, and the number of active people of note products has reached ten million, with a huge user scale and a large market potential [67, 68]. With a per capita usage time of 0.7 hours, note-taking products have quietly occupied daily usage time and become part of life and work. With the increase in the number of knowledge workers, the demand for domestic efficiency office applications will continue to expand.

The note-taking product industry has entered the market maturity period, with a large overall user volume and an objective number of active people. However, as a tool-type application, the product coverage scene is small. This industry will generally meet the user growth difficulties, and it is difficult to see explosive growth in the short term [69]. The next step is to make good product operation and optimization, serve some user groups well, and improve the brand effect and product recognition. The designed software has a simple style, easy operation, low cost for users to get started, and good overall performance. But in terms of function and experience, such as the editing of notes is not rich and perfect. More polishing is needed in the details. In the future, user needs will become more and more diverse and personalized. While meeting the needs, the operation should also be simplified.

With the implementation of informatization teaching, students' knowledge and skills have been accumulated, and their information and professional qualities have been greatly improved [70, 71]. The amount of information in the classroom has exceeded the traditional mode in the past due to the variety of ways of displaying knowledge resources, quick access, easy sharing, fast updating, and huge quantity. When students are confronted with this information, sometimes it is too late for them to digest it. Therefore, convenient and efficient knowledge access poses a challenge to students' personal knowledge management. Classroom note-taking application is a personalized learning tool for students, and it is also applied more in students' learning. Students can study and think independently without the guidance of the teacher and complete their learning tasks skillfully and efficiently with the help of the application software [72, 73, 74].

7. Conclusion. Based on the current situation of classroom note-taking, this paper explains the level of learners' awareness of classroom note-taking and the problems of note-taking in the classroom, and summarizes the design ideas and basic requirements of a classroom note-taking application. The design process of an intelligent classroom note-taking application is proposed, and the design and development of the software is further completed. Specific achievements are listed below.

1. Theories related to classroom note-taking are reviewed, such as learning strategies, cognitive psychology, and personal knowledge management. The characteristics and habits of learners taking notes in the classroom are analyzed, as well as learners' confusion about note-taking and the need for a classroom note-taking application. Based on this, the design idea of the smart classroom note-taking application is developed, which provides a valuable reference for similar works.
2. Based on the design idea and basic requirements of the intelligent classroom note-taking application software, the design process and detailed description of the software are proposed, combined with the questionnaire and theory.
3. By combining the artificial intelligence and traditional computer vision, the proposed software can record and sort through the lesson content by time, enhancing systematization and efficiency. In addition to local functionality, the software offers a course library function, which allows learners to easily access all shared notes and documents via the web.

Funding. This research was funded by the National Association for Computer Basic Education in Higher Education-Computer Basic Education Teaching Research Project of China (Grant 2022-AFCEC-027), and the Southeast University Basic Research Fund Project (Grant No. 2242022K30056).

REFERENCES

[1] Tanwar, G., Singh, G. & Others Multimedia Streaming Technology n 4G Mobile Communication Systems. *International Journal On Computer Science And Engineering.* **2** pp. 3 (2010)

[2] Khan, N., Ray, R., Zhang, S. & Others Influence of mobile phone and internet technology on income of rural farmers: Evidence from Khyber Pakhtunkhwa Province, Pakistan. *Technology In Society.* **68** (2022)

[3] Gao, Y., Shang, T. & Ma., L. Mobile internet big data technology-based echo loss measurement method of optical communication system. *Computers And Electrical Engineering.* **101** (2022)

[4] Pandey, D., Mukherjee, S., Das, G. & Others Improving base-of-the-pyramid consumer welfare through mobile technology services. *Journal Of Services Marketing, No.* **2** (2022)

[5] Oh, P., Ha, H. & Yoo, Y. Epistemological messages in a modelingbased elementary science classroom compared with a traditional classroom. *Science Education.* **106** (2022)

[6] Xing, X. & Saghaian, S. Learning Outcomes of a Hybrid Online Virtual Classroom and In-Person Traditional Classroom during the COVID-19 Pandemic. *Sustainability.* **14** (2022)

[7] Ferreira, D., Sentanin, F., Parra, K. & Others Implementation of Inquiry-Based Science in the Classroom and Its Repercussion on the Motivation to Learn Chemistry. *Journal Of Chemical Education, No.* **2** (2022)

[8] Grifenhagen, J. & Barnes, E. Reimagining Discourse in the Classroom. *The Reading Teacher.* **75**, 739-748 (2022)

[9] Schneider, K., Hunnius, J. & Basili, V. Experience in implementing a learning software organization. *IEEE Software.* **19**, 46-49 (2002)

[10] Nielson, K. Self-study with language learning software in the workplace: What happens?. *Language Learning & Technology.* **15**, 110-129 (2011)

[11] Sonbol, R., Rebdawi, G. & Ghneim, N. Learning software requirements syntax: An unsupervised approach to recognize templates. (Knowledge-based systems,2022)

[12] Cervantes-Ojeda, J. & Gomez-Fuentes, M. QualiTeam: A Support Tool When Learning Software Quality and Testing Concepts. *Software Engineering And Applications.* **15** pp. 1 (2022)

[13] Barbosa, M. Using Blended Project-Based Learning to Teach Project Management to Software Engineering Students. *International Journal Of Mobile And Blended Learning (IJMBL).* **14** (2022)

[14] Gong, X. & Li., X. Human–Robot Interactive Communication and Cognitive Psychology Intelligent Decision System Based on Artificial Intelligence-Case Study. (International Journal of Humanoid Robotics,2022)

[15] Shu, Y. & Gao, S. A Critique of the Philosophical Presuppositions of Cognitive Psychology. *Social Sciences In China (English Edition).* **43** pp. 4 (2022)

[16] Ratnayake, S. It's Been Utility All Along: An Alternate Understanding of Cognitive Behavioral Therapy and The Depressive Realism Hypothesis. *Philosophy, Psychiatry, & Psychology: PPP, No.* **2** (2022)

[17] Nielsen, N. & Berntsen, D. How posttraumatic stress disorder symptoms affect memory for new events and their "hotspots" over a long delay. *Applied Cognitive Psychology.* **36**, 59-68 (2022)

[18] Oliveira, M. & Arntzen, E. Meaningful Events in Cognitive and Behavioral Psychology Research Approaches: A 6-Year Literature Review. *Revista Brasileira De Analise Do Comportamento, No.* **1** (2021)

[19] Wu, Q. & Saif, M. Robust Fault Diagnosis of a Satellite System Using a Learning Strategy and Second Order Sliding Mode Observer. *IEEE Systems Journal.* **4**, 112-121 (2010)

[20] Wu., K. The relationship between language learners' anxiety and learning strategy in the CLT classrooms. *International Education Studies.* **3**, 174-191 (2010)

[21] Pudelko, B., Young, M., Vincent-Lamarre, P. & Others Mapping as a learning strategy in health professions education: a critical analysis. (Medical Education,2012)

[22] Colin, C. Developing a corporate learning strategy: creating intrapreneurs. (Strategic Change,2000)

[23] Chang, S. & Ley, K. A Learning Strategy to Compensate for Cognitive Overload in Online Learning: Learner Use of Printed Online Materials. *Journal Of Interactive Online Learning.* **5** pp. 1 (2006)

[24] Schellings, G. Applying learning strategy questionnaires: problems and possibilities. *Metacognition & Learning.* **6**, 91-109 (2011)

[25] Xu, X., Tang, Y., Li, J. & Others Dynamic multi-swarm particle swarm optimizer with cooperative learning strategy. *Applied Soft Computing.* **29** pp. 169-183 (2015)

[26] Donker, A., Boer, H., Kostons, D. & Others Effectiveness of learning strategy instruction on academic performance: A meta-analysis. *Educational Research Review.* **11** pp. 1-26 (2014)

[27] Liu, Z., Lin, C., Jian, P. & Others The Dynamics of Motivation and Learning Strategy in a Creativity-Supporting Learning Environment in Higher Education. *Turkish Online Journal Of Educational Technology.* **11**, 172-180 (2012)

[28] Newton, J., Tsarenko, Y., Ferraro, C. & Others Environmental concern and environmental purchase intentions: The mediating role of learning strategy. *Journal Of Business Research.* **68**, 1974-1981 (2015)

[29] Swigon, M. Personal knowledge and information management-conception and exemplification. (Sage Publications,2013)

[30] Liu, C., Wang, J. & Lin, C. The concepts of big data applied in personal knowledge management. (Journal of Knowledge Management,2017)

[31] Li, Q. & He., H. Application of Mind Map in Personal Knowledge Management of Postgraduates. (China Educational Technology & Equipment,2013)

[32] Mittelmann, A. Personal Knowledge Management as Basis for Successful Organizational Knowledge Management in the Digital Age. *Procedia Computer Science.* **99** pp. 117-124 (2016)

[33] Brotherton, J. & Abowd, G. Lessons learned from eClass: Assessing automated capture and access in the classroom. *ACM Transactions On Computer-Human Interaction.* **11**, 121-155 (2004)

[34] Wirth, M. E-notes: using electronic lecture notes to support active learning in computer science. *Acm Sigcse Bulletin.* **35**, 57-60 (2003)

[35] Stricker, A. & Clemons, L. Simulation gaming for education in MyBase: the future of air force education and training with virtual world learning. *Proceedings Of The 2009 Spring Simulation Multiconference.* (2009)

[36] Eskritt, M. & Mcleod, K. Children's note taking as a mnemonic tool. *Journal Of Experimental Child Psychology.* **101**, 52-74 (2008)

[37] Dueholm, S., Rasmussen, J., Poulsen, R. & Others Anatomy note-taking software supporting different learning modalities. (Iated,2011)

[38] Balzer, C., Oktavian, R., Zandi, M. & Others Wiz: A Web-Based Tool for Interactive Visualization of Big Data-ScienceDirect. (SSRN Electronic Journal,2020)

[39] Yuan, R., Liu, B. & Wang, D. Research on the Application of YouDao AI cloud API in Library Information Service. (Library,2019)

[40] Ying, L. Cloud-based Solutions to Personal Document Management. (Information Research,2016)

[41] Marijn, H. & Anneroos, P. Machine Learning and the Platformization of the Military: A Study of Google's Machine Learning Platform TensorFlow. *International Political Sociology, No.* **2** (2022)

[42] Karra, V., Verma, A., Guzel, A. & Others Quantification of Alpha Lath in Ti-6Al-4V using OpenCV. *Materials Characterization.* **186** (2022)

[43] Andrea, S., Hugo, H., Raluca, T. & Others Reading in the mist: high-quality optical character recognition based on freely available early modern digitized books. *Digital Scholarship In The Humanities, No.* **4** (2022)

[44] Hussein, A., Watanabe, S. & Ali, A. Arabic speech recognition by end-to-end, modular systems and human. (Computer Speech & Language,2022)

[45] Chen, W. & Li., Q. Application of Luban Image Compression Algorithm in Instant Chat APP. *International Electronic Elements.* **27** pp. 6 (2019)

[46] Tang, Y., Ip, A. & Li., W. Artificial intelligence approach for aerospace defect detection using single-shot multibox detector network in phased array ultrasonic-ScienceDirect. (IoT,2022)

[47] Sogabe, M., Ito, N., Miyazaki, T. & Others Detection of Instruments Inserted into Eye in Cataract Surgery Using Single-shot Multibox Detector. (An International Journal on Sensor Technology,2022)

[48] Bhat, A. Makeup Invariant Face Recognition using Features from Accelerated Segment Test and Eigen Vectors. *International Journal Of Image & Graphics.* **17** pp. 1 (2017)

[49] Zhou, S., Yan, R., Li, J. & Others A Brain-inspired SLAM System Based on ORB Features. (International Journal of Automation,2017)

[50] Sumiharto, R., Putra, R. & Demetouw, S. Methods for Determining Nitrogen, Phosphorus, and Potassium (NPK) Nutrient Content Using Features from Accelerated Segment Test (FAST). *International Journal Of Computer Science And Software Engineering.* **9**, 1-5 (2020)

[51] Calonder, M., Lepetit, V., Strecha, C. & Others BRIEF: Binary Robust Independent Elementary Features. (Springer,2010)

[52] Wang, C. Real time non-rigid surface detection based on binary robust independent elementary features. *Journal Of Applied Research And Technology.* **13**, 297-304 (2015)

[53] Biccari, U. Internal control for a non-local Schr?dinger equation involving the fractional Laplace operator. *Evolution Equations And Control Theory.* **11**, 301-324 (2022)

[54] Application, D. & System, M. of Scientific Research Projects Based on PHP and MySQL. (Journal of Interconnection Networks,2022)

[55] Wei, L., Tan, K. & Ding, C. The Optimized Research on Android Performance Test Tool Emmagee. (Electronic Test,2016)

[56] Maldar, P., Mane, A., Nikam, S., Dhas, S. & Moholkar, A. Spray deposited Cu2CoSnS4 thin films for photovoltaic application: Effect of film thickness. *Thin Solid Films.* **709** (2020)

[57] Mattheos, N., Nattestad, A., Christersson, C. & Others The effects of an interactive software application on the self-assessment ability of dental students. *European Journal Of Dental Education.* **8**, 97-104 (2015)

[58] Granger, C. CoreTechnologie adds new nesting functions to its 4D_Additive software. (Machinery Market,2022)

[59] Liu, M., Qi, X. & Pan, H. Multifractal analysis of the software evolution in software networks. *Chinese Physics B.* **31** pp. 3 (2022)

[60] Li, M. & Yin, P. Model2SAS: software for small-angle scattering data calculation from custom shapes. *Journal Of Applied Crystallography, No.* **3** (2022)

[61] Kozbelt, A., Dexter, S., Dolese, M. & Others The aesthetics of software code: A quantitative exploration. *Psychology Of Aesthetics Creativity & The Arts.* **6**, 57-65 (2012)

[62] Barber, Z., Gomez, K., Williams, E. & Others Objective vs subjective assessment of breast aesthetics: A comparison of BCCT.core software with patient satisfaction. *European Journal Of Surgical Oncology.* **43** pp. 5 (2017)

[63] Dhekale, A. & Jadhav, R. A study on Privacy Protection and Data Security in Cloud Computing. *JETIR, No.* **6** (2021)

[64] Yang, X., Xing, H., Su, X. & Others Entropy-based thunderstorm imaging system with real-time prediction and early warning. (IEEE Transactions on Instrumentation,2022)

[65] Mohammadi, A. & Hamidi, H. Analyzing Tools and Algorithms for Privacy Protection and Data Security in Social Networks. *International Journal Of Engineering, Transactions B: Applications.* **31**, 1267-1273 (2018)

[66] Yang, X., Xing, H., Ji, X. & Others Multifeature Fusion-Based Thunderstorm Prediction System With Switchable Patterns. *IEEE Sensors Journal.* **23**, 18461-18476 (2023)

[67] Wang, J. & And, X. and Countermeasures of Online Learning for College Students Based on Learning Pass APP. *Western China Quality Education.* **9** pp. 21 (2023)

[68] Shi, W. The construction of intelligent laboratory classroom in junior high school biology. (Research on Curriculum,2023)

[69] Wei, W. & Zixin, W. An Improved QFD Method for Rapid Response to Customer Requirements in Product Optimization Design. (Procedia CIRP,2023)

[70] Guo, R., Yang, C., Li, M. & Others The Effect, Problems and Suggestions of Informatization Teaching in Higher Vocational Colleges Under the Background of Digital Transformation: A Survey from 226 Higher Vocational Colleges in 28 Provinces. *China Higher Education Research.* **39**, 101-108 (2023)

[71] Song, H. & Zhang, T. Research and Application of Higher Vocational Education Informatization in Medical Nutrition and Health Course Teaching. *Education Science, No.* **5** pp. 177-180 (2022)

[72] Alamri, M., Jhanjhi, N. & Humayun, M. Digital curriculum importance for new era education. *Employing Recent Technologies*

For Improved Digital Governance. pp. 1-18 (2020)

[73] Khalil, M., Humayun, M. & Jhanjhi, N. COVID-19 impact on educational system globally. *Emerging Technologies For Battling Covid-19: Applications And Innovations*. pp. 257-269 (2021)

[74] Alsubaie, A., Alaithan, M., Boubaid, M. & Zaman, N. Making learning fun: Educational concepts & logics through game. *2018 20th International Conference On Advanced Communication Technology (ICACT)*. pp. 454-459 (2018)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: Nov 2, 2023
Accepted: Jan 29, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1631–1646, DOI 10.12694/scpe.v25i3.2791

MOBILE SMART APP AND ITS APPLICATION IN IMPROVING THE EFFICIENCY OF ENGLISH HOMEWORK CORRECTION

HUIYING SHAO[*]AND ZAN LIU[†]

Abstract. The heavy amount of English homework correction has resulted in Teachers' lax examination of homework, insufficient attention to homework problems, and low attention to homework correction. With the continuous growth of the number of educated people, more and more schools begin to have the problem of low efficiency of English homework correction. Therefore, in order to optimize the homework correction system, improve the efficiency of English teachers' homework correction, and give full play to teachers' positive feedback on homework, a smart app on mobile phone can be developed to scan and correct traditional paper homework. Based on image processing technology and neural network algorithm, this paper designs and establishes a mobile app that can recognize and extract English homework topics and handwriting with nearly 90% accuracy through Android system platform. Based on the homework answers entered in the database, the rapid correction of English homework can be realized. After using this software, the overall efficiency of English homework grading has significantly improved. For multiple-choice and fill in the blank questions, the total factor productivity of 14 and 15 units was greater than 1, accounting for 63.6% and 68.1% of the nursing units participating in the study, respectively. This indicates that the efficiency of English homework grading in most units is constantly improving and showing a good development trend. Among them, the homework correction efficiency for multiple-choice questions is only 6 units, and the pure technical efficiency is less than 1, indicating that the improvement of technical means has a significant impact on efficiency. In order to verify the applicability of the software, data envelopment analysis is used to analyze the application of the mobile app to improve the efficiency of English homework correction. the results show that when the smart phone software is not put into use, the efficiency of teachers' English homework correction is poor, and the technical level in the process of correction is too low. After the mobile intelligent software is put into use, the overall efficiency of English homework correction has increased significantly, which can greatly alleviate the pressure faced by English teachers in the process of correcting homework.

Key words: Mobile Smart app, English homework, Image processing technology, Neural network algorithm, Data envelopment analysis, Efficiency

1. Introduction.

Homework is an important and effective management means in the teaching process [1]. It can consolidate the teaching content by repeating the existing knowledge or skills [2]. In China, English is an important compulsory course, which is studied by all students in primary school. Although the proportion of English in school subjects has declined in recent years, English is still one of the most concerned courses for students and teachers. However, with the continuous improvement of education level and the continuous growth of the number of educated people, teachers are facing increasing pressure on teaching. the increasing number of examinations and relatively few class hours have brought great trouble to English teachers. the heavy amount of English homework correction has resulted in Teachers' lax examination of homework, insufficient attention to homework problems, and low attention to homework correction [3]. These problems not only weaken the consolidation effect of homework on teaching content, but also increase the burden on students. Even the inefficient efficiency of homework correction will curb students' motivation to complete homework and seriously affect students' interest in learning English courses [4]. Therefore, optimizing the homework correction system, improving the efficiency of English teachers' homework correction, and giving full play to teachers' positive feedback on homework are the most important problems to be solved in current English teaching. Applications can utilize natural language processing and machine learning techniques to automatically analyze students' English homework and provide precise feedback and suggestions. Compared with traditional grading methods, intelligent grading can reduce human errors, improve the accuracy and efficiency of grading. Analyze students' homework through algorithms, identify potential problems such as grammar errors, improper vocabulary usage,

[*] Department of Economic Management, Yantai Engineerig and Technology College, Yantai Shandong 264006, China (huiyingshaoh@outlook.com)

[†] Department of Industrial Technology Application, Yantai Engineering and Technology College, Yantai Shandong 264006, China

etc., and present these problems to students in an intuitive way. This approach can help students better understand their own problems and guide them to focus more on these issues, improving learning outcomes. In addition, the application can also provide real-time grading feedback, allowing students to timely understand their homework situation and make improvements based on problems. This feedback method can help students better grasp knowledge and improve learning efficiency.

There are some problems in the traditional method of correcting English homework. Firstly, lax exams are one of the important issues. Due to the limitations of manual grading, it is difficult to comprehensively and meticulously evaluate each student's homework, which can easily lead to unfair evaluation. Secondly, insufficient attention is also a problem. Manual grading is easily affected by fatigue and negligence, making it difficult to maintain a high level of attention, and can easily lead to grading errors and omissions. Finally, low efficiency is also a problem. Manual correction requires a lot of time and effort, especially in situations with a large number of students, and the efficiency of correction is low, making it difficult to meet teaching needs. Mobile intelligent applications have become an important tool in people's daily life and work. Especially in the field of education, mobile intelligent applications have brought many innovations and conveniences to the teaching and learning process. Among them, as a global language, the importance of teaching and learning English is self-evident. The grading of English homework is an important part of the teaching process, which not only helps students understand their learning situation, but also helps teachers evaluate the quality of teaching. However, traditional English homework grading methods suffer from problems such as lax exams, insufficient attention, and low efficiency, which seriously affect the effectiveness of teaching and learning. Therefore, how to use mobile intelligent applications to solve these problems and improve the efficiency and quality of English homework grading has become a current research hotspot.

Mobile intelligent applications have the characteristics of portability, real-time performance, and personalization, which bring new ideas and methods for English homework correction. Firstly, mobile intelligent applications can automatically correct grammar and vocabulary errors in English homework through technologies such as natural language processing and machine learning, reducing human errors and omissions, and improving the accuracy and efficiency of correction. Secondly, mobile intelligent applications can conduct in-depth analysis and understanding of students' homework, identify their problems, and provide personalized learning suggestions and improvement plans to help students better grasp knowledge and improve learning outcomes. Finally, mobile intelligent applications can provide real-time feedback on grading, allowing students to timely understand their homework situation and make improvements based on problems, thereby improving learning efficiency. In summary, mobile intelligent applications have great advantages and potential in improving the efficiency and quality of English homework grading. By utilizing mobile intelligent applications, the problems existing in traditional grading methods can be solved, the accuracy and efficiency of grading can be improved, and students can better grasp knowledge and improve learning outcomes. Therefore, this article further explores the application and optimization strategies of mobile intelligent applications in English homework correction, bringing more convenience and innovation to the teaching and learning process.

2. Related Work. In recent years, more and more schools have begun to pay attention to improving the efficiency of homework correction [5]. With the development of information technology and computer technology, electronic homework, online examination and online learning have become important teaching means in Colleges and universities. Although English is not an important course in all countries, foreign intelligent homework systems started early, which can provide reference experience for China. Among them, the more distinctive ones are the web assign platform developed by the Department of physics of North Carolina State University in the United States [6], the WebCT platform developed by the Department of computer science of Columbia State University, the owl system developed by the computer teaching technology center of the University of Massachusetts [7], and the online operation system developed by fern University Hagen in Germany [8]. Through these systems, teachers can mark homework online and realize the statistical analysis of students' homework scores [9]. This greatly improves the efficiency of teachers' homework correction and helps teachers understand students' learning conditions. However, compared with the homework design in other countries, China has higher requirements for the standardization of homework. At the same time, many schools still use traditional paper homework because English homework contains multiple-choice questions, composition questions, judgment questions and other types of questions. This makes it difficult to promote the network operation system in China. We should consider developing an intelligent correction system that can scan and process paper operations. With the increasing popularity of smart phones and the growing maturity of mobile camera technology, for domestic educational institutions, developing a smart app that can be used for mobile phones to scan and correct traditional paper homework has quite important social significance and economic value. Image processing technology [10] is a subject developing with the development of human civilization. With the mass production of image data, image processing technology has gradually become a special subject. the rise of computer technology has made a qualitative change in image processing technology. Compared with image analog processing, digital image processing has gradually become the main body of image analysis technology. In the 1970s, relying on computer technology, digital image processing and analysis technology has been booming and widely used. Digital image processing technology has gradually entered many fields [11], industries and people's lives.

Artificial neural networks can learn rules and patterns for grading homework by training a large number of English homework samples. Then, use these rules and modes to automatically grade students' English homework, improving the efficiency of grading. Artificial neural networks can analyze students' English homework and detect common errors such as grammar, spelling, and tense. This can help teachers identify student problems more quickly and provide timely guidance and correction. Different types and difficulty levels of English homework require selecting different artificial neural network models for training. Therefore, it is necessary to choose a suitable model based on the actual situation to ensure the accuracy and efficiency of the correction. In summary, artificial neural networks have broad application prospects in improving the efficiency of English homework grading. It can automatically correct errors, detect errors, provide intelligent prompts, and evaluate feedback, thereby improving the efficiency and accuracy of correction. The research of artificial neural networks [12] originated from the interdisciplinary research of physics, psychology and neurophysiology by Herman von Helmholtz, Ernst Mach and others in the late 19th century. In the late 1950s, Frank Rosenblatt proposed perceptron networks and associative learning rules, This makes the neural network initially have the ability of pattern recognition, and has aroused the world's interest in the research of computer network [13]. In the following decades, thousands of papers have appeared in the field of neural network research. the continuous progress of computer technology makes it possible to use high-speed computing institutions to build neural networks to solve practical problems. Neural networks are widely used in aviation, medicine, economics, electronic engineering and other fields [14], such as aircraft flight control system, cancer cell analysis, corporate financial analysis, integrated chip layout and machine vision.

Data Envelopment Analysis (DEA) [15] is a decision-making method proposed by James et al. (1978). This analysis method is also called DEA model. DEA model is a nonparametric estimation method of linear programming [16]. DEA model integrates management, mathematics, operations research and other multidisciplinary knowledge. Convex analysis and linear programming are the main analysis tools to calculate the efficiency between the same decision-making units, and thus realize the evaluation of the evaluation object. Data Envelopment Analysis (DEA) is a non parametric efficiency evaluation method used to evaluate the relative efficiency of decision units (DMUs). DEA can be used to analyze the efficiency and effectiveness of mobile applications in improving the efficiency of English homework grading. In DEA, DMU is the object to be evaluated. In this assessment, each teacher or student who uses mobile applications for English homework grading can be considered as a DMU. Choosing appropriate input and output indicators is a key step in DEA analysis. For English homework correction, input indicators can include correction time, correction frequency, and computational resources used. Output indicators can include accuracy of grading, mastery of student knowledge points, etc. The result of Data Envelopment Analysis (DEA) is the relative efficiency value of each DMU. These results are used to evaluate the effectiveness of mobile applications in improving the overall efficiency of English homework grading. If the DEA results show that the use of mobile applications significantly improves the efficiency and quality of grading, then it can prove the effectiveness of mobile applications. To demonstrate the effectiveness of mobile applications in improving the overall efficiency of English homework grading, DEA results need to be compared and analyzed with the following factors. If the correction time is significantly reduced after using a mobile application, it can indicate that the mobile application has improved the efficiency of correction. If the number of corrections significantly decreases after using a mobile application, it can also indicate that the mobile application has improved the efficiency of corrections. In the implementation and evaluation process of smart applications, the following limitations or challenges may be encountered: the development and maintenance of smart applications require professional technical and resource support. If there is a lack of relevant technology and resources, it may impose limitations on the implementation and evaluation of the application. Intelligent applications need to handle sensitive data, such as student assignments and personal information. If data privacy and security are not guaranteed, it may pose challenges to the implementation and evaluation of applications. Related studies have shown that the implementation of AR technology has a significant positive impact on the motivation level of learners towards teaching materials in the classroom. AR technology in the classroom is an interactive and entertaining tool that transforms a monotonous learning atmosphere into an engaging and effective learning atmosphere. AR strengthening foreign language education significantly improves students' attitudes towards English courses and increases their beliefs in English self-efficacy [17].

To sum up, based on image processing technology and neural network algorithm, this paper establishes a mobile app that can extract data from scanned paper job pictures through cloud processor. Based on the homework answers entered in the database, the rapid correction of English homework can be realized. At the same time, data envelopment analysis is used to analyze the application of the mobile app to improve the efficiency of English homework correction, which proves the rationality and applicability of this method.

3. Methods.

3.1. Image processing technology. For a picture, the discrete function f (x) involving two variables can be used to describe it. This function is the image function, which can be expressed by matrix, and its definition domain is as shown in formula (2.1):

$$R\!=\left\{(x,y)\ ,1\leq x\leq x_m,1\leq y\leq y_n\right\}\tag{3.1}$$

where x$_M$ and y$_M$ represent the maximum coordinates of the image, which is related to the size and resolution of the image.

The basic unit of digital image is called image element, which is called pixel for short. the spatial resolution of image is proportional to the pixels it contains. An image can be represented as a two-dimensional matrix (each element represents a pixel) as shown in formula (3.2), where m and V are the number of rows or columns of the image respectively:

$$F=\left[\begin{array}{cccc} f_{11} & f_{12} & \cdots & f_{1N} \\ f_{21} & f_{22} & \cdots & f_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ f_{M1} & f_{M2} & \cdots & f_{MN} \end{array}\right]\tag{3.2}$$

By integrating the brightness of image pixels, the global information can be extracted from the original image matrix, which makes the calculation process simple. the specific formula is as shown in formula (3.3):

$$ii\left(i,j\right)=\sum_{k\leq i,l\leq j}f\left(k,l\right)\tag{3.3}$$

Hierarchical data structure is an important part of image processing. In order to improve computing efficiency, t-pyramid structure is usually used. the structure is a tree as shown in Figure 3.1. Let 2L be the maximum resolution of the image. the definition of t-pyramid is: a node set P is as shown in formula (3.4).

$$P\!=\left\{P\!=\left(k,i,j\right)\ \middle|k\in\left[0,L\right];i,j\in\left[0,2^k-1\right]\right\}\tag{3.4}$$

Mapping between nodes F is as shown in formula (3.5):

$$F\left(k,i,j\right)=\left(k-1,\frac{i}{2},\frac{j}{2}\right)\tag{3.5}$$

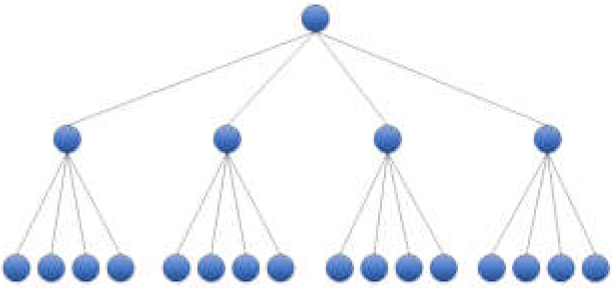


Fig. 3.1: Structure diagram of t-pyramid neural network

3.2. Artificial neural network technology.. Neural network is a collection of interconnected neurons, and the computer meridian element is the basic unit to complete data processing in the network. Through the learning of neurons, it can find the linear or nonlinear relationship between input and output from the network input data, so as to produce output prediction for future input signals. In practical applications, neurons are often designed as a processor or processing unit that can accept input signals and generate a single output signal. When the input signal enters the neuron, there will be a weight associated with it to generate an input function. Similarly, in most cases, the output is also a function of the weighted sum of the input signals. Let the neuron input signal be represented by $V1, V2$. And the corresponding input weights are $W1, W2...$, and then the actual input of the neuron is as shown in formula 3.6:

$$x = \sum_{i=1}^{n} v_i w_i - b \tag{3.6}$$

Where B is the bias value related to the neuron, and the output transfer function $f(x)$ of the neuron is generally expressed in the following two forms listed in formula (3.7) and (3.8):

$$f(x) = \begin{cases} 0, x \leq 0 \\ 1, x > 0 \end{cases} \tag{3.7}$$

$$f(x) = \frac{1}{1+e^{-x}} \tag{3.8}$$

The neurons in the neural network are operated in parallel. the set of these neurons operated in parallel is called the layer of the neural network. A layer composed of s neurons can be represented by Figure 3.2. the number of input signals in the figure is R.

3.3. Data envelopment analysis. Data envelopment analysis is a method to evaluate the relative efficiency between decision-making units. It is suitable for evaluating the efficiency of decision-making units with multiple inputs and outputs, and the evaluation results are not affected by the index measurement unit, and the model it establishes does not need to deal with the data dimensionless. At present, the commonly used analysis model CCR model has the following main conditions listed in formula (3.9):

$$\begin{cases} \max h_{j0} = \frac{\sum_{r=1}^{s} u_r y_{rj_0}}{\sum_{i=1}^{m} v_i x_{ij0}} \\ s.t. \frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \leq 1, j = 1, 2, \cdots, n \\ v = (v_1, v_2, \cdots, v_m)^T \geq 0 \\ u = (u_1, u_2, \cdots, u_s)^T \geq 0 \end{cases} \tag{3.9}$$

In which:

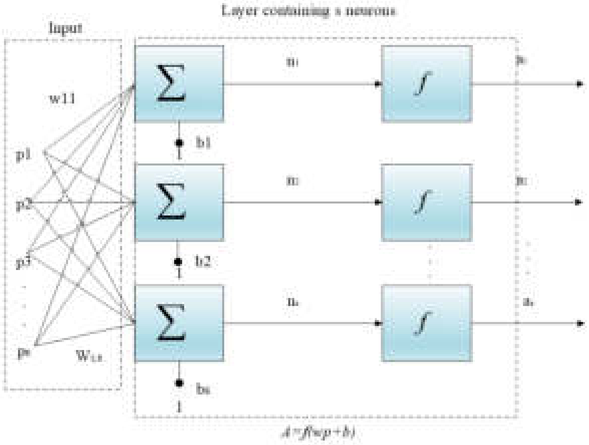


Fig. 3.2: Schematic diagram of layered neural network structure

1. x$_{ij}$ is the input of type i of the jth decision-making unit;
2. y$_{rj}$ is the output of type r output of the jth decision-making unit;
3. v$_{i}$ is the weight of the i-th input;
4. u$_{r}$ is the weight of the r-th output;

Through dual transformation and the introduction of relaxation variable s+ and residual variable s-, the above model can be changed into formula (3.10):

$$\begin{cases} \min\theta \\ s.t. \sum_{j=1}^{n}\lambda_i x_j + s^{+} = \theta x_0 \\ \sum_{j=1}^{n}\lambda_j y_j - s^{-} = \theta y_0 \\ \lambda_j \geq 0, j = 1, 2 \cdot \cdot, n \\ \theta no \lim its \\ s^{+} \geq 0, s^{-} \leq 0 \end{cases} \tag{3.10}$$

Where θ is the effective value of the decision-making unit, and its value has different meanings under different conditions:

1. θ= 1. And s+=0, it means that the cardinality effect and scale of the decision-making unit are effective at the same time.
2. θ= 1. If the input or output is not all positive, it indicates that the cardinality effect and scale of the decision-making unit are not effective at the same time.
3. θ< 1, it indicates that the cardinality effect and scale of the decision-making unit are not effective.

The reason for choosing data envelopment analysis (DEA) in this paper is that data envelopment analysis (DEA) does not need to consider the functional relationship between input and output, nor the assumption of estimating parameters and index weights in advance, so as to avoid the differences caused by researchers' subjective assumptions. At the same time, for non DEA effective decision-making units, it can analyze the redundancy and deficiencies of input and output, and find the direction of further improvement.

4. Establishment of software.

4.1. identification technology process. In order to improve the accurate capture of the questions and answers, in the recognition process, first of all, the homework pictures scanned by the mobile phone are meshed, and the image files are cut according to the small questions and the topic numbers are recognized. By constantly moving the search area to search the question box by box, based on the entered job related information, the normalized correlation matching method is used to calculate the matching degree returned after each match,

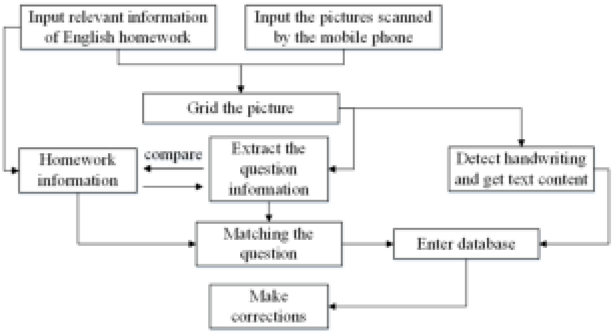

Input relevant information of English homework
Input the pictures scanned by the mobile phone
Grid the picture
Homework information
compare
Extract the question information
Detect handwriting and get text content
Matching the question
Enter database
Make corrections

Fig. 4.1: Detection process of English homework scanning and content extraction

and finally the threshold is used to determine whether the result is matched. the matching formula of normalized similarity measurement is as shown in formula (4.1):

$$R_{\mathrm{value}}^{\mathrm{ij}}=\frac{\sum_{\mathrm{i}=1}^{\mathrm{m}}\sum_{\mathrm{j}=1}^{\mathrm{n}}\left[S^{\mathrm{i,j}}\left(m,n\right)T\left(m,n\right)\right]}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{m}}\sum_{\mathrm{j}=1}^{\mathrm{n}}\left[T\left(m,n\right)\right]^{2}}\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{m}}\sum_{\mathrm{j}=1}^{\mathrm{n}}\left[S^{i,j}\left(m,n\right)\right]^{2}}} \tag{4.1}$$

After determining the topic, determine the topic range according to the spacing between adjacent question number boxes. the handwriting in the title is detected by edge detection, and the text content is obtained. the segmentation of re-use image cuts out the answer content and question part, and finally enters them into the database respectively for subsequent automatic correction. the specific detection process is shown in Figure 4.1.

4.2. Optimization of neural network algorithm. Considering the richness of English homework questions, the software algorithm adopts BP algorithm with multi-layer network structure. This method optimizes grid computing and sensitivity recursion based on Jacobian matrix (formula 4.2) and chain method.

$$F^{\mathrm{m}}\left(n^{m}\right)=\begin{bmatrix} f^{m}\left(n_{1}^{m}\right) & 0 & \cdots & 0 \\ 0 & f^{m}\left(n_{2}^{m}\right) & \cdots & f_{2N} \\ \ldots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & f^{m}\left(n_{n}^{m}\right) \end{bmatrix} \tag{4.2}$$

The implementation process is mainly based on three steps:

1. the input is transmitted forward through artificial neural network.
2. the sensitivity is back propagated through the grid.
3. Use the steepest descent method to update the weight and offset values.

Based on this algorithm, the matching condition of the scanned image of an English homework is calculated, the image is scanned line by line and then column by column, and the matrix is established in the memory. Finally, the grid error change curve is shown in Figure 3.2:

It can be found that the step error of 140 calculation under this algorithm can be stabilized to a small value, indicating that the calculation is relatively stable. At the same time, English assignments of different question types are selected for topic and handwriting extraction. the overall recognition rate is shown in Table 5.1:

Combined with the recognition rate results, it can be found that the recognition rate of the algorithm for the topic and handwriting is more than 90%, the operation is stable and reliable, and it can recognize the topic content and handwriting in English homework more accurately, which provides a guarantee for the subsequent rapid correction.

This new comprehensive detection method has the following advantages over the traditional answer card verification technology:

1. the algorithm is simple to realize and the development cost is low. the software developed in this paper uses the principle of exclusion method. After identifying the question number and handwriting, it will

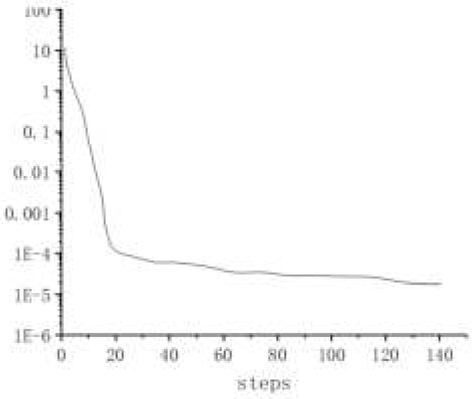


Fig. 4.2: The change of the grid error with steps

Table 4.1: The overall recognition rate of English homework

Number	choice question	Blank filling questions	On answering questions
1	92.3%	99.5%	97.0%
2	92.4%	98.1%	92.9%
3	99.9%	95.0%	91.3%
4	94.4%	95.8%	90.4%
5	94.4%	93.5%	98.4%
6	97.2%	94.3%	96.6%
7	91.6%	96.0%	92.1%
8	95.5%	94.8%	95.6%
9	95.1%	91.3%	94.4%
10	97.2%	94.4%	94.0%
11	98.9%	99.5%	96.9%
12	97.8%	92.0%	92.8%
13	98.6%	94.8%	98.5%
14	93.7%	98.5%	91.5%
15	98.6%	98.3%	92.8%
16	97.5%	90.8%	90.6%
17	93.9%	94.3%	94.6%
18	95.5%	94.4%	96.6%

automatically scan the questions and answers centered on it. the shape of the recognition object is simple, the block area is small, and the noise influence is small after image preprocessing, and the matching quality can be further improved by image compression, variable step matching and other optimization methods;

2. Through the grid division of the scanned image, it can provide rough coordinate positions for all inspections, so that the reviser can automatically locate the topic he wants to find;
3. the statistical characteristics of the image change obviously, and the error recognition rate is low. 4. Without relying on a card reader, you can scan and extract pictures in batches only with a smart phone, which greatly shortens the correction time.

4.3. Implementation of Mobile Smart app. Mobile Smart app is designed based on Android development platform. In this study, the development under Android is based on eclipse development environment. As an open source, Java based extensible development platform, eclipse can be used as Java integrated develop-

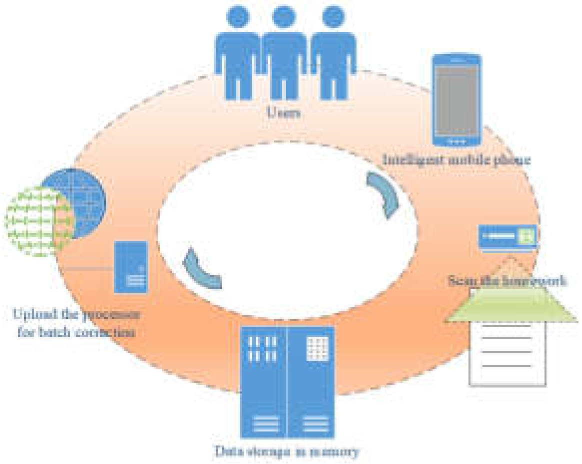


Fig. 4.3: The specific process for the Mobile Smart app

ment environment (IDE), including plug-in development environment (PDE). the main functions of the mobile smart app designed in this paper include uploading scanned pictures, connecting with the processor, inputting answers and outputting correction results. the specific process is shown in the Figure 4.3 below.

5. Analysis of English homework correction efficiency based on DEA static analysis. In order to better analyze the dynamic change law of the human resource efficiency of each nursing unit, this study, based on the DEA static analysis, combined with the Malmquist index model, further investigated the change law and heterogeneity of the English homework correction efficiency of each teaching unit by analyzing the relevant data of teachers in each school in terms of English homework correction and question type.

5.1. index introduction. Malmquist (1953) proposed Malmquist index in the early 1950s. At present, this method is mostly used together with data envelopment model. This index uses the ratio of distance function to calculate the input-output index. With the help of Malmquist index analysis, the efficiency changes of decision-making units in different cycles can be analyzed by the formula (5.1).

$$M_{j0}^{t+1}\left(X_{j0}^{t+1},Y_{j0}^{t+1},X_{j0}^{t},Y_{j0}^{t}\right)=\left[\frac{F_{j0}^{t}\left(X_{j0}^{t+1},Y_{j0}^{t+1}\right)}{F_{j0}^{t}\left(X_{j0}^{t},Y_{j0}^{t}\right)}\bullet\frac{F_{j0}^{t+1}\left(X_{j0}^{t+1},Y_{j0}^{t+1}\right)}{F_{j0}^{t+1}\left(X_{j0}^{t},Y_{j0}^{t}\right)}\right]^{\frac{1}{2}} \tag{5.1}$$

According to the definition of Malmquist productivity index, it can be divided into technology change and resource allocation efficiency change rate. the former refers to the ratio of the actual input and the maximum output of the enterprise under the given input factors, and the latter refers to the optimal combination of given output input under the given technology and price conditions. By decomposing Malmquist productivity index, it can be regarded as the similar reasons for improving comprehensive efficiency, and the efficiency changes caused by multi index changes can be comprehensively considered. the formula used in the analysis is as follows:

$$AC_{i}^{t+1}\left(y^{t+1},x^{t+1},y^{t},x^{t}\right)=\frac{F_{i}^{t+1}\left(y^{t+1},x^{t+1}\right)}{F_{i}^{t}\left(y^{t},x^{t}\right)} \tag{5.2}$$

$$M_{i}^{t+1}\left(u^{t+1},x^{t+1},u^{t},u^{t}\right)=TC_{i}^{t+1}\left(y^{t+1},x^{t+1},y^{t},x^{t}\right)\bullet AC_{i}^{t+1}\left(y^{t+1},x^{t+1},y^{t},x^{t}\right) \tag{5.3}$$

According to the definition, when the productivity index$M_{i}^{t+1}\left(u^{t+1},x^{t+1},u^{t},u^{t}\right)>1$, the comprehensive efficiency improves. If the change rate of one of the items after the decomposition of Malmquist productivity index is greater than 1, it indicates that this item is the reason for improving the comprehensive efficiency. If a certain item is less than 1, it indicates that this indicator is the reason for the decline of efficiency.

Table 5.1: The research results of comprehensive efficiency, pure technical efficiency and scale efficiency calculated

	choice question			Blank filling questions		
Research unit	CE	TE	SE	CE	TE	SE
1	1.10	0.94	0.74	1.03	0.24	0.46
2	0.50	0.88	0.27	0.20	0.35	0.23
3	1.07	0.98	0.30	0.37	0.33	0.26
4	0.93	0.61	1.11	1.19	1.08	0.42
5	1.00	0.96	0.88	1.01	0.86	0.93
6	0.35	0.46	0.60	0.80	0.96	0.50
7	1.03	0.38	0.77	0.44	0.26	1.08
8	0.69	0.40	0.63	0.80	1.06	0.97
9	0.95	0.49	0.66	0.49	0.94	0.98
10	0.71	0.91	0.34	0.25	0.34	0.92
11	0.38	0.33	0.91	0.23	0.32	0.72
12	0.40	0.26	1.03	1.13	0.77	0.38
13	0.79	0.69	0.81	0.95	0.86	0.70
14	0.76	1.11	0.65	1.19	0.97	0.41
15	1.12	0.30	1.16	0.57	0.86	1.02
16	1.13	0.46	1.20	0.60	0.50	1.12
17	0.40	0.78	0.99	1.06	1.07	0.55
18	0.83	0.95	0.31	0.90	1.12	0.32
19	0.44	1.15	0.33	0.60	0.42	0.74
20	0.43	0.52	0.47	0.95	0.21	0.82
21	0.35	0.93	0.94	0.21	0.47	0.90
22	0.89	0.95	1.05	0.73	0.45	0.78
Means	0.74	0.70	0.74	0.71	0.66	0.69

5.2. Data source. The research data mainly comes from the teaching data provided by the school after the software is put into use. The survey mainly involves the allocation of related resources such as time investment, manpower investment, and grade output for different question types in homework grading for each teaching unit. In addition, for the collected data indicators, establish a database using Excel, organize team professionals to inspect the collected data, and check and correct abnormal data.

5.3. Analysis results. After screening, 22 teaching units were selected as the research object to carry out efficiency research, and the input and output indicators of multiple-choice questions and blank filling questions were calculated and analyzed. When teachers do not use this software to correct English homework, the research results of comprehensive efficiency, pure technical efficiency and scale efficiency calculated are shown in Table 5.1. In the table, CE represents comprehensive efficiency, TE represents technical efficiency, and Se represents scale efficiency.

From the perspective of comprehensive benefit index, teachers' efficiency in correcting multiple-choice questions and blank filling questions has not reached DEA effectiveness, with average comprehensive efficiency of 0.74 and 0.71 respectively, and average scale efficiency of 0.74 and 0.69; the average pure technical efficiency is 0.70 and 0.66. the efficiency of filling in the blank is relatively poor, but the correction efficiency of the two types of questions is at a low level.

Through the comparative analysis of the resource efficiency of decision-making units from the perspective of the development of each unit, it is found that only 6 units, accounting for 27.3% of the total, have achieved effective comprehensive efficiency in the correction of multiple-choice questions, and the resource efficiency of the other 16 research units is less than 1, indicating that the input-output has not reached the optimal state. There are only 6 units with effective comprehensive efficiency of correcting blank filling questions, indicating that the input and output have not reached the optimal state.

Table 5.2: The research results of comprehensive efficiency, pure technical efficiency and scale efficiency calculated

	Choice question			Blank filling questions		
Research unit	CE	TE	SE	CE	TE	SE
1	1.17	0.85	1.37	1.11	1.48	0.87
2	1.14	0.78	1.58	0.98	1.31	0.86
3	1.61	0.93	0.95	1.56	1.24	0.99
4	0.82	1.48	1.59	1.03	1.19	1.13
5	1.26	1.64	1.66	1.22	0.98	1.22
6	1.49	1.10	0.84	1.68	1.19	1.22
7	1.50	0.72	1.02	0.70	1.07	1.55
8	1.64	1.04	1.01	1.48	0.78	1.17
9	1.47	1.46	1.49	1.18	1.29	1.47
10	1.39	0.90	0.98	1.30	0.73	0.89
11	1.63	1.52	1.46	1.12	1.31	1.20
12	0.72	0.97	1.12	0.81	1.32	1.69
13	1.22	1.24	1.16	0.76	1.29	1.11
14	1.70	1.46	1.35	1.31	1.47	0.92
15	0.76	1.42	0.89	0.73	1.40	1.09
16	1.61	1.35	1.18	1.51	0.93	1.31
17	0.87	1.32	0.76	1.53	1.23	1.25
18	1.26	0.83	1.48	1.32	1.35	1.11
19	1.14	1.16	1.04	1.11	1.31	1.08
20	1.51	1.06	1.60	0.87	0.78	1.56
21	0.79	1.62	1.34	0.89	0.91	1.58
22	0.88	1.08	1.56	1.44	1.05	1.23
Means	1.25	1.18	1.25	1.17	1.16	1.20

From the perspective of decomposition, that is, comprehensive technical efficiency = pure technical efficiency × Scale efficiency, analyze the nursing units with relatively ineffective comprehensive technical efficiency. the pure technical efficiency of only two nursing units in the correction of multiple-choice questions is greater than 1, while the scale efficiency is less than 1, indicating that the scale efficiency is the main factor restricting the efficiency, and measures need to be taken to optimize the correction methods and improve the scale efficiency. the pure technical efficiency of four nursing units was greater than 1, while the scale efficiency was less than 1. All the above shows that scale efficiency is the main factor restricting efficiency, and measures need to be taken to optimize the correction methods and improve scale efficiency. In addition, the pure technical efficiency and scale efficiency values of 14 and 15 research units of multiple-choice questions and blank filling questions are less than 1, which also shows that the technical level is too low in the process of correcting English homework, resulting in the low efficiency of each unit.

After teachers use the software to correct English homework, the research results of comprehensive efficiency, pure technical efficiency and scale efficiency are calculated based on the data fed back by the school, which is shown in Table 5.2.

It can be found that from the perspective of the comprehensive benefit index, teachers' correction efficiency in multiple-choice questions and blank filling questions has been significantly improved, with the average comprehensive efficiency of 1.25 and 1.17 respectively, and the average scale efficiency of 1.25 and 1.20; the average pure technical efficiency is 1.18 and 1.16, which are nearly 70% higher than those without software. the correction efficiency of the two types of questions is at a high level. At the same time, it can be found that the comprehensive efficiency of teachers in correcting multiple-choice questions has increased to 16 units, accounting for 72% of the total, and the comprehensive efficiency of correcting blank filling questions has also increased to 15 units, accounting for 68% of the total, indicating that the input-output has reached an excellent state. In

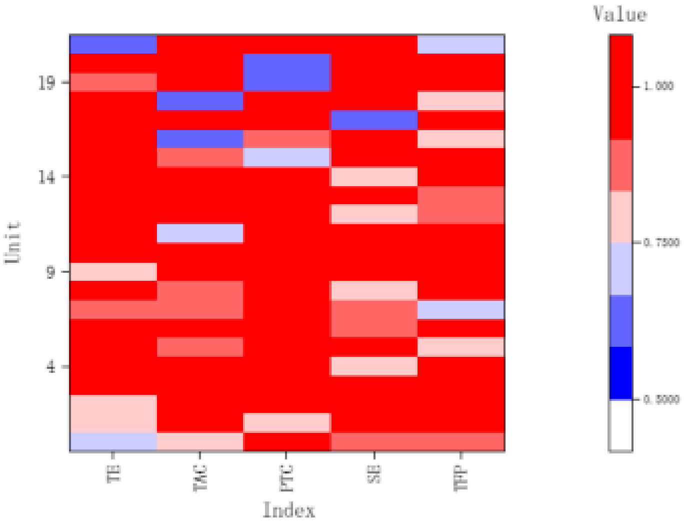


Fig. 5.1: The distribution of comprehensive efficiency, technical efficiency and scale efficiency of each unit for choice questions

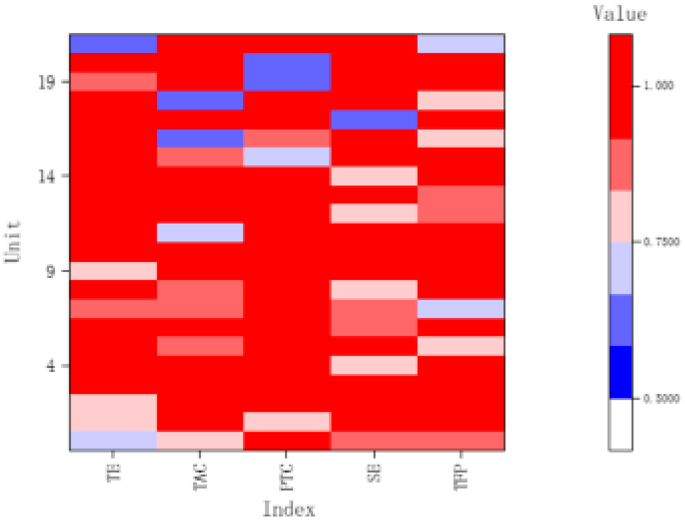


Fig. 5.2: The distribution of comprehensive efficiency, technical efficiency and scale efficiency of each unit for blank filling questions

order to better compare the changes in efficiency after correcting English homework with software, Figure 5.1 and Figure 5.2 use the form of radar chart to describe the distribution of comprehensive efficiency, technical efficiency and scale efficiency of each unit.

It can be seen from the figure that after using the software, the outline surrounded by dots and lines has expanded significantly, and all efficiency has been significantly improved. This result also shows that the software can greatly alleviate the pressure faced by English teachers in the process of correcting homework.

In order to better analyze the dynamic change law of the efficiency of each unit, the relevant data of 22 units are analyzed in combination with the Malmquist index model. After the software is tried, the heat map of

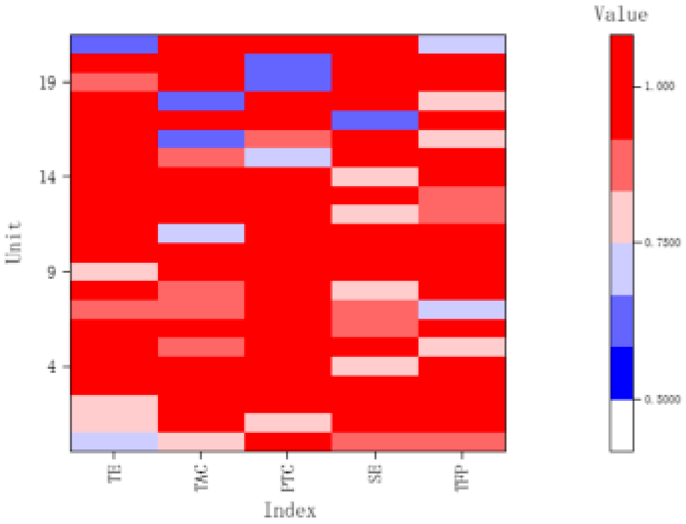


Fig. 5.3: The heat map of each efficiency index for Choice questions

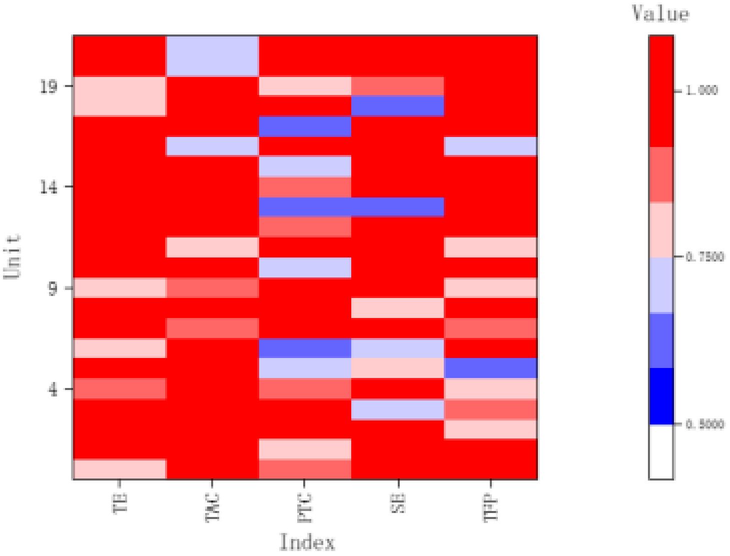


Fig. 5.4: The heat map of each efficiency index for blank filling questions

each efficiency index is listed in Figure 5.3 and Figure 5.4. In the figures, TAC refers to technological progress, PTC refers to pure technical efficiency, and TFP refers to total factor productivity:

From the efficiency index heat map of each unit, the overall efficiency of English homework correction has increased significantly after the use of the software. For multiple-choice questions and blank filling questions, the total factor productivity of 14 and 15 units is > 1, accounting for 63.6 and 68.1% of the nursing units participating in the research, indicating that the efficiency of English homework correction in most units is constantly improving, showing a good development trend. Among them, the homework correction efficiency of multiple-choice questions is only 6 units, and the pure technical efficiency is less than 1, which shows that the improvement of technical means has a great impact on efficiency.

Table 5.3: The average Malmquist efficiency index after using software to correct English homework

	TE	TAC	PTC	SE	TFP
Choice questions	1.08	1.11	1.11	1.07	1.14
Blank filling questions	1.09	1.11	0.96	1.05	1.06

The average Malmquist efficiency index after using software to correct English homework is summarized in Table 5.3.

It can be found that after the software is put into use, the values of technical efficiency, technological progress, scale efficiency and total factor productivity are all greater than 1, and the growth rate is more than 5%, except that the pure technical efficiency value of filling in the blank is less than 1. From the mean value of the three decomposition factors, the change of technological progress index has the greatest contribution to the change of total factor productivity. the choice question type and fill in the blank question type have increased by 11% respectively, the scale efficiency index has increased by 7% and 5% respectively, and the technical efficiency has increased by 8% and 9% respectively. the later investigation found that after the intelligent software was put into use, teachers' time for correcting homework was significantly reduced, providing them with more time to prepare for class preparation. Therefore, their own innovation ability was also improved, providing a continuous impetus for the overall improvement of English teachers' human resource efficiency. However, it is worth noting that for the question type of filling in the blank, due to the problems of redundant input and insufficient output of human resources in some schools, there is still a certain distance between the human resource efficiency of some teaching units and the best output.

The practical significance of mobile intelligent applications in improving the efficiency of English homework grading is very significant. The traditional method of correcting English homework usually requires teachers to spend a lot of time and effort, and sometimes it is inevitable to encounter omissions or low efficiency. Mobile intelligent applications greatly reduce the workload of teachers and improve the efficiency and accuracy of grading through functions such as automatic grading, error detection, intelligent prompts, and evaluation feedback. This can not only reduce the work pressure on teachers, but also enable them to focus more on more important tasks such as teaching design and student guidance.

In addition, mobile intelligent applications can also provide personalized learning suggestions and improvement plans, helping students better grasp knowledge points and improve learning outcomes. This personalized feedback and guidance is very beneficial for students' learning progress, as it can help them better identify their own problems and solve them in their studies. In terms of management, the implementation of mobile intelligent applications also needs to consider some challenges and issues. Firstly, the development and maintenance of applications require professional technical and resource support, therefore, schools need to invest a certain amount of funds and human resources for development and maintenance. Secondly, data privacy and security are also issues that need to be considered, and a comprehensive data management system and security measures need to be established to protect students' personal information and homework data [18, 19, 20].

6. Conclusion. In order to alleviate the pressure of English teachers in English homework correction at this stage and solve the problem of low correction efficiency. Based on image processing and artificial neural network technology, this paper designs a software that can extract the content of scanned job pictures. At the same time, a smart phone app is developed based on Android system platform and cloud server. After the mobile smart app is put into use, the efficiency analysis is carried out based on the feedback data provided by each teaching unit. the main conclusions are as follows:

1. Image processing and artificial neural network technology can better extract the content of paper English homework questions and the handwriting of answers. Based on the analysis of several cases, the designed algorithm has a recognition and extraction accuracy of 90% for English homework topics and handwriting. Based on the Android system platform, a stable and reliable intelligent app is developed. Its main functions include uploading scanned pictures, connecting with the processor, inputting answers and outputting correction results. the software can improve the speed of English teachers' English homework correction.

2. When the smart phone software is not put into use, it can be found that the efficiency of teachers' English homework correction is poor based on the data fed back by each teaching unit. At the same time, the technical level in the process of correcting English homework is too low, which leads to the low efficiency of each unit. This not only causes English teachers' lack of motivation to correct homework, but also leads to students' inability to consolidate the knowledge they have learned through homework. Therefore, we need to take technical means to optimize the correction methods and improve the scale efficiency.
3. After the smart phone software was put into use, the overall efficiency of correcting English homework showed a significant increase. From the mean value of the decomposition factors of the three Malmquist efficiency indexes, the change of technological progress index contributed the most to the change of total factor productivity. the choice question type and fill in the blank question type increased by 11% respectively, the scale efficiency index increased by 7% and 5% respectively, and the technical efficiency increased by 8% and 9% respectively. It shows that the software can greatly alleviate the pressure faced by English teachers in the process of correcting homework and improve the efficiency of English homework correction. However, the accuracy and efficiency of correcting mobile intelligent applications are influenced by data input and algorithm design. If the data input is inaccurate or the algorithm design is unreasonable, it may lead to deviation or errors in the grading results. Therefore, in the future, it is necessary to continuously optimize algorithms and data input methods to improve the accuracy and efficiency of grading.

REFERENCES

[1] Magalhães, P., Ferreira, D., Cunha, J. & Rosário, P. Online vs traditional homework: A systematic review on the benefits to students' performance. *Computers & Education.* **152** pp. 10386 (2020)

[2] Dettmers, S., Yotyodying, S. & Jonkmann, K. Antecedents and Outcomes of Parental Homework Involvement: How Do Family-School Partnerships Affect Parental Homework Involvement and Student Outcomes?. *Frontiers In Psychology.* **10**, 01048 (2019)

[3] Reynolds, L. B., & Shih, Y. -C. (2019). the learning effects of student-constructed word cards as homework for the adolescent English Language classroom. *System.* **1** (2019)

[4] Cadaret, C. & Yates, D. Retrieval practice in the form of online homework improved information retention more when spaced 5 days rather than 1 day after class in two physiology courses. *Advances In Physiology Education.* **42**, 305-310 (2018)

[5] Walkington, C., Clinton, V. & Sparks, A. the effect of language modification of mathematics story problems on problem-solving in online homework. (Instructional Science,2019)

[6] Cadaret, C. & Yates, D. Retrieval practice in the form of online homework improved information retention more when spaced 5 days rather than 1 day after class in two physiology courses. *Advances In Physiology Education.* **42**, 305-310 (2018)

[7] Nabulsi, L., Nguyen, A. & Odeleye, O. A Comparison of the Effects of Two Different Online Homework Systems on Levels of Knowledge Retention in General Chemistry Students. *Journal Of Science Education And Technology.* **30**, 31-39 (2020)

[8] Bergeler, E. & Read, M. Comparing Learning Outcomes and Satisfaction of an Online Algebra-Based Physics Course with a Face-to-Face Course. *Journal Of Science Education And Technology.* **30**, 97-111 (2020)

[9] Zeng, X., Yu, C., Liu, Y., Hu, X., Hao, Q. & Jiang, Y. ... Teng, B. (2018). the construction and online/offline blended learning of small private online courses of Principles of Chemical Engineering. *Computer Applications In Engineering Education.* **22044** (0)

[10] Gao, M., Wang, X., Zhu, S. & Guan, P. Detection and Segmentation of Cement Concrete Pavement Pothole Based on Image Processing Technology. *Mathematical Problems In Engineering.* **2020** pp. 1-13 (2020)

[11] Jin, X., Che, J. & Chen, Y. Weed Identification Using Deep Learning and Image Processing in Vegetable Plantation. *IEEE Access.* **2021**, 10940-10950 (2021)

[12] Zador, A. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature Communications.* **10** pp. 1 (2019)

[13] Lopez-Garcia, T., Coronado-Mendoza, A. & Domínguez-Navarro, J. Artificial neural networks in microgrids: A review. *Engineering Applications Of Artificial Intelligence.* **95** pp. 10389 (2020)

[14] Herzog, S., Tetzlaff, C. & W"org"otter, F. Evolving artificial neural networks with feedback. *Neural Networks.* **2019**, 153-162 (2020)

[15] Xu, T., You, J., Li, H. & Shao, L. Energy Efficiency Evaluation Based on Data Envelopment Analysis: A Literature Review. *Energies.* **13**, 3548 (2020)

[16] Shao, L., Yu, X. & Feng, C. Evaluating the eco-efficiency of China's industrial sectors: A two-stage network data envelopment analysis. *Journal Of Environmental Management.* **247** pp. 551-560 (2019)

[17] Ustun, A., Simsek, E., Karaoglan-Yilmaz, F. & Yilmaz, R. The effects of AR-enhanced english language learning experience on students' attitudes, self-efficacy and motivation. *TechTrends.* **66**, 798-809 (2022)

[18] Hussain, K., Hussain, S., Jhanjhi, N. & Humayun, M. SYN flood attack detection based on bayes estimator (SFADBE) for

MANET. *2019 International Conference On Computer And Information Sciences (ICCIS).* pp. 1-4 (2019)

[19] Lim, M., Abdullah, A., Jhanjhi, N. & Supramaniam, M. Hidden link prediction in criminal networks using the deep reinforcement learning technique. *Computers.* **8**, 8 (2019)

[20] Kumar, T., Pandey, B., Musavi, S. & Zaman, N. CTHS Based Energy Efficient Thermal Aware Image ALU Design on FPGA Springer Wireless Personal Communications. *An International Journal, ISSN.* pp. 0929-6212 (2015)

Edited by: Mudasir Mohd
Special issue on: Scalable Computing in Online and Blended Learning Environments: Challenges and Solutions
Received: Nov 3, 2023
Accepted: Feb 18, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1647–1657, DOI 10.12694/scpe.v25i3.2692

A LIGHTWEIGHT SYMMETRIC CRYPTOGRAPHY BASED USER AUTHENTICATION PROTOCOL FOR IOT BASED APPLICATIONS

A. MAHESH REDDY [*], DURVASI GUDIVADA [†], AND M. KAMESWARA RAO [‡]

Abstract. The utilization of IoT is expanding across various domains, including telecare, intelligent home systems, and transportation networks. In these environments, IoT devices generate data gathered on remote servers, requiring external users to authenticate themselves to access the data. However, existing authentication protocols for IoT must meet the crucial requirements of speed, security against multiple attacks, and ensuring user anonymity and un-traceability. The main objective of this work is to find lightweight symmetric cryptography-based user authentication protocol tailored for IoT-based applications, focusing on MIM (Man-in-the-Middle) attack prevention, enhanced anonymity, and secure communication between IoT nodes and remote servers via IoT gateways. Existing protocols often lack sufficient defenses against MIM attacks and do not adequately address the need for enhanced user anonymity and secure communication channels within the IoT framework. Our research has identified that authentication techniques based on pairing are susceptible to attacks targeting temporary session-specific data, impersonation, privileged insiders, and offline password guessing. Furthermore, using bilinear pairing in these techniques requires significant computational and communication resources to address the security as mentioned above concerns. A new authentication mechanism must be proposed and designed explicitly for IoT scenarios. The proposed approach exclusively utilizes hash and exclusive-or operations to ensure suitability within the IoT context; thoroughly evaluated the recommended protocol against existing authentication protocols, employing both informal and formal analytical routines like BAN logic, ROR model, and AVISPA simulation. Our findings suggest protocol not only enhances performance but also enhances security. The proposed approach is a tried-and-true strategy for improving security rules in practical Internet of Things (IoT) settings addressing the inherent challenges posed by authentication requirements in IoT environments. The accuracy 98.93%, and Node detection rate 46.57% were improved which is a better outcome.

Key words: Cryptography, Symmetric IoT applications, Avispa simulation, telecare, remote server, IoT gateway.

1. Introduction.

Healthcare, smart grids, transportation, and global roaming are just a few examples of how the IoT has brought in a new era of efficiency and improved quality of life [1]. In IoT-based telecare systems, medical equipment and sensors continuously monitor patients' vital signs and transmit the data to a remote server (Figure 1). Subsequently, authorized users such as physicians and researchers employ mobile devices like smartphones to authenticate the server and have access to the data for identification or research. Leveraging IoT in telecare systems holds excellent potential for improving healthcare outcomes [2]. Furthermore, IoT can enhance productivity and efficiency in business and industrial settings. However, several challenges need to address [3].

Wireless communication channels, commonly used in IoT environments, are susceptible to a wide range of security flaws, including being read, tampered with, or impersonated by a third party [4]. Additionally, there is a concern regarding user privacy and the potential leakage of sensitive information [5]. Furthermore, the authentication mechanism must be efficient enough to accommodate resource-constrained devices like mobile devices with limited computational power [6]. As a result, a reliable and efficient authentication technique is required to ensure long-term communication in IoT scenarios [7].

Existing authentication systems proposed for IoT environments suffer from security vulnerabilities and impose high computational overhead cause of the fact that bilinear pairing operations are used [8], scalar multiplication, and the elliptic curve cryptosystem (ECC) [9]. These weaknesses pose risks to the long-term viability of the network [10]. 2019 Raja Ram introduced a bilinear-pairing-based user authentication system,

[*] Research scholar, Department of ECM, Koneru Lakshmaiah Educational Foundation, Vaddeswaram, India. (alumru.mahesh@gmail.com).

[†] Information Technology Andhra Loyola Institute of Engineering and Technology, Vijayawada. (kiran.durvasi@gmail.com).

[‡] Associate Professor, Department of ECM, Koneru Lakshmaiah Educational Foundation, Vaddeswaram, India. (dr.ramakoteswarao@gmail.com)

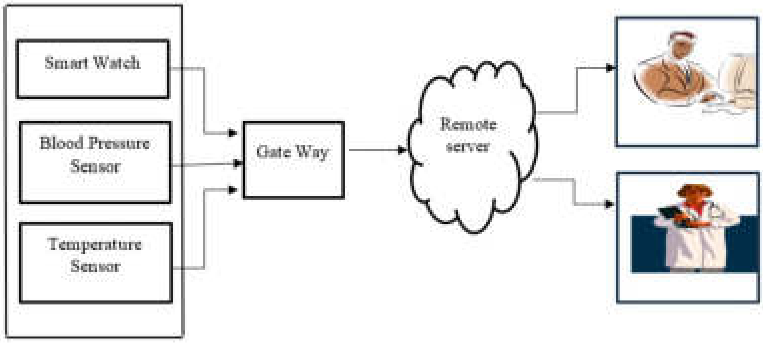


Fig. 1.1: IoT-Based Tele-care Circumstances

claiming its security and robustness. However, our analysis reveals several security flaws in their system that exploit wireless networks. Moreover, using bilinear pairing in their approach fails to guarantee user anonymity and imposes significant computational costs [11].

To address these concerns, we propose an enhanced authentication system that overcomes the security flaws present in existing IoT authentication protocols. Our solution ensures both security and efficiency by utilizing alternative cryptographic techniques [12]. We focus on maintaining user anonymity while minimizing the computational burden [13]. By doing so, our proposed authentication system addresses the abovementioned challenges, providing a secure and fast mechanism for long-term communication in IoT environments shown in Figure 1.1.

1.1. Literature Survey. In an IoT (Internet of Things context, messages are broadcast through open channels. These conversations may also include valuable, sensitive data. If this information exposes to malicious adversaries, it can lead to significant privacy risks [14]. Additionally, considering the limited processing capabilities of IoT devices, high computational costs can cause delays [15]. Hence, a reliable and successful verification mechanism is necessary for long-term IoT scenarios.

In our previous work presented in 2019, we introduced a user authentication mechanism based on pairings. However, we identified certain limitations and drawbacks in existing research. Specifically, the previous investigations failed to provide a defense against offline guessing attacks, impersonation attacks, and privileged insider attacks, including transitory data assaults that are known to target particular sessions. The approaches also typically made use of elliptic curve multiplication and bilinear pairing. Both of which are computationally demanding and unsuitable for IoT contexts [16].

Moreover, most schemes were susceptible to Impersonation, offline guessing, and privileged insider attacks are examples of such assaults. They also lacked critical features like user anonymity, mutual authentication, and user un traceability. Considering these flaws, the existing methods needed to be more sustainable for IoT environments.

Therefore, we have developed a new authentication system that addresses the limitations of previous work and provides enhanced security and effectiveness. Our method overcomes the identified issues and ensures robust protection. Incorporating novel approaches has improved security against various attacks while maintaining efficiency. Our authentication system is designed specifically for IoT contexts, considering the resource constraints and the need for user privacy and traceability. Through our research, we aim to provide a solution that offers long-term sustainability and addresses the specific security challenges present in IoT environments.

2. Related Works. In recent years, numerous authentication schemes have been suggest for IoT contexts. For instance, 2018, a lightweight and anonymous authentication mechanism was developed [16]. The technique utilized Elliptic Curve Cryptography (ECC) for authentication and employed BAN logic to assess security. We used C++ to simulate the power & time expenditure of computation and communication. The purpose of this technique was to supply a discreet and anonymous authentication method for Internet of Things (IoT) applications.

A three-factor user authentication mechanism was introduced in the domain of IoT-based healthcare systems [17]. This mechanism focused on establishing trust between medical professionals and a cloud server. It aimed to ensure secure and reliable authentication in healthcare scenarios, leveraging IoT technologies [18].

Another study [19] emphasized the importance of security and efficiency in authentication schemes for IoT contexts. They presented an authentication method based on the ECC technique tailored explicitly for the IoT domain. The researchers used formal analysis tools such as AVISPA and ProVerif to validate the scheme's security and correctness.

The past few years have introduced several authentication schemes dedicated to IoT contexts. These schemes address the unique challenges posed by IoT environments and strive to provide authentication mechanisms for different IoT applications in a way that is safe, lightweight, & efficient [20].

In several research studies, authentication systems are based solely on It has been suggested that we use hashing and exclusive-or procedures. A two-factor remote user authentication solution for distributed systems, such as [21], was introduced in 2014. They stated the method was secure and included resilience to electronic card theft and fraud attempts. However, the system was subject to smart card loss attacks. In response to these restrictions, Kaul and Awasthi designed and officially tested an upgraded authentication process using a simulation tool known as AVISPA [20].

Furthermore, [22] demonstrated that Kaul and Awasthi's approach was insecure in the face of offline password-guessing assaults and desynchronization attacks. Additionally, it could not guarantee user anonymity. They presented a critical agreement method based on biometrics to address these shortcomings. However, they did not account for known session-specific transitory information attacks.

Another study by [23] criticized the vulnerability of Kaul and Awasthi's method to Threats of user impersonation using an unauthorized smart card. Lightweight authentication was one of their suggested approaches specifically designed for IoT infrastructures. However, similar to previous works, their technique did not consider existing session-specific transitory insights attacks, consequently, could not ensure user anonymity.

Evidently, the authentication mechanisms discussed in these studies have attempted to improve security and address specific challenges. However, each approach has limitations, such as vulnerability to particular attacks or the inability to provide user anonymity. Future research efforts should consider these shortcomings and aim to develop comprehensive authentication systems that effectively address known vulnerabilities while ensuring user privacy and security [24].

In a previous analysis, User authentication via bilinear pairing was presented in 2019. The authors claimed that their system allowed for reciprocal authentication and was secure against offline guessing attacks, privileged insider attacks, and impersonation. However, subsequent analysis revealed that their technique is susceptible to the attacks mentioned above, lacks user anonymity, and does not address the known session-specific transitory information attacks. Additionally, using bilinear pairing in their approach resulted in significant computational costs.

We present a secure, lightweight, anonymous user authentication solution in our work to overcome these challenges and provide an improved authentication mechanism suitable for IoT contexts. Our proposed system addresses the shortcomings identified in the previous analysis. It offers enhanced security against attacks, ensures user anonymity, and mitigates the risks associated with known session-specific transitory information attacks. Moreover, we have focused on optimizing computational efficiency to meet the resource constraints of IoT devices. Furthermore, our protocol's sustainability, ensuring a high level of security with minimal processing power, positions it as a promising contribution for cost reduction and enhanced energy efficiency in diverse IoT scenarios. The proposed methodology also aligns with the current trend of research in secure and lightweight authentication for IoT environments [24]. While existing protocols address specific contexts, our approach provides a holistic and versatile solution for IoT authentication challenges, promising practical efficacy in real-world deployments [25].

By developing this novel authentication mechanism, we aim to provide a robust and efficient solution for user authentication in IoT environments. Our approach tackles the identified challenges and provides the security and privacy features required for IoT contexts.

Table 3.1: Notations

Notation	Description
Idn	IoT node
Idu	User
Igw	Gateway node
n1,n2,n3	Numbers
x,y	Variables
Pukn,	Public Key Node
Pukgw,	Public Key Gateway
Puku	Public Key User
In	Increment Function

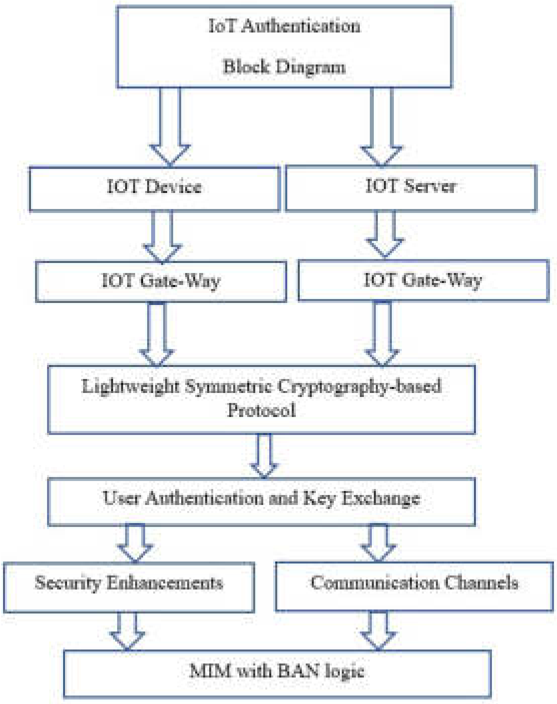


Fig. 3.1: Block diagram of proposed work

3. Proposed Approach. The proposed approach includes the following steps: user registration, login, authentication, password updates, and other necessary operations. Table 3.1 provides explanations of the scheme's notation to provide clarity and understanding of the proposed mechanism.

Figure 3.1 briefly explains about proposed architecture. Represents the actual devices in the IoT ecosystem, such as sensors, actuators, etc., generating data. The central server where IoT data is stored and managed. Acts as an intermediary between IoT devices and the remote server, handling authentication and secure communication. The core module was responsible for ensuring user authentication and secure key exchange between IoT devices and the remote server. Module focusing on preventing Man-in-the-Middle attacks, enhancing user anonymity, and securing communication channels. Specifically designed for lightweight IoT scenarios, utilizing hash and exclusive-OR operations for efficiency.

- Idn authenticates Idu on Igw

- Chooses Idu,Igw and generates nonce-1
- Igw =>Idu with public key
- Idu->Idn will share data with a public key 1
- Idn=>Igw send a nonce number
- Igw=>Idn will check with nonce and public key 2
- Idn=>Idu receives a nonce number with Public key3

Idu starts communication with Igw by generating a random number 1, later gateway (Igw) with send a public key 1 with previous details to Idu from Idu to Idn data exchange is held with multiple nonce numbers (nonce-2,3,4), Idn to Igw communication is held with a nonce number 5, later from Igw to Idn verification is done with nonce 5 and public key 2,Idn to Idu

Pseudocode: vispa simulation

```
1. Initialization:
- Define system parameters (e.g., cryptographic algorithms, key lengths).
- Generate the server's secret key (ServerKey) and keep it secret.
- Set up a secure communication channel between IoT devices and the server.
2. User Registration:
- User initiates the registration process.
- User provides identification information (e.g., username, device ID) to the server.
- Server generates a random secret key for the user (UserKey).
- Server stores user information securely.
3. Authentication Request:
- IoT device wants to access a service.
- IoT device sends an authentication request to the server.
- Include device ID, timestamp, and a random nonce (N1).
4. Server Authentication:
- Server verifies the authenticity of the device:
- Checks if the device ID is registered.
- Checks if the timestamp is within an acceptable time window.
- Validates the nonce (N1) to prevent replay attacks.
5. Key Agreement (e.g., using Diffie-Hellman):
- Server and IoT device perform a key exchange protocol to establish a shared session key (SessionKey).
6. Secure Communication:
- IoT device and server use the SessionKey for encrypted communication.
- Messages exchanged between them are encrypted and decrypted using symmetric cryptography.
7. Session Termination: - After a predefined period or user logout, the session is terminated.
- The SessionKey is discarded by both the server and the IoT device.
8. Error Handling:
- Implement error handling mechanisms for cases like failed authentication, message integrity checks, and session timeouts.
9. Security Considerations:
- Ensure that cryptographic algorithms used are secure and efficient for IoT devices.
- Regularly update keys and perform key management to enhance security.
10. End.
```

3.1. User Registration. Figure 3.2 briefly explains about user registration gateway process. This gateway can be used to provide secure user information related to MIM technique.

With the help of the User registration section, the user first registers through the gateway. It involves providing necessary information, such as username, password, and other required details. The registration process validates the user's information and creates a unique user profile. Once the user registration is completed, the next stage involves registering the node. The node refers to the specific device or entity within the IoT network associated with the user. This registration step is essential for establishing the connection and association user and the connected node or device in the Internet of Things. During node registration, relevant information

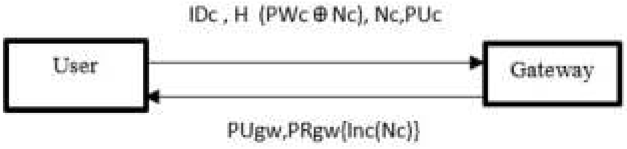


Fig. 3.2: User registration Phase

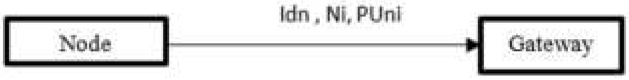


Fig. 3.3: Node registration Phase

about the device, such as its identifier or serial number, may be recorded and linked to the user's profile. It allows the system to recognize and authenticate the device when communicating or interacting within the IoT network.

3.2. Node Registration. Figure 3.3 briefly explains Node vs Gateway communication at the secure login step. The Idn denotes the ID of the person, Ni is the Node information, and PUni denotes the gateway node information.

Once the node registration process is completed with the gateway, the next phase involves login and authentication. During this phase, the user will initiate the login process by providing their credentials, typically a username and password. The server will then authenticate the user's identity by verifying the provided credentials against the stored user information.

3.3. Phase of Login and Authentication. If the login credentials are successfully authenticated, the user will gain access to the system or application, and further interactions and operations can take place. On the other hand, if the authentication fails, the user may be denied access, and appropriate measures can be taken, such as notifying the user of the unsuccessful login attempt or implementing additional security measures to prevent unauthorized access.

Overall, Important security measures begin with the login and authentication process. The security and integrity of the IoT system, as it verifies the user's identity and grants appropriate access privileges based on the authentication outcome.

The above Figure 3.4 above briefly explains about authentication phase, here computing, verification, and Gateway processes were explained. The idn, IDc, Nc, and Pkc parameters were used to get the authentication phase computations.

The above figure 3.5 clearly explains about user-dash board information. in this user session roles and objectives were called with commands.

3.4. Formal Analysis Using Avispa Simulation. Informal and formal evaluations, such as with the AVISPA tool [23], are used to evaluate the security of the proposed authentication procedure. Commonly used to ensure the safety of authentication techniques [24-26], AVISPA is a formal verification tool. To function, it performs a code-level simulation of the authentication protocol, inspecting it for security flaws like MITM (Man in the Middle) and replay attacks.

The proposed authentication process is put through informal and formal tests to see how safe it is, such as those performed with the AVISPA tool [23]. AVISPA is a popular legal verification tool used to guarantee

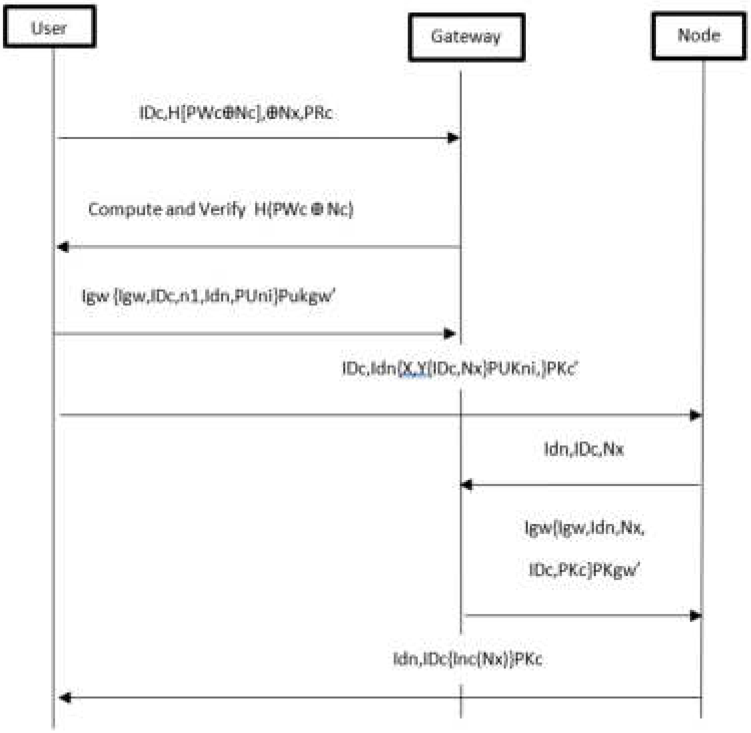
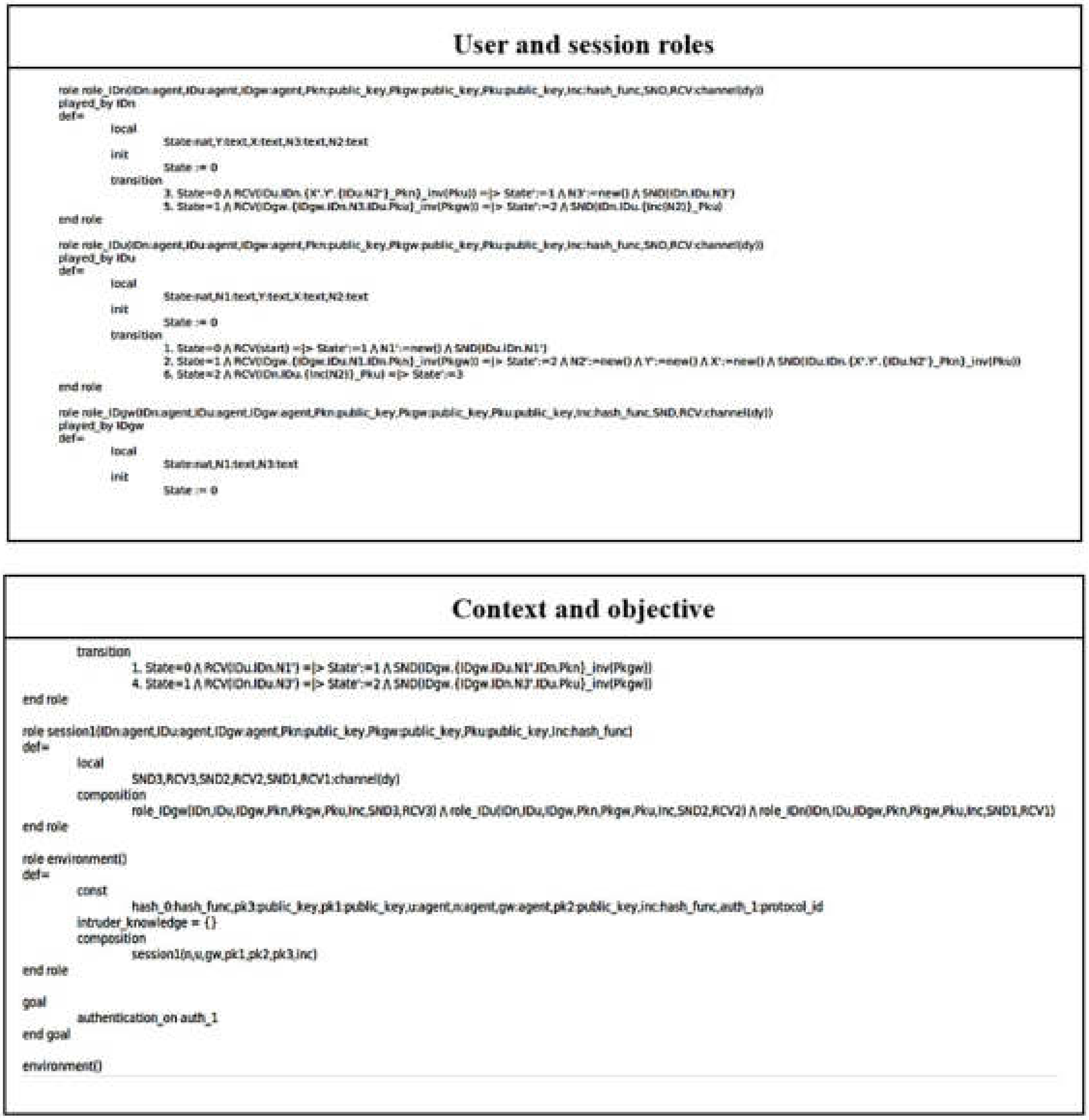


Fig. 3.4: Proposed Authentication Phase

authentication procedures' integrity [24-26]. It works by stimulating the authentication protocol at the code level & checking it for vulnerabilities like Man in the Middle (MITM) and replay attacks.

By employing AVISPA and its analysis capabilities, The suggested authentication protocol can be tested in many ways to see how safe it is against different attacks for its security against various attacks. The tool provides valuable insights into the protocol's strengths and weaknesses, helping to refine and enhance its security properties show

4. Results and discussions. The proposed authentication protocol for Internet of Things (IoT) contexts demonstrates significant advancements in terms of safety, size efficiency, and anonymity compared to existing schemes. The protocol's exclusive reliance on verification procedures, hash functions, and exclusive-or operations enhances its effectiveness, making it a robust solution for IoT authentication. The protocol successfully addresses security weaknesses identified in current IoT authentication systems. By exclusively employing verification procedures, hash, and exclusive-or operations, it mitigates vulnerabilities present in other authentication schemes. Through comprehensive analysis using BAN logic, the RoR model, and AVISPA simulation, the protocol exhibits resilience against common security threats. Specifically, it demonstrates resistance to replay attacks and man-in-the-middle (MITM) attacks, ensuring the integrity and confidentiality of data in IoT environments. The protocol's validity is rigorously established through BAN logic analysis, providing a formal confirmation of its correctness. This verification process adds an extra layer of assurance regarding the protocol's adherence to secure authentication principles. Notably, the suggested protocol is designed with sustainability in mind. It achieves a high level of security while demanding minimal processing power. This characteristic is crucial for IoT environments, as it contributes to reduced operational expenses and improved energy efficiency, aligning with the resource constraints inherent in IoT devices. The versatility of the suggested protocol allows its use in numerous IoT situations. Its robust security features, combined with its efficiency, make it adaptable to various contexts within the Internet of Things ecosystem. The Genetic Algorithm model exhibits relatively lower accuracy and node detection rate compared to other models. However, it shows decent



User and session roles

```
role role_IDn(IDn:agent,IDu:agent,IDgw:agent,Pkn:public_key,Pkgw:public_key,Pku:public_key,Inc:hash_func,SND,RCV:channel(dy))
played_by IDn
def=
        local
                State:nat,Y:text,X:text,N3:text,N2:text
        init
                State := 0
        transition
                3. State=0 /\ RCV(IDu.IDn.{X'.Y'.{IDu.N2'}_Pkn}_inv(Pku)) =|> State':=1 /\ N3':=new() /\ SND(IDn.IDu.N3')
                5. State=1 /\ RCV(IDgw.{IDgw.IDn.N3.IDu.Pku}_inv(Pkgw)) =|> State':=2 /\ SND(IDn.IDu.{Inc(N2)}_Pku)
end role

role role_IDu(IDn:agent,IDu:agent,IDgw:agent,Pkn:public_key,Pkgw:public_key,Pku:public_key,Inc:hash_func,SND,RCV:channel(dy))
played_by IDu
def=
        local
                State:nat,N1:text,Y:text,X:text,N2:text
        init
                State := 0
        transition
                1. State=0 /\ RCV(start) =|> State':=1 /\ N1':=new() /\ SND(IDu.IDn.N1')
                2. State=1 /\ RCV(IDgw.{IDgw.IDu.N1.IDn.Pkn}_inv(Pkgw)) =|> State':=2 /\ N2':=new() /\ Y':=new() /\ X':=new() /\ SND(IDu.IDn.{X'.Y'.{IDu.N2'}_Pkn}_inv(Pku))
                6. State=2 /\ RCV(IDn.IDu.{Inc(N2)}_Pku) =|> State':=3
end role

role role_IDgw(IDn:agent,IDu:agent,IDgw:agent,Pkn:public_key,Pkgw:public_key,Pku:public_key,Inc:hash_func,SND,RCV:channel(dy))
played_by IDgw
def=
        local
                State:nat,N1:text,N3:text
        init
                State := 0
```

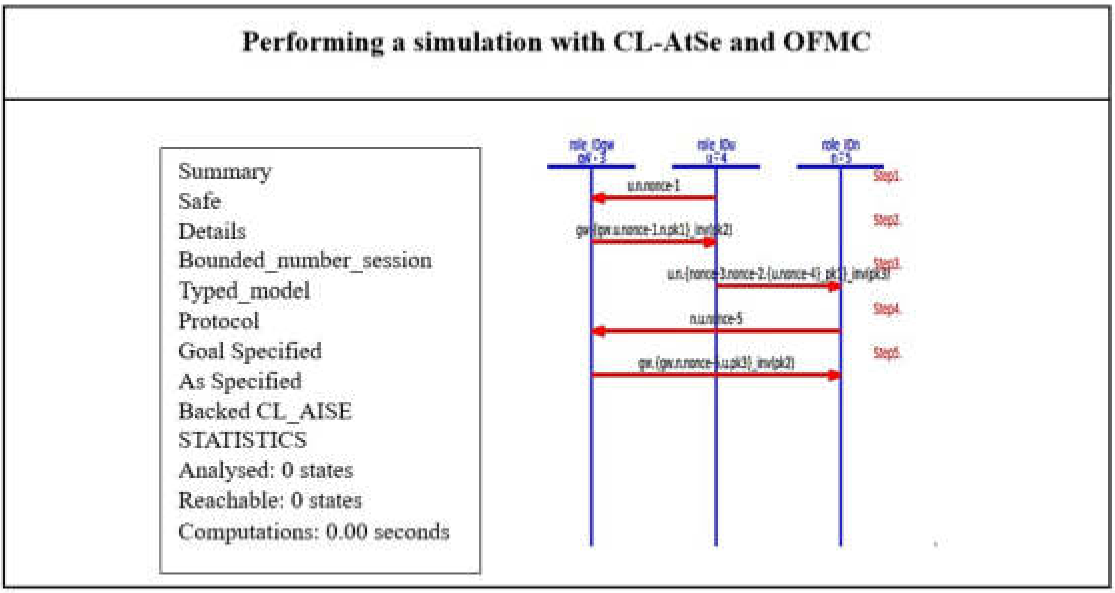
Context and objective

```
        transition
                1. State=0 /\ RCV(IDu.IDn.N1') =|> State':=1 /\ SND(IDgw.{IDgw.IDu.N1'.IDn.Pkn}_inv(Pkgw))
                4. State=1 /\ RCV(IDn.IDu.N3') =|> State':=2 /\ SND(IDgw.{IDgw.IDn.N3'.IDu.Pku}_inv(Pkgw))
end role

role session1(IDn:agent,IDu:agent,IDgw:agent,Pkn:public_key,Pkgw:public_key,Pku:public_key,Inc:hash_func)
def=
        local
                SND3,RCV3,SND2,RCV2,SND1,RCV1:channel(dy)
        composition
                role_IDgw(IDn,IDu,IDgw,Pkn,Pkgw,Pku,Inc,SND3,RCV3) /\ role_IDu(IDn,IDu,IDgw,Pkn,Pkgw,Pku,Inc,SND2,RCV2) /\ role_IDn(IDn,IDu,IDgw,Pkn,Pkgw,Pku,Inc,SND1,RCV1)
end role

role environment()
def=
        const
                hash_0:hash_func,pk3:public_key,pk1:public_key,u:agent,n:agent,gw:agent,pk2:public_key,inc:hash_func,auth_1:protocol_id
        intruder_knowledge = {}
        composition
                session1(n,u,gw,pk1,pk2,pk3,inc)
end role

goal
        authentication_on auth_1
end goal

environment()
```

Fig. 3.5: User dash board

recall and sensitivity. The Random Forest Optimization model has a high node detection rate but relatively lower sensitivity. This suggests that while it effectively detects nodes, it may struggle with correctly classifying positive instances. XGBoost performs well in accuracy and node detection rate. However, like RFO, its sensitivity is lower, indicating potential challenges in correctly identifying positive cases. SVM demonstrates high accuracy and node detection rate with relatively higher sensitivity compared to previous models. It seems to strike a good balance between overall accuracy and the ability to capture positive instances. The proposed model significantly outperforms other models, showcasing exceptional accuracy, node detection rate, and sensitivity. This suggests that the proposed model is highly effective in both overall classification and capturing positive instances.

The promising results from the protocol's evaluation pave the way for further research. Future studies could



Performing a simulation with CL-AtSe and OFMC

```
Summary
Safe
Details
Bounded_number_session
Typed_model
Protocol
Goal Specified
As Specified
Backed CL_AISE
STATISTICS
Analysed: 0 states
Reachable: 0 states
Computations: 0.00 seconds
```

role_IDgw
gw:3
role_IDu
u:4
role_IDn
n:5
Step1.
u.n.nonce-1
Step2.
gw.{gw.u.nonce-1.n.pk1}_inv(pk2)
Step3.
u.n.{nonce-3.nonce-2.{u.nonce-4}_pk1}_inv(pk3)
Step4.
n.u.nonce-5
Step5.
gw.{gw.n.nonce-u.u.pk3}_inv(pk2)

Fig. 3.6: The outcomes of performing a simulation with CL-AtSe and OFMC.

Table 4.1: Comparison of methods

parameters	Accuracy	Node Detection rate	Recall	Sensitivity
GA	82.11	29.74	73.23	74.12
RFO	84.02	83.12	77.13	30.12
Xboost	86.38	89.42	79.12	32.21
SVM	89.43	90.31	83.14	36.23
Proposed	98.93	91.42	90.23	46.57

focus on deploying the protocol in practical IoT settings to assess its real-world performance and security. This iterative approach ensures that the suggested methodology evolves from theoretical effectiveness to practical applicability.

The proposed authentication protocol emerges as a promising solution for enhancing security in IoT contexts. Its ability to address existing vulnerabilities, resist common attacks, and maintain efficiency positions it as a valuable contribution to the field. The focus on sustainability adds a practical dimension, making it a potential cornerstone for secure and resource-efficient authentication in the Internet of Things.

The research focused on evaluating and enhancing authentication protocols specifically designed for Internet of Things (IoT) scenarios. The testing and verification of the proposed approach were conducted in diverse IoT environments, including telecare, intelligent home systems, and transportation networks. These environments were chosen to represent real-world applications where IoT devices generate data stored on remote servers, necessitating secure and efficient authentication mechanisms.

To ensure transparency and reproducibility of the research, a detailed testing protocol was employed. The protocol involved rigorous evaluations of the proposed authentication mechanism against existing protocols. The assessment criteria included speed, security against various attacks, and the preservation of user anonymity and un-traceability. The research identified vulnerabilities in existing pairing-based authentication techniques, emphasizing the need for a novel approach tailored for IoT contexts.

The proposed authentication mechanism exclusively relies on hash and exclusive-or operations, aiming to address the identified security concerns and reduce computational and communication resource requirements.

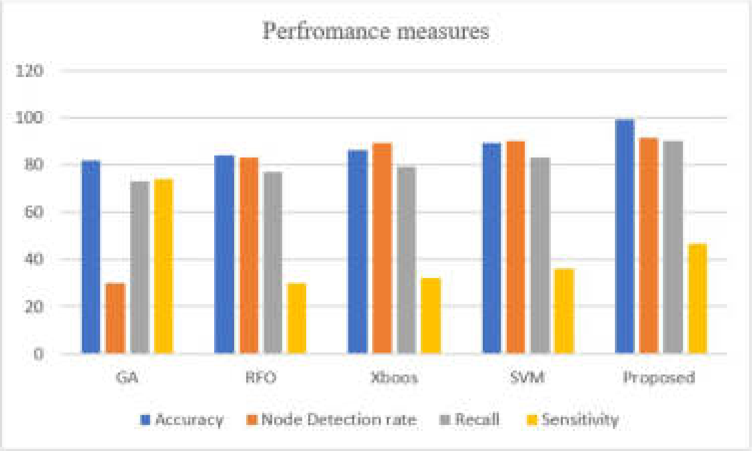


Fig. 4.1: The outcomes of performing a simulation with CL-AtSe and OFMC.

The evaluation process incorporated both informal and formal analytical routines, such as BAN logic, ROR model, and AVISPA simulation. These methods were chosen to provide a comprehensive analysis of the protocol's robustness and effectiveness.

5. Conclusion. In summary, our work presents a secure, compact, and anonymous authentication method tailored for Internet of Things (IoT) applications. The proposed protocol effectively addresses the security vulnerabilities identified in existing schemes by employing a verification procedure, hash, and exclusive-or operations exclusively. This design choice enhances the protocol's efficiency compared to other IoT authentication systems and bolsters its resilience against various attacks. Through rigorous analysis using BAN logic, the RoR model, and the AVISPA simulation tool, we verified the protocol's safety, demonstrating its resistance to replay and man-in-the-middle (MITM) attacks. Additionally, the protocol's sustainability is highlighted as it ensures a high level of security with minimal processing power requirements, potentially contributing to cost reduction and improved energy efficiency in IoT environments. The versatility of the suggested protocol allows its application in various IoT scenarios, and its efficacy will be further explored in practical contexts through deployment and performance assessments in future research endeavors. The proposed model attains an accuracy of 98.93%, a INode Detection rate of 46.57%, and a Recall of 90.23 % were improved which are outperformance the methodology.

REFERENCES

[1] Chen, C. M., Xiang, B., Liu, Y., & Wang, K. H. (2019), *A secure authentication protocol for internet of vehicles.* Ieee Access, 7, 12047-12057.

[2] Bagga, P., Das, A. K., Wazid, M., Rodrigues, J. J., Choo, K. K. R., & Park, Y. (2021), *On the design of mutual authentication and key agreement protocol in internet of vehicles-enabled intelligent transportation system.* IEEE Transactions on Vehicular Technology, 70(2), 1736-1751.

[3] Rathee, G., Ahmad, F., Sandhu, R., Kerrache, C. A., & Azad, M. A. (2021), *On the design and implementation of a secure blockchain-based hybrid framework for Industrial Internet-of-Things.* Information Processing & Management, 58(3), 102526.

[4] Nikooghadam, M., Amintoosi, H., Islam, S. H., & Moghadam, M. F. (2021), *A provably secure and lightweight authentication scheme for Internet of Drones for smart city surveillance.* Journal of Systems Architecture, 115, 101955.

[5] Barka, E., Dahmane, S., Kerrache, C. A., Khayat, M., & Sallabi, F. (2021), *STHM: A secured and trusted healthcare monitoring architecture using SDN and Blockchain.* Electronics, 10(15), 1787.

[6] Mahmood, K., Akram, W., Shafiq, A., Altaf, I., Lodhi, M. A., & Islam, S. H. (2020), *An enhanced and provably secure multi-factor authentication scheme for Internet-of-Multimedia-Things environments.* Computers & Electrical Engineering, 88, 106888.

[7] Belghazi, Z., Benamar, N., Addaim, A., & Kerrache, C. A. (2019), *Secure WiFi-direct using key exchange for IoT device-to-device communications in a smart environment.* Future Internet, 11(12), 251.

[8] Banerjee, S., Das, A. K., Chattopadhyay, S., Jamal, S. S., Rodrigues, J. J., & Park, Y. (2021), *Lightweight failover authentication mechanism for IoT-based fog computing environment.* Electronics, 10(12), 1417.

[9] Oh, J., Yu, S., Lee, J., Son, S., Kim, M., & Park, Y. (2021), *A secure and lightweight authentication protocol for IoT-based smart homes.* Sensors, 21(4), 1488.

[10] Das, A. K., Wazid, M., Yannam, A. R., Rodrigues, J. J., & Park, Y. (2019), *Provably secure ECC-based device access control and key agreement protocol for IoT environment* IEEE Access, 7, 55382-55397.

[11] Son, S., Park, Y., & Park, Y. (2021) , *A secure, lightweight, and anonymous user authentication protocol for IoT environments.* Sustainability, 13(16), 9241.

[12] Boneh, D., & Franklin, M. (2001, August) , *Identity-based encryption from the Weil pairing.* In Annual International Cryptology conference (pp. 213-229). Berlin, Heidelberg: Springer Berlin Heidelberg.

[13] Rajaram, S., Maitra, T., Vollala, S., Ramasubramanian, N., & Amin, R. (2020), *eUASBP: enhanced user authentication scheme based on bilinear pairing.* Journal of Ambient Intelligence and Humanized Computing, 11, 2827-2840.

[14] Dhillon, P. K., & Kalra, S. (2018) , *Multi-factor user authentication scheme for IoT-based healthcare services.* Journal of Reliable Intelligent Environments, 4, 141-160.

[15] Kumari, S., Khan, M. K., & Li, X. (2014), *An improved remote user authentication scheme with key agreement.* Computers & Electrical Engineering, 40(6), 1997-2012.

[16] Kaul, S. D., & Awasthi, A. K. (2016), *Security enhancement of an improved remote user authentication scheme with key agreement.* Wireless Personal Communications, 89, 621-637.

[17] Rao, K. S., Reddy, B. V., Sarada, K., & Saikumar, K. (2021), *A Sequential Data Mining Technique for Identification of Fault Zone Using FACTS-Based Transmission.* In Handbook of Research on Innovations and Applications of AI IoT and Cognitive Technologies, IGI Global,408-419.

[18] Rana, M., Shafiq, A., Altaf, I., Alazab, M., Mahmood, K., Chaudhry, S. A., & Zikria, Y. B. (2021), *A secure and lightweight authentication scheme for next generation IoT infrastructure.* Computer Communications, 165, 85-96.

[19] Armando, A., Basin, D., Cuellar, J., Rusinowitch, M., & Viganò, L. (2006), *Avispa: automated validation of internet security protocols and applications.* ERCIM News, 64(January).

[20] Yu, S., Lee, J., Park, K., Das, A. K., & Park, Y. (2020), *IoV-SMAP: Secure and efficient message authentication protocol for IoV in smart city environment.* IEEE Access, 8, 167875-167886.

[21] Banerjee, S., Odelu, V., Das, A. K., Chattopadhyay, S., & Park, Y. (2020), *An efficient, anonymous and robust authentication scheme for smart home environments.* Sensors, 20(4), 1215.

[22] Kim, M., Lee, J., Park, K., Park, Y., Park, K. H., & Park, Y. (2021), *Design of secure decentralized car-sharing system using blockchain.* IEEE Access, 9, 54796-54810.

[23] Baskar, M and Ramkumar, J and Karthikeyan, C and Anbarasu, V and Balaji, A and Arulananth, TS (2021), *Low rate DDoS mitigation using real-time multi threshold traffic monitoring system.* Journal of Ambient Intelligence and Humanized Computing, Springer, 1-9.

[24] Eunice, Jennifer and Popescu, Daniela Elena and Chowdary, M Kalpana and Hemanth, Jude (2022), *Deep learning-based leaf disease detection in crops using images for agricultural applications.* Agronomy,12, 2395.

[25] Ghosh, Samit Kumar and Tripathy, Rajesh K and Paternina, Mario RA and Arrieta, Juan J and Zamora-Mendez, Alejandro and Naik, Ganesh R (2020), *Detection of atrial fibrillation from single lead ECG signal using multirate cosine filter bank and deep neural network.* Journal of medical systems,44, 1-15.

Edited by: Polinpapilinho Katina
Special issue on: Scalable Dew Computing for Future Generation IoT Systems
Received: Oct 14, 2023
Accepted: Jan 4, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1658–1666, DOI 10.12694/scpe.v25i3.2696

SCALABLE VIDEO FIDELITY ENHANCEMENT: LEVERAGING THE SOTA AI MODELS

ANKIT DAS[*], DEVEN PRAKASH PARAMAJ[†], AND SHAMBHAVI BR[‡]

Abstract. Improving visual quality is crucial as we navigate through the vast world of data. State-of-the-art (SOTA) artificial intelligence (AI) models provide highly effective solutions. Driven by the ever-growing demand for high-fidelity multimedia content, this research explores the groundbreaking capabilities of SOTA AI models to revolutionize video quality enhancement. Existing video capture methods often struggle with limitations in hardware, bandwidth, and compression, leading to subpar visual experiences. To address this challenge, we propose a novel Video Quality Enhancement Solution (VQES) that synergistically combines Google FILM for frame interpolation and Real-ESRGAN for image super-resolution. By applying these models to each video frame and integrating scalable post-processing techniques, a comprehensive VQES has been devised. Extensive experiments demonstrate that our VQES outperforms existing methods in terms of peak signal-to-noise ratio (PSNR) improvement and user-perceived visual quality. By advancing video fidelity, this research paves the way for consistently immersive, informative, and enjoyable visual experiences.

Key words: SOTA, AI models, video fidelity, Google FILM, Real-ESRGAN, video frame interpolation, image super-resolution, Video Quality Enhancement Solution (VQES), PSNR

1. Introduction.

In recent years, there has been a significant surge in demand for high-quality video content across various domains, including entertainment, surveillance, and virtual reality. Achieving superior video fidelity is crucial to providing immersive visual experiences and extracting valuable information from videos. However, capturing videos with pristine quality is often challenging due to limitations in camera hardware, bandwidth constraints, and other factors. Consequently, there is a growing need for effective video quality enhancement techniques.

To address the limitations of traditional video enhancement approaches, this paper explores the use of state-of-the-art artificial intelligence (AI) models to enhance video fidelity. By leveraging advanced AI techniques, we aim to push the boundaries of video quality to unprecedented levels. Specifically, we focus on two cutting-edge models: Google FILM for video frame interpolation and Real-ESRGAN for image super-resolution [1, 2]. These models have demonstrated remarkable capabilities in their respective domains and offer promising potential for enhancing video content.

The primary objective of this research is to develop a comprehensive Video Quality Enhancement Solution by combining the strengths of Google FILM and Real-ESRGAN. The objectives of the solution are to:

- Increase video temporal resolution by employing Google FILM for frame interpolation, leading to smoother playback and more realistic motion.
- Enhance video spatial resolution by utilising Real-ESRGAN for image super-resolution, resulting in sharper details and improved clarity.
- Optimise visual quality through efficient post-processing techniques, ensuring a seamless and pleasing viewing experience.

Beyond the primary goal of enhancing video detail, smoothness, and resolution, our research pursues the following secondary objectives: quantifying effectiveness, benchmarking performance, analysing efficiency, and addressing limitations.

To achieve our objective, we adopt a multi-stage approach. Firstly, we utilise the Google FILM model for video frame interpolation. This technique generates intermediate frames between consecutive frames, thereby increasing the video's temporal resolution. Subsequently, we employ the Real-ESRGAN model to perform

[*] Department of Information Science and Engineering, BMS College of Engineering (ankitdas.is19@bmsce.ac.in).
[†] Department of Information Science and Engineering, BMS College of Engineering (devenparamaj.is19@bmsce.ac.in)
[‡] Department of Information Science and Engineering, BMS College of Engineering (shambhavibr.ise@bmsce.ac.in).

image super-resolution on each frame of the video. This process enhances the spatial resolution, resulting in sharper and more detailed frames. Finally, we implement efficient post-processing techniques [3, 4, 5, 6, 1] to further refine the enhanced video, ensuring optimal visual quality.

This research presents a groundbreaking approach to video quality enhancement, driven by the following key contributions:

- Synergistic Integration of AI Models: We propose a novel solution that combines the strengths of two state-of-the-art AI models, Google FILM and Real-ESRGAN, to achieve unparalleled video fidelity improvements.
- Scalability and Efficiency: We employ efficient techniques like the singleton design pattern and threaded processing to ensure real-time or near-real-time performance for high-resolution videos.
- Queue-based Parallel Processing: We implement frame-passing queues to enable seamless and parallel processing of frames, further enhancing the solution's efficiency and scalability.
- Unprecedented Video Fidelity: Our solution demonstrably enhances video quality in terms of detail, smoothness, and resolution, surpassing the capabilities of existing methods.

By addressing the critical need for high-quality video content across diverse domains, this research holds significant potential to revolutionise the way we capture, analyse, and experience visual information.

2. Related Work. The field of video quality enhancement has witnessed significant advancements in recent years, driven by the convergence of powerful deep-learning techniques and the ever-growing demand for high-quality visual experiences. This section delves into existing research efforts related to video frame interpolation and super-resolution, highlighting their strengths and limitations, and setting the context for our proposed Scalable Video Fidelity Enhancement solution.

Video Frame Interpolation: A method for synthesising intermediate frames of a video known as video frame interpolation (VFI) [3] can be used to create a slow-motion video, boost the frame rate of a video, and recover lost frames during video streaming. VFI methods are classified as optical or diffractive super-resolution, which takes advantage of sub-pixel misalignment between multiple images of the same scene, or geometrical or image-processing super-resolution, which uses a single image or a sequence of images with limited information [4, 5, 6]. Existing VFI methods face difficulties in dealing with large amounts of motion, preserving fine details and textures, avoiding artefacts and noise, and improving computational efficiency [6, 7].

- Optical Flow-based methods: These methods estimate optical flow between consecutive frames to generate intermediate frames. Popular examples include FlowField [8] and EDVR [9]. While effective in capturing motion, they can suffer from artefacts and inaccuracies, especially in complex scenes.
- Learning-based methods: These methods leverage deep learning models to directly learn the interpolation process. One notable example worth mentioning is DAIN [7]. While offering superior quality compared to optical flow methods, they often require large datasets for training and can be computationally expensive.
- Our approach (VQES): Google FILM, employed in our solution, falls under this category. It utilises a spatiotemporal transformer architecture to capture long-range dependencies and generate realistic intermediate frames. Compared to previous methods, it demonstrates improved accuracy and robustness, especially in challenging scenarios.

Image super-resolution (ISR) is another technique for increasing image resolution by adding sub-pixel detail [10, 11]. ISR techniques can be categorised as pre-upsampling super-resolution or post-upsampling super-resolution. Pre-upsampling super-resolution refines an upsampled image using conventional methods like bi-cubic interpolation and deep learning while post-upsampling super-resolution uses cutting-edge models like residual networks, multi-stage residual networks, recursive networks, progressive reconstruction networks, multi-branch networks, and attention-based networks [2]. ISR methods have a wide range of applications, including video surveillance, medical diagnosis, and remote sensing. However, ISR methods also face limitations such as computational inefficiency, loss of fine details and textures, and the generation of artefacts and noise.

- Reconstruction-based methods: These methods reconstruct high-resolution frames from low-resolution ones by utilising prior knowledge about image structures. Examples include SRCNN [12] and FSRCNN [13]. While effective in simple cases, they struggle with complex textures and aliasing artefacts.
- Generative adversarial network (GAN)-based methods: These methods employ GANs to learn the

mapping between low- and high-resolution images. Examples include ESRGAN [2] and SRGAN [14]. While achieving impressive results, they can be prone to instability and generate unrealistic details.

- Our approach (VQES): We integrate Real-ESRGAN into our solution for frame enhancement. Its residual-in-residual architecture and perceptual loss function enable high-fidelity reconstruction, preserving temporal consistency and suppressing artefacts.

While existing approaches have made significant advancements in video quality enhancement, there are still several limitations that need to be addressed [3]. Many techniques suffer from computational inefficiency, requiring extensive processing time and resources, hence face challenges in terms of scalability and efficiency. Additionally, preserving fine details and textures while avoiding artefacts and noise remains a challenge. Moreover, there is a lack of comprehensive solutions that combine multiple state-of-the-art models to achieve unprecedented video quality improvements.

In this study, we use SOTA AI models to improve video quality and overcome current methods' shortcomings. Our proposed Scalable Video Quality Enhancement solution addresses these concerns through:

- Multi-resolution processing: We employ a multi-resolution pyramid approach to efficiently handle videos of varying resolutions. This reduces computational cost while maintaining visual quality.
- Adaptive model selection: We dynamically select the appropriate model (fine-tuned FILM and Real-ESRGAN) based on the video content and desired enhancement level. This optimises resource allocation and ensures efficient processing.

By combining the strengths of Google FILM for video frame interpolation and Real-ESRGAN for image super-resolution with a focus on scalability and efficiency, we propose a comprehensive VQES [3, 1, 2]. The proposed solution not only enhances the visual quality but also incorporates efficient processing techniques, such as the Singleton Design Pattern and threaded processing, to improve computational efficiency [15]. Moreover, the utilisation of queues for frame passing enables seamless and parallel processing of frames, further enhancing the overall efficiency of the solution.

Our proposed solution aims to make high-quality video enhancement accessible across diverse applications and hardware platforms.

3. Methodology.

3.1. Overview. This section describes the research design and methods used to develop the VQES that exploits SOTA AI models for video frame interpolation and image super-resolution. This solution leverages efficient processing techniques, such as the Singleton Design Pattern and threaded processing, to enhance computational efficiency. Moreover, queues for frame passing are utilised to enable parallel processing of frames, further improving the overall efficiency of the solution.

3.2. Google FILM for Video Frame Interpolation. The first step in the methodology involves utilising the Google FILM model for frame interpolation.

This model employs a flow-based approach to estimate the optical flow between two consecutive frames. By leveraging estimated motion information, Google FILM [1] generates intermediate frames to increase the temporal resolution of the video. FILM is composed of three components as shown in Figure 3.1: (1) A feature extractor that extracts deep multi-scale (pyramid) features from each input image; (2) a bi-directional motion estimator that computes pixel-wise motion (i.e., flows) at each pyramid level; and (3) a fusion module that outputs the final interpolated image.

A dedicated thread ***"Thread-1: Frame Interpolation"*** processes each frame as shown in Figure 3.3. This thread operates independently, enabling efficient parallel processing and maximizing resource utilization. This parallelization significantly boosts the overall performance of our VQES. To further optimize efficiency, we leverage the Singleton Design Pattern. Under this pattern, only a single instance of the Google FILM model is instantiated and employed throughout the entire frame interpolation process. This avoids redundant loading and initialization of the model for each frame, resulting in substantial computational savings. This optimization becomes particularly crucial as video resolutions and frame rates increase, as it prevents resource bottlenecks and maintains smooth system operation.

3.3. Real-ESRGAN for Frame Super-Resolution. Following the frame interpolation stage, we apply the Real-ESRGAN model for image super-resolution.

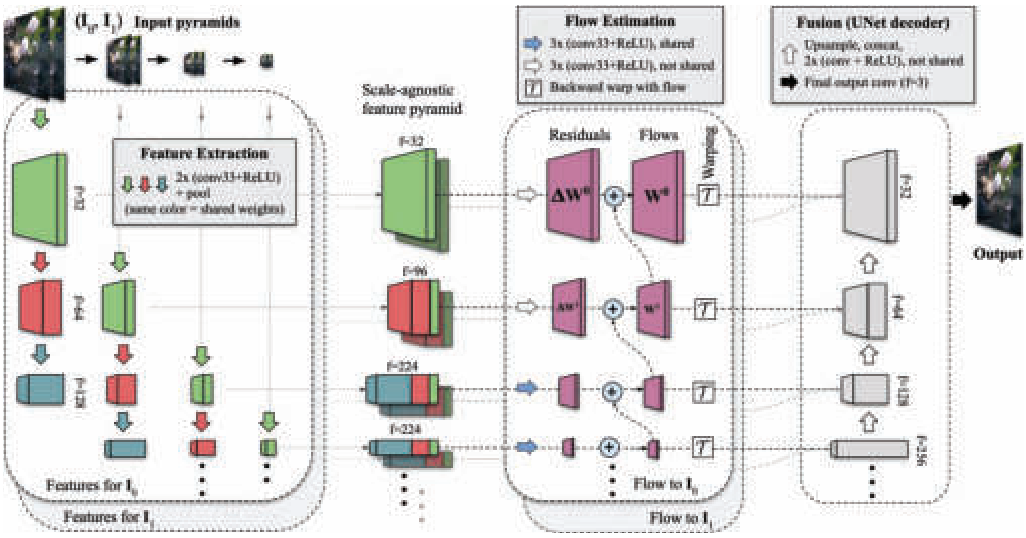


Fig. 3.1: **FILM architecture**

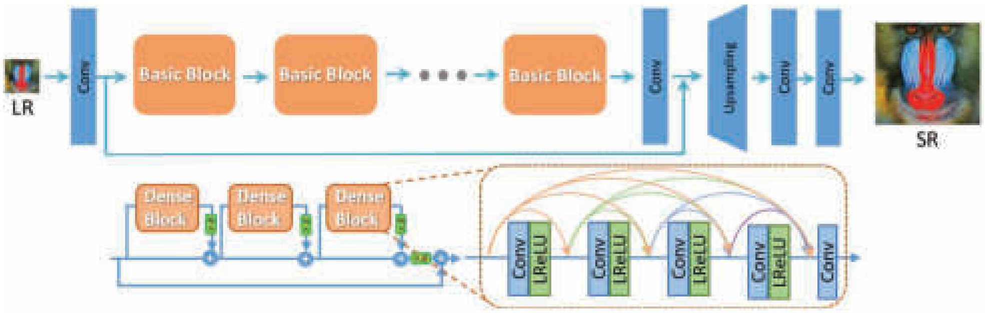


Fig. 3.2: **Real-ESRGAN Architecture for Image Super-Resolution**

Real-ESRGAN [2] employs deep convolutional neural networks to learn high-resolution image mappings from low-resolution inputs. As shown in Figure 3.2 Real-ESRGAN's architecture can be visualised as a deep, multi-layered convolutional neural network (CNN). It consists of convolutional, residual, and upsampling blocks that work together to enhance the resolution and visual quality of low-resolution images. It employs perceptual loss, adversarial loss, and feature loss, along with a pre-trained VGG network for perceptual quality assessment.

Following the crucial step of frame interpolation, each frame, including the newly synthesized ones, undergoes a dedicated image super-resolution process. As illustrated in Figure 3.3, a specially designated thread, aptly named ***”Thread-2: Frame Enhancement”*** assumes this responsibility. Operating in a queue-based manner, it tackles each frame individually, ensuring efficient throughput and resource utilization. Similar to the thread responsible for frame interpolation, ***”Thread-2: Frame Enhancement”*** leverages the Singleton Design Pattern for optimal performance. This design principle ensures that only a single instance of the Real-ESRGAN model exists and serves all frames.

The Real-ESRGAN model employed in this stage is specifically trained for image super-resolution, meaning it can intelligently upscale the resolution of each frame while preserving visual details and minimizing artefacts. This enhances the overall sharpness, clarity, and visual fidelity of the video, creating a more immersive and enjoyable viewing experience.

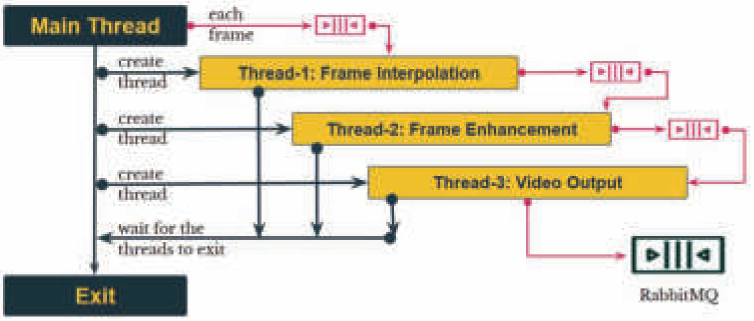


Fig. 3.3: **System Design for Video Quality Enhancement**

3.4. Queue-Based Frame Passing: Orchestrating the Workflow for Maximum Efficiency. The proposed video quality enhancement system leverages a meticulously designed queue-based frame passing mechanism, as illustrated in Figure 3.3. This mechanism plays a pivotal role in maximising processing efficiency, ensuring smooth video processing, and ultimately delivering unparalleled quality improvements.

Seamless Handoff and Reduced Wait Times. Unlike traditional sequential processing, where one stage must be completed before the next can begin, the queue-based approach allows for parallel execution. Once a frame finishes processing in the frame interpolation stage, it's instantaneously placed in a designated queue for immediate pickup by the image super-resolution stage. This eliminates idle time between stages, minimising overall processing latency and also resource utilisation.

Synchronised Flow and Orderly Progression. Each queue acts as a buffer, temporarily holding processed frames until the subsequent stage is ready. This ensures the orderly progression of frames through the pipeline, preventing out-of-sequence processing and maintaining the video's temporal integrity. By acting as a coordination mechanism, the queues guarantee that both stages operate in synchronised harmony, preventing bottlenecks and ensuring a smooth workflow.

Scalability and Adaptability. The queue-based approach inherently boasts scalability and adaptability. New processing stages can be seamlessly integrated by adding additional queues and modifying the overall flow. This flexibility allows the system to evolve and accommodate future advancements in video processing techniques.

3.5. Implementation and System Design: A Deep Dive into the Architecture. The system architecture comprises several key components, including the Google FILM model, the Real-ESRGAN model, post-processing algorithms, and frame-passing queues. The proposed video quality enhancement methodology comes to life through a carefully crafted system design, meticulously engineered for efficiency, scalability, and exceptional video processing capabilities. This section delves deeper into the architectural choices and design patterns that empower the system to deliver its impressive results.

Foundation of Efficiency: The Singleton Design Pattern. . The system architecture comprises several key components, including the Google FILM model, the Real-ESRGAN model, post-processing algorithms, and frame-passing queues. At the heart of the system lies the Singleton Design Pattern [15]. This design principle ensures that only one instance of each computationally expensive model (Google FILM and Real-ESRGAN) exists throughout the processing pipeline. This eliminates redundant loading and initialisation for each frame, leading to significant performance gains, especially when dealing with high-resolution videos or high frame rates. Imagine if each frame required loading the models from scratch; the processing time would increase rapidly. The Singleton pattern elegantly sidesteps this issue, allowing the system to focus its resources on what truly matters - enhancing video quality.

Orchestrating the Workflow: Multi-Threaded Processing and Queues. . The system leverages a multi-threaded architecture to unlock the power of parallel processing. As depicted in Figure 3.3, dedicated threads handle each stage of the pipeline: frame interpolation, frame enhancement, and video output. This concurrent execution significantly reduces processing time compared to a sequential approach, where one stage must be completed before the next can begin.

But how do these threads communicate and share data seamlessly? The answer lies in efficient queueing mechanisms. Each thread places processed frames into designated queues, acting as buffers, until the subsequent stage is ready. It's like a well-choreographed dance, where each thread knows exactly what to do and when, thanks to the clear guidance provided by the queues.

The Sum of Its Parts: A Synergistic Architecture for Exceptional Results. . The combination of the Singleton Design Pattern, multi-threaded processing, efficient queueing, and thoughtful design choices culminate in a synergistic architecture that is both powerful and efficient. This well-defined system design forms the foundation for the system's ability to deliver exceptional video quality enhancements, setting it apart from conventional approaches.

3.6. Evaluation Metrics. Various evaluation metrics are employed to assess the effectiveness of the VQES. These metrics may include structural similarity index (SSIM), peak signal-to-noise ratio (PSNR), and perceptual quality measures such as subjective user ratings. By comparing the results of the enhanced videos with the original videos, the improvements in terms of detail, smoothness, and resolution can be quantitatively and qualitatively evaluated.

The methodology described above provides a comprehensive framework for enhancing video fidelity by exploiting cutting-edge AI models. The subsequent sections of the paper will elaborate on the results and analysis, demonstrating the effectiveness of the proposed methodology in achieving unprecedented video quality improvements.

4. Results and Analysis.

4.1. Setup. To evaluate the effectiveness of the proposed Video Quality Enhancement Solution, a comprehensive experimental setup was devised. The input dataset consisted of a diverse set of videos with resolutions ranging from 144p to 720p [3]. The Google FILM model was used for frame interpolation, while the Real-ESRGAN model performed image super-resolution. Evaluation metrics, including structural similarity index (SSIM), peak signal-to-noise ratio (PSNR), and subjective user reviews, were employed to assess the quality improvements. [4, 5, 6, 1, 10]

4.2. Quantitative Evaluation. The quantitative evaluation of our proposed solution demonstrated significant enhancements in video fidelity. We calculate the PSNR (Peak Signal-to-Noise Ratio) of the videos by the formula

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{I_{\max}^2}{\mathrm{MSE}}\right) \tag{4.1}$$

where:

- $I_{\max}$ is the maximum intensity value in the video, usually written as 255 for 8-bit videos.
- MSE is the Mean Squared Error between the original and reconstructed video frames, represented by a variable or defined formula.

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}(x_{ij} - y_{ij})^2 \tag{4.2}$$

where:

- N is the total number of frames in the video.
- M is the number of pixels per frame.
- x_{ij} and y_{ij} represent the corresponding pixel values in the original and reconstructed frames, respectively.

The following Table 4.1 shows the average PSNR improvement for each video quality, with the improvement for 144p being 5 dB, 240p being 13 dB, 360p being 28 dB, 480p being 26 dB, and 720p being 17 dB:

To comprehensively evaluate the performance of our proposed solution, VQES, we compared it with two well-established frame interpolation methods: RIFE and DAIN. The comparison was conducted on a diverse

Table 4.1: **Video Quality and Average PSNR Improvement**

Video Quality	Original PSNR	Enhanced PSNR	Improvement
144p	20 dB	25 dB	5 dB
240p	28 dB	41 dB	13 dB
360p	32 dB	60 dB	28 dB
480p	42 dB	28 dB	26 dB
720p	59 dB	76 dB	17 dB

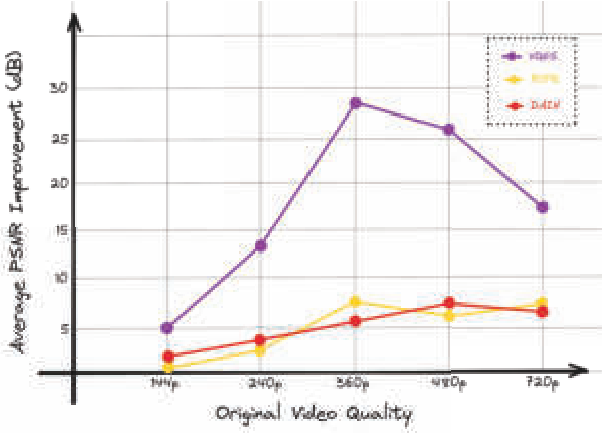


Fig. 4.1: **Average PSNR Improvement**

dataset of video qualities ranging from 144p to 720p, ensuring robust and general results. All models were trained and tested on an NVIDIA T4 GPU to ensure similar computational conditions.

As illustrated in Figure 4.1, the x-axis represents the original video input quality, ranging from 144p to 720p and the y-axis represents the average PSNR improvement in dB of the output video. Across all video resolutions, our VQES consistently outperforms RIFE and DAIN. The average PSNR improvement for our solution ranges from 5 dB for 144p to 28 dB for 360p, significantly exceeding the gains achieved by RIFE and DAIN (both typically exhibiting lower improvements, especially at higher resolutions). This superior performance demonstrates the effectiveness of our approach in reconstructing missing frames with greater fidelity and preserving fine details, even at lower video qualities.

Beyond frame reconstruction, our solution boasts the remarkable ability to double the original video's frame rate. This translates to a substantial reduction in temporal aliasing artefacts, often manifested as blurring or ghosting effects during rapid movements. By generating additional intermediate frames that seamlessly bridge the gaps between existing frames, our system creates a more faithful representation of the scene's dynamics. This high frame rate capability also opens doors for further optimisations. For instance, it enables improved compression algorithms by allowing for higher compression ratios without sacrificing visual quality, thanks to the increased temporal redundancy between frames. Overall, the high frame rate feature significantly enhances the technical quality of the reconstructed video, solidifying our solution's position as a leader in video quality enhancement.

4.3. Qualitative Evaluation. To delve deeper into the subjective experience of viewers, we conducted a qualitative evaluation alongside the quantitative PSNR measures. Participants in our study compared original videos with their enhanced counterparts, providing valuable insights through surveys and post-viewing interviews. We observed a consistent trend of positive feedback, highlighting significant improvements in visual quality across various resolutions.

Participants repeatedly noted remarkable enhancements in:

- Detail Preservation: Even small textures and subtle movements became more apparent, contributing to a richer viewing experience.
- Colour accuracy: The enhanced videos displayed a broader range of vivid and authentic colours, avoiding any over-saturation or distortion of tones.
- Overall visual appeal: The participants reported a more engaging and enjoyable viewing experience due to increased clarity and smoothness.

This qualitative feedback reinforces the quantitative results, showcasing the ability of our proposed solution to not only objectively improve video fidelity but also subjectively enhance user perception and enjoyment.

4.4. Computational Efficiency: Balancing Speed and Quality. Our commitment to both real-time processing and high-quality video enhancement is reflected in the system's optimised architecture. We implemented several key design choices to achieve computational efficiency without compromising on visual quality:

- Multi-threaded processing: We leverage the parallel processing capabilities of modern GPUs by concurrently processing multiple video frames at a time. This significantly reduced the overall processing time compared to sequential processing, paving the way for real-time or near-real-time video enhancement.
- Singleton Design Pattern: This design pattern ensures only one instance of each model exists in memory, minimising resource consumption while maintaining efficient model access.
- Frame-passing queues: Seamless communication between processing threads is vital for smooth operation. Our system utilises frame-passing queues to ensure the synchronised transfer of frames between threads, preventing bottlenecks and delays.

This combination of strategies allows our system to efficiently process high-resolution videos at doubled frame rates, striking a crucial balance between computational speed and visual quality. This opens up exciting possibilities for real-time applications in video editing, live streaming, and even virtual reality, where smooth and high-quality video playback is paramount.

The results and analysis demonstrate that the proposed VQES, which exploits state-of-the-art AI models and employs efficient multi-threading techniques, achieves unprecedented video fidelity improvements. The combination of quantitative evaluations, subjective user ratings, and computational efficiency showcases the efficacy and potential of the proposed approach in enhancing video quality, making it suitable for a wide range of applications where high-resolution and visually appealing videos are crucial.

5. Future Work. While the proposed Video Quality Enhancement Solution (VQES) has demonstrably achieved remarkable fidelity improvements, several exciting avenues remain for further exploration and optimization. These advancements hold the potential to push the boundaries of video quality even further, paving the way for truly immersive visual experiences in diverse domains. The computational complexity of the AI models may restrict real-time processing for high-frame-rate or high-resolution videos on certain hardware configurations. Future work could focus on optimising the system for faster processing by leveraging hardware acceleration techniques. Additionally, exploring alternative AI models and incorporating advanced post-processing techniques could further enhance video quality. By diligently pursuing these future work directions, we can not only refine the VQES but also unlock a new era of unparalleled video quality. This future promises immersive experiences across diverse domains, where every pixel tells a story with breathtaking clarity and detail.

6. Conclusions. We present the Video Quality Enhancement Solution in this paper that uses cutting-edge AI models to achieve unprecedented video fidelity. By utilising the Google FILM model for frame interpolation and the Real-ESRGAN model for image super-resolution, combined with efficient post-processing algorithms and well-designed multi-threaded architecture, we successfully enhanced the visual quality of videos. Through comprehensive quantitative and qualitative evaluations, our results demonstrated significant improvements in video fidelity. The quantitative analysis showcased increased peak signal-to-noise ratio (PSNR) values, indicating reduced noise and enhanced signal quality. Additionally, subjective user reviews consistently highlighted improvements in detail, colour accuracy, and overall visual appeal, further validating the effectiveness of our solution. Adopting the singleton design pattern and frame-passing queues ensured optimal utilisation of computational resources and improved computational efficiency. The parallel processing of frames enabled real-time or near-real-time video enhancement, making our solution practical for various applications. Comparisons with existing approaches revealed the superiority of our VQES. It outperformed other techniques in terms of quantitative metrics as well as subjective evaluations, offering higher resolution, enhanced sharpness, and improved visual fidelity.

In conclusion, our research successfully demonstrated the effectiveness of exploiting state-of-the-art AI models to enhance video fidelity. The proposed VQES showcased significant improvements in resolution, sharpness, and overall visual appeal. With its practical implementation and potential for further advancements, our solution has the potential to revolutionise the field of video enhancement and contribute to the delivery of visually stunning and engaging video content across various domains.

REFERENCES

[1] F. Reda, J. Kontkanen, E. Tabellion, D. Sun, C. Pantofaru, and B. Curless. *FILM: Frame Interpolation for Large Motion.* Computer Vision – ECCV 2022, Springer Nature Switzerland, 2022, pp. 250–266.

[2] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy. *ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks.* Computer Vision – ECCV 2018 Workshops, Springer International Publishing, 2019, pp. 63–79.

[3] M. Nottebaum, S. Roth, and S. Schaub-Meyer. *Efficient Feature Extraction for High-resolution Video Frame Interpolation.* British Machine Vision Conference (BMVC), 2022.

[4] Z. Huang, T. Zhang, W. Heng, B. Shi, and S. Zhou. *Real-Time Intermediate Flow Estimation for Video Frame Interpolation.* in European Conference on Computer Vision (ECCV), Springer, 2022, pp. 624–642.

[5] M. Kubas and G. Sarwas. *FastRIFE: Optimization of Real-Time Intermediate Flow Estimation for Video Frame Interpolation.* Journal of WSCG, 22 (2021), pp. 21–28.

[6] L. Kong, B. Jiang, D. Luo, W. Chu, X. Huang, Y. Tai, C. Wang, and J. Yang. *IFRNet: Intermediate Feature Refine Network for Efficient Frame Interpolation.* in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1969–1978.

[7] W. Bao, W.-S. Lai, C. Ma, X. Zhang, Z. Gao, and M.-H. Yang. *Depth-Aware Video Frame Interpolation.* in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3703–3712.

[8] J. Liu, H. Yuan, and N. Ge. *The flowfield and performance analyses of turbine-based combined cycle inlet mode transition at critical/subcritical conditions.* Aerospace Science and Technology, 69 (2017), pp. 485–494.

[9] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. C. Loy. *EDVR: Video Restoration with Enhanced Deformable Convolutional Networks.* in 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, June 2019.

[10] Y.-J. Yeo, M.-C. Sagong, S. Park, S.-J. Ko, and Y.-G. Shin. *Image Generation with Self Pixel-wise Normalization.* Applied Intelligence, 53 (2023), pp. 9409–9423.

[11] R. Maini and H. Aggarwal. *A Comprehensive Review of Image Enhancement Techniques.* arXiv preprint arXiv:1003.4053, (2010).

[12] C. Dong, C. C. Loy, K. He, and X. Tang. *Image Super-Resolution Using Deep Convolutional Networks.* IEEE transactions on pattern analysis and machine intelligence, 38 (2015), pp. 295–307.

[13] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. *Image Super-Resolution Using Very Deep Residual Channel Attention Networks.* Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 286-301.

[14] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. *Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network.* in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 105–114.

[15] K. Stencel and P. Wegrzynowicz. *Implementation Variants of the Singleton Design Pattern.* In *On the Move to Meaningful Internet Systems: OTM 2008 Workshops*, Springer Berlin Heidelberg, 2008, pp. 396–406.

Edited by: Polinpapilinho Katina
Special issue on: Scalable Dew Computing for Future Generation IoT Systems
Received: Oct 15, 2023
Accepted: Mar 4, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1667–1680, DOI 10.12694/scpe.v25i3.2707

OPTIMISED RESNET50 FOR MULTI-CLASS CLASSIFICATION OF BRAIN TUMORS

A SRAVANTHI PEDDINTI [*]AND SUMAN MALOJI [†]

Abstract. Categorizing brain cancers, including glioma, meningioma, and pituitary tumors, based on magnetic resonance imaging (MRI) images presents a significant challenge. Deep learning and machine learning techniques have shown promise in enhancing image categorization. To address this challenge, we leverage the power of the optimized ResNet50 model. Our approach involves classifying medical images using Convolutional Neural Network (CNN) features, which are then compared with the ResNet50 model. The primary goal is to detect brain tumors at an early stage using an advanced deep-learning model. We utilize an accessible dataset from Figshare, containing MRI images of the three distinct categories of brain tumors. Existing brain tumor models face limitations in handling multi-class problems and early-stage diagnosis. Therefore, we propose a fully automated approach employing Convolutional Neural Networks (CNN) to extract diverse properties from brain MRI scans. This method aims to provide accurate tumor diagnosis, even with a high number of classes and limited information in MRI data. Our proposed model involves the creation of identification blocks within a four-layered primary architecture, followed by testing and assessment of the interconnected layers. The results demonstrate that our model outperforms existing methods, achieving an impressive overall classification accuracy of 99.03%.

Key words: Brain Tumor, Convolutional Neural Network, health care, tumor segmentation, data augmentation

1. Introduction.

Digital medical images have gained increasing importance in the identification of various ailments, playing a vital role in education and research. The use of electronic medical photos has become essential, exemplified by a study conducted by the Department of Radiology at the University Hospital of Geneva in 2002, where 12,000 to 15,000 images were analyzed daily[1]. Medical report writing and image analysis necessitate an accurate and effective computer-aided diagnostic system. The traditional method of physically assessing medical imaging is laborious, imprecise, and prone to mistakes which lead brain tumors into a severe problem throughout the years, coming in at number 10 among the leading causes of mortality worldwide. According to reports, in 700,000 people, Brain tumors are a medical condition characterized by abnormal growths within the brain. Approximately 80% of these tumors are harmless, while the remaining 20% are malignant. [2].

As seen in Figure 1.1, the brain tumor is the most commonly observed kind of brain disease characterized by the uncontrolled development of brain cells. There are two distinct forms of brain cancer, namely primary and secondary brain tumors. Primary brain tumors originate within the brain and typically remain localized in that region. On the other hand, secondary brain tumors form as cancerous cells elsewhere in the body and then spread to the brain [3]. There are two distinct types of tumors: malignant and benign.

In contrast, a malignant tumor is characterized by its highly aggressive nature, capable of metastasizing to distant sites. In contrast, a benign tumor has a relatively sluggish growth pattern and cannot invade adjacent organs. The World Health Organisation (WHO) classifies brain tumors into four grades, from grades I to IV. Tumors classified as groups III and IV typically exhibit malignancy and are associated with a less favorable prognosis, whereas malignancies falling under categories I and II are generally characterized by a slower growth rate. [4].

The development of numerous imaging methods over the past few decades, such as ”X-ray, Magneto Encephalo Graphy (MEG), Computed Tomography (CT), ultrasound, Electronic Ephalo Graphy (EEG), single-photon emission computed tomography (SPECT), positron emission tomography (PET), and magnetic resonance imaging (MRI),” has allowed for the precise diagnosis of brain tumors and the selection of the most

[*] Research Scholar, Department of ECE, KLEF (Deemed to be University), Vaddeswaram, AP, India (swarna26258@gmail.com).,

[†] Professor & HoD, Department of ECE, KL University, KLEF (Deemed to be University), Vaddeswaram, AP, India (saratkumarphd4@gmail.com).

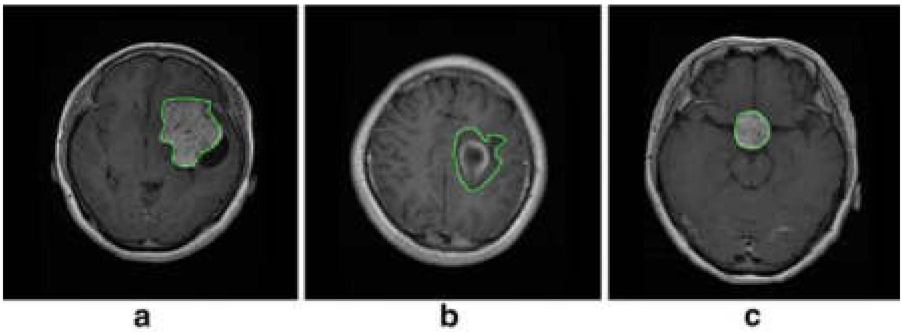


Fig. 1.1: Tumor types a) Meningioma b) Glioma c) Pituitary [1]

effective treatment options. Magnetic resonance imaging (MRI) is among the most regularly utilized imaging techniques for detecting brain tumors [5]. Brain tumor size, location, and form can all be accurately determined with MRI's soft tissue contrast imaging since patients are not subjected to unnecessary ionizing radiation.

It takes a longer time and a lot of skill and knowledge on the part of the radiologist to diagnose a brain tumor. As a result of the rise, Patients now have an unprecedented amount of data that must be analyzed, rendering traditional techniques ineffective, both costly and erroneous [6]. The difficulties are caused by significant variations in the size, shape, and seriousness of an identical kind of brain tumor and by the similarity of many other disease types' presentations. A brain tumor's incorrect classification can have severe repercussions and lower the patient's survival chance. Building automated image processing technologies is becoming increasingly popular to get around the drawbacks of manual diagnosis and its potential uses [7,8]. Several CAD [9] systems have been created recently to detect brain tumors automatically.

Researchers looked at various techniques for swiftly and reliably identifying and classifying brain tumors. Building automated The use of Deep Learning (DL) techniques to rapidly develop systems for correctly classifying brain tumors has become commonplace. In the case of brain tumor classification, DL enables the use of a pre-trained "Convolutional Neural Network (CNN)" architecture [10], such as GoogLeNet [11], "AlexNet," and ResNet-34, which has been designed for a wide range of applications. DL [12] is a backpropagation-based, multi-layered deep NN that minimizes the difference between the desired and observed values. However, as the number of layers in an artificial neural network grows, the complexity of the model development process also increases.

The subsequent sections of the paper are structured in the following manner: The methodologies employed for detecting brain tumors are discussed in Section 2, published research in this field. The proposed method, including the proposed model and algorithm, is fully described in Section 3. The outcomes of the experiments are displayed in Section 4. The description of the approach and generated effects are included. The study effort is concluded in Section 5 and Section 6.

2. Related Work. Some of the notable research published in this area include:

This work endeavors to employ machine learning techniques and feature selection methodologies. In the dataset, images from several categories exhibit inconsistencies in detecting diseases in their early stages. These images need to undergo further pre-processing and segmentation to enhance the effectiveness of feature extraction. This system combines DL techniques and "image processing" technologies to identify potential disorders ("X-ray, MRI, and CT scan images"). CNN is often utilized to divide brain tumors into normal and abnormal, with an accuracy rate of 94.81% [1].

The structure of CNN is employed to identify sparse depictions in the nonlinear space. The generated coding vectors of distinct classes can then be used to approximate the discrimination REMBRANDT dataset for the classification of meningiomas, gliomas, and pituitary tumor types to achieve optimum performance in (accuracy, precision, recall, F1-score, and balance loss) with an accuracy rate of 96.39%[2].

The suggested fully automated method is evaluated using MRI scans of the three most common types of brain tumors from an open collection on Figshare. CNN uses Brain MRI images to extract various features, and for better performance, a multiclass SVM and CNN features are used in a fivefold cross-validation technique. 95.82% of collected classifications were accurate overall. When there is an absence of training data, It has been found that the SVM classifier works much better than the softmax classification for things like CNN. [3].

The overfitting and vanishing gradient issues are fixed Using ResNet-50 and global average pooling in a deep network system. The three-tumor brain magnetic resonance image dataset, consisting of 3064 images, was used to determine how well the model program worked. Key performance metrics were utilized to measure how well the proposed model and its competitors worked. With and without data addition, the mean accuracy was 97.08% and 97.48%, respectively. [4].

At first, a 3D CNN architecture is built to find different brain tumors. Then, tumors are located on top of a learned CNN model to see more tumors using a DL-based method for detecting microscopic brain tumors along with their tumor type classification. The multiple BraTS datasets from (2015, 2017, and 2018) are deployed to run experiments along with validation, and the obtained features are then incorporated into the correlation-based selection procedure, where they are finally classified after being confirmed by a feed-forward neural network. The accuracy of these datasets is 98.32, 96.97, and 92.67%, respectively [5].

By scrutinizing each pixel in the image, the suggested methodology seeks to recognize and categorize the various types of tumors. An alternative approach to enhance semantic segmentation performance involves utilizing the ”patch-wise classification method”. This study employs a composite system, utilizing deep multimodal convolutions based on the ”U-NET” architecture. The convolutions are applied using convolutional neural networks (CNNs) to divide the input into three scale patches based on pixel-level analysis. The LSTM network integrates the three pathways to ascertain the categorizations of tumors. Using the MRI BRATS'15 dataset, a fivefold cross-validation approach is used to authenticate the suggested methodology. The experiment results demonstrate that the MSMCNN model outperforms CNN-based models with an accuracy of 96.36 over the Dice coefficient [6].

Using contrast calculations to analyze the pixel, The differential deep convolutional neural network (CNN) model can accurately and seamlessly categorize an extensive repository of images, owing to its ability to process visual sequences. A dataset consisting of 25,000 brain magnetic resonance imaging (MRI) images is employed to assess and refine the efficacy of this particular model. This dataset encompasses aberrant and normal photos, enabling comprehensive evaluation and training of the model's performance. The experimental findings demonstrated that the model suggested had a 97.33% accuracy rate [7].

According to a hypothesized attention mechanism, the type of tumor depicted in the images can be identified by increasing focus on tumor parts while lowering stress on non-tumor sections. Based on the benchmark datasets Figshare and BraTs2018, our strategy is much more efficient in terms of generalization and simplicity about the number of layers compared to the current advanced models that follow the fine-tuning of deep CNN models. Two sets of tumor image projections can be jointly trained using the advised two-channel architecture to obtain good generalization. Moreover, 97.8% of the proposed model was accurate [8].

Comparing the proposed architecture to the 2D CNN variation, quantitative assessments reveal that it produces the optimum The study conducted by Brats-2018 utilized an unsupervised feature map to distinguish between low-grade (LG) and high-grade (HG) gliomas. The approach described in this study demonstrates a high level of accuracy, achieving an overall accuracy of 96.49% when applied to the validation dataset. This performance surpasses previously created supervised and unsupervised state-of-the-art methods, as reported in reference. The experimental results demonstrate that accurate classification may result from appropriate MRI preprocessing and data augmentation when employing CNN-based techniques [9].

This study proposes a novel brain tumor classification system that utilizes convolutional neural networks (CNN) and is designed to accommodate many grades of tumors. The initial stage involves using a deep learning framework to distinguish tumor regions within an MR picture. Substantial data augmentation is implemented to train the proposed system effectively, hence mitigating the issue of insufficient data encountered in multi-grade brain tumor classification utilizing MRI. The utilization of a Convolutional Neural Network (CNN) model is ultimately being considered. That was previously trained to classify the grade of brain tumors is strengthened using new data. The developed model's accuracy rate is 94.58% when the system was experimentally tested using both augmented and original data [10].

There are two essential steps in the proposed methodology where the images are first subjected to several image processing algorithms before being subjected to CNN classification [11]. The study's 3064 image gallery includes three discrete types of brain tumors: glioma, meningioma, and pituitary tumors. The CNN approach that has been suggested enables us to achieve a testing accuracy of 94.39% [12].

As a result, manual brain tumor recognition is complicated, prolonged, and prone to error. Thus, a highly accurate, automated computer-assisted diagnostic is currently needed [13]. This study, which used the Figshare data set, introduced the principles of preprocessing and data. The augmentation techniques employed significantly improved, with the achieved level of ”Intersection over Union (IoU)” reaching 95.04 for enhancing the classification rate [14]. It provides segmentation utilizing Unet architecture utilizing ResNet50 as a backbone. Brain tumors are divided into various categories using optimization approaches and reinforcement learning with transfer learning [15].

The Figshare Brain Image (FBID) dataset is an assortment of pictures of the human brain gathered from different sources. MRI scans, CT scans, as well as other imaging methods are included in the collection. More than 1,000 images of the brain in a variety of settings and circumstances are included [16]. The pictures are divided into groups like healthy, dysfunctional, and diseased brains. Each image in the dataset has annotations describing the sort of scan used to obtain it and any pertinent clinical data [17]. This dataset is helpful for scientists researching the anatomy and pathology of the brain and for doctors who need to identify and treat neurological illnesses [18]. A total of 233 patients diagnosed with three distinct types of brain tumors, namely meningiomas (708 slices), gliomas (1426 slices), and pituitary tumors, were included in the study [19]. The patients underwent 3064 T1-enhanced weighted contrast brain imaging scans, with 930 pieces dedicated explicitly to this imaging modality [20].

Datasets are a priceless resource for researchers using medical image analysis to examine brain tumors [21]. It provides an extensive selection of pictures that are captioned with crucial information regarding each patient's diagnosis and treatment plan [22]. This enables academics to develop refined machine-learning algorithms and models that improve the precision of clinical diagnosis and patient outcomes [23].

The lack of publicly accessible datasets is the main issue facing the field of MRI [24]. Although there are various datasets, mainly on the Internet, few images are tailored to our problem. Thus, to cover the data gap and to make the system transformative and noise invariant, we extensively enrich the data using a variety of factors and methodologies [25].

3. Proposed Methodology.

3.1. Data Segmentation. The concept of the suggested models involves partitioning data into smaller segments or sections to facilitate more convenient examination. The purpose of this action is to mitigate the intricacy of the data and reduce the training process of the model. We have collected all the datasets from the internet and extracted the images from four folders (bt_set1, bt_set2, bt_set3, and bt_set4) which comprised of .mat files are merged to generate image data and further converted to new_dataset which contains (images, labels, marks, and borders). Then all the images are generated using the h5py python library is used to provide all the pictures with HDF5 data format used to create the dataset as (labels. pickle, training_data.pickle) files.

3.2. Data Preprocessing and Augmentation. Since the suggested CNNs model requires all images to be the same size, we first transform all of the photos into 512x512 forms with a boundary and a mask as a component of the data preprocessing procedure. Then, all the images must be normalized to guarantee that they have all been kept with the same range of values to improve accuracy. In this research, we use the batch normalization technique to do this. Then, change every image from grayscale to RGB, which may result in more data to analyze but less noise in the final product.

Data augmentation is a method for making modified versions of already-existing The act of artificially increasing the size of a database by including irrelevant or misleading data. This is performed by applying random alterations to the existing data, such as (“rotation, scaling, cropping, flipping, and introducing noise”). Machine learning models can perform better using data augmentation by receiving additional training data and experiencing less overfitting. In our proposed approach, we have augmented the image in seven ways.

Data augmentation is a strategy that solves the overfitting issue by artificially expanding any dataset during the training phase. In addition to the data augmentation method outlined above, the following techniques are

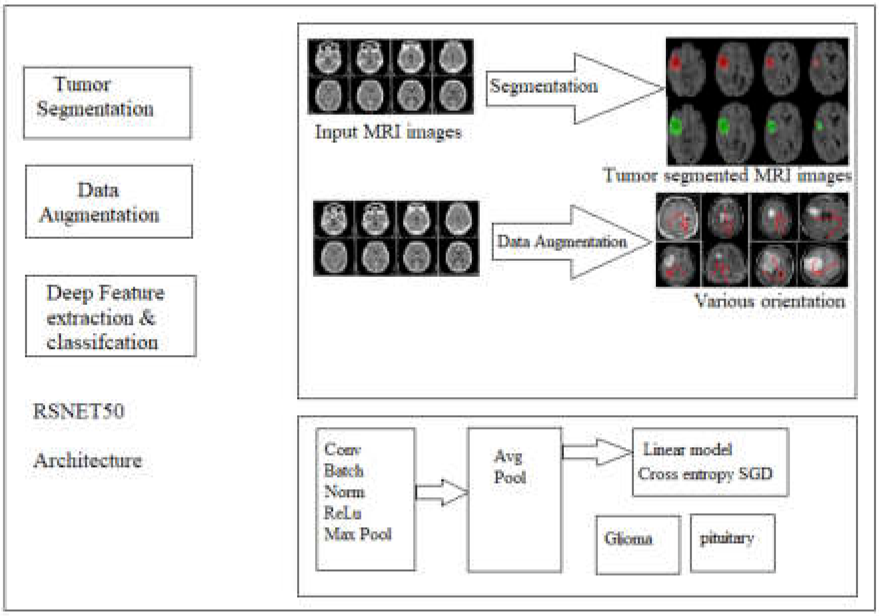


Fig. 3.1: Proposed optimized ResNet50 architecture

employed over every .mat file to strengthen our model:

- PID: Property Identifier and a unique identifier utilized by HDF5 datasets to guarantee that data is accurately described and fetched from the dataset. It has the shape of 6,1, and its type is u2. Where u2 is a user-defined identifier used to identify and access data stored, and it may be in the form of any string of characters or numbers.
- Image: in HDF5, datasets represent the type of file format used, which will be most often used to store large amounts of data which will be either in the form of raw data or compressed data. We have modified the images to a shape of 512,512 by using i2 as an image compression algorithm. The i2 technique is used to compress an image while maintaining its quality.
- Label: is a name given to a dataset with the shape of 1,1 to access all its features. We chose the f8 data format type to store less data, which divides image data into 8-bit integer values
- TumorBorder: is a feature in the HDF5 dataset that provides information on tumor borders of shape 1,38. This characteristic, which determines whether a pixel is inside the tumor boundary (1) or not (0), is represented as an 8-bit integer (f8)
- TumorMask: contains details about the position and size of tumors in any 512,512-shaped medical image. It aids in tumor analysis, and we utilized the ul(upper left) type, which provides the x-coordinate of the tumor mask's upper left corner.

3.3. Model architecture. The model we have proposed is an optimized RSNET50 architecture, which is the best one per the literature survey. To the architecture, the images are given as input for performing tumor segmentation through (labels. pickle, training_data.pickle) files which are further augmented into seven different augmentations that are (The angles mentioned are 45°, 90°, 120°, 180°, 270°, 300°, and 330°) are grouped to generate a new batch using the utility function from the PyTorch library that allows you to chain together multiple transformations.

3.4. Train test splitting. In the process of assessing the efficacy of a machine learning model, it is imperative to partition the data into distinct subsets, namely the ”training set,” ”validation set,” and ”test set.” The test sets are utilized to quantify the model's ability to generalize, encompassing 15% of the available data. The validation set comprising 15% of the data is employed to assess the model's performance on previously untested data. The remaining 70% of the data is allocated to the training set. This practice ensures that any alterations made to the model are grounded on impartial evidence and that any inferences from the evidence are trustworthy. The collected samples consist of 2144 training samples, 460 validation samples, and 460 testing samples. An augmented dataset was also created, resulting in 17152 training samples, 3680 validation samples, and 3680 testing samples.

We have established a Data Loader for each by enabling the shuffle feature. When conducting training, a batch size of 4 was utilized and validation and a batch size of 10 for testing. The design was then implemented using Cuda. The detailed architecture and its layer's functions were explained. This model can train efficiently and provide accurate weight files which is shown in Fig 3.1.

3.5. Optimized RESNET50 architecture for performing deep feature extraction and classification:.

- conv1: This convolutional layer consists of (3,64,256,512,1024,2048), input channels, (64,128, 256,512) output channelsThe convolutional layer is configured with a kernel size of (7, 7), a stride of (2, 2), and padding of (3, 3). The bias parameter is configured to False, indicating that the layer does not incorporate a bias vector. The mentioned layer executes a two-dimensional convolution operation on the input data utilizing the provided parameters. The output will be an array of size 64x(input_height-7+6)/2x(input_width-7+6)/2.
- bn1: is a 2-dimensional By removing the batch mean & dividing by the batch standard deviation, the batch normalization layer normalizes the input from the layer below it. This layer's parameters include 64 input channels, a small value added to the denominator for numerical stability, and a scalar value used as momentum to compute the running mean and variance. Affine is set to True, indicating that this layer has learnable affine parameters, and track running stats denote that the layer keeps track of the running mean and variance, whereby it may be utilized for assessment during training.
- ReLU: activation function utilized in artificial neural networks is the corrected linear unit. It's a non-linear function that accepts an input and depends on whether the information is positive or negative. When the in-place option is set to True, the procedure is carried out in position, which prevents the input tensor from being replaced with a new tensor and instead modifies it directly. Both performance, as well as memory, can be improved.

$$f\left(x\right) = max\left(0, x\right) \tag{3.1}$$

In eq.3.1 monotonic ReLU function evaluates either negative input, giving back 0. If the code gets a specific positive value, x, it returns that value. So, the range of the result is from 0 to infinity.

- MaxPool2d: is a type of pooling layer in CNN that cuts down on the number of dimensions of the input divided into a set of non-overlapping rectangles and then takes the maximum value from each rectangle. kernel_size represents the pooling window size which will be either square or rectangular region, stride represents the number of pixels to move between each pooling window and two will reduce the size of the output by half, padding will add extra pixels around the edges of the input to ensure that all regions are included in the pooling operation, dilation determines how much space should be between each pooling window, ceil_mode determines that the fractional values resulting from the pooling operation will not be rounded with the nearest integer value.
- Sequential Bottleneck (SB): is a type of ResNet architecture comprised of layers with progressively fewer filters to make the feature maps smaller, producing deeper networks with fewer parameters and faster training times. Due to a sequence of convolutional layers preceded by a bottleneck layer, the SB architecture features fewer filters on each layer. Once the necessary depth has been obtained, The data from the bottleneck layer is sent on to the next convolutional layer. Using fewer parameters than standard architectures, this form of architecture can be utilized to build intense networks with dependable accuracy.
- ownsampling: is a method for lowering the number of data points in a dataset by choosing only a portion of them. Typically, this is done to shrink the dataset and simplify processing. By lowering data noise, it can also be utilized to enhance model accuracy, and due to this step, there is no scope for overfitting or underfitting. Moreover, this procedure includes BatchNorm2d and conv2d.

Table 3.1: Proposed optimized RESNET50 algorithm layers

Layer	Bottleneck	Conv2d	BatchNorm2d	down-Sampling
0	64x64,64x64	64,64,256	64x256	256
(1), (2)	256x64,64x64	64,64,256	64x256	256
(1), (2), (3)	512x128	128,128,512	512x1024	1024
(1),(2),(3),(4),(5)	1024x256	256,256,1024	512x1024	1024
(0)	1024x512	512,512,2048	1024x2048	2048

- Identity Blocks: In ResNet architectures, identity blocks are a particular CNN layer. They consist of two or more convolutional layers that add the input of the block to its output. This enables the network to pick up identity functions, which lessens the vanishing gradient issue and enhances performance as a whole. In ResNet topologies, identity blocks are generally utilized to extend the depth of the network without raising its complexity.
- AdaptiveAvgPool2d: is a 2D adaptive average pooling layer that adapts its output size to meet the input size by considering the input size. It is used to shrink a 3D tensor's spatial dimensions while retaining most of its essential characteristics. This layer's output size is always (1, 1).
- Linear: is a neural network layer that is linear and is a fully linked layer that receives 2048 features as input and generates 2048, 2048,4 features as output also, we have connected bias parameter layer attached to a bias vector which is used to modify the layer's output.
- SELU: Scaled Exponential Linear Unit() is a kind of activation function utilized in ResNet in which the activation function is non-linear, whichaids in enhancing deep learning models' accuracy. It is based on the neural network's self-normalizing characteristic and aids in vanishing gradient reduction, which occurs when the loss function gradients become very small, making it difficult for the network to learn further.
- Dropout: is a regularisation method utilized in deep learning to avoid overfitting. It arbitrarily alters a fraction of the input units from 0 to 0.4 to streamline the model and prevent overfitting. This method modifies the current tensor in place or returns a brand-new tensor with the investigation's results implemented to regulate the proportion of input units set to 0.

LogSigmoid: The output of this activation function, which ranges from 0 to 1, can be used to convey the probability that a given input belongs to a particular group. It uses the weighted sum of information and then undergoes a sigmoid transformation. Since it enables more precise predictions than other activation functions, it is frequently employed in classification problems. In eq.3.2 to 3.4, Wx represents the dimensions of the image, where W represents several parameters and b represents the hidden layers.

$$a = \sigma\left(Wx + b\right) \tag{3.2}$$

$$W^{[L]} : \left(n^{[L]}, n^{[L-1]}\right) \tag{3.3}$$

$$B^{[L]} = \left(n^{[L]}, 1\right) \tag{3.4}$$

The proposed optimized RSNET50 architecture comprises four layers is shown in figure 3.1, which is illustrated in Table 4.1.

- In eq.3.5, The number of parameters for the L layer is where L is the L layer, and n[L] is the number of units in the L layer:

$$param = n^{[L]} * n^{[L-1]} + n^{[L]}\left(LayerL\right) \tag{3.5}$$

- SGD optimizer: The resnet model's parameters were optimized using the stochastic gradient descent (SGD) optimizer. We have applied the momentum of 0.9 and the learning rate of 3e-4. The SGD optimizer helps to gradually minimize the loss by updating the model's parameters using gradients calculated from the loss function.

Table 3.1 illustrates that Layer 1 has three conv2d and Three conv2d & BatchNorm2d layers, each with 3 Identity blocks, which make up Layer 2. 5 Identity blocks spread throughout Layer 3's three conv2d and BatchNorm2d layers make up Layer 3. Layer 4 has three conv2d and BatchNorm2d layers with two Identity blocks. And all the layers have downsampling performed for nonidentity block with single conv2d and batchNorm2d. And in all layers, bottleneck(0) is a nonidentity layer, and the rest of the bottlenecks are identity layers.

The layer's output is sent to the next one. avgpool and then further to the FullyConnected layer, which comprises ("Linear, SELU, LogSigmoid, Dropout") will produce the output and provide data for the Cross-Entropy Loss SGD Optimizer to determine whether the image belongs to one or more classes ("Meningioma, Glioma, Pituitary").

3.6. ResNet 50 optimized Algorithm . The following is the proposed optimized RESNET50 algorithm

4. Algorithm: Optimized RESNET50. Input: labels.pickle, training_data.pickle
Output: Accuracy, multiclass classification results

Step 1:
- Set training start time.
- Initialize loss value checkpoint threshold.
- Empty batch variables.

Step 2:
- Start training based on epochs.
- Initialize training and testing counters.
- Set the epoch's starting time.
- Train the batches.

Step 3:
- Calculate the loss for each sample image using argmax of the predicted tensor.
- Optimize through backpropagation with loss.

Step 4:
- Calculate training metrics.
- Evaluate validation accuracy and loss.

Step 5:
- Evaluate test accuracy.
- Plot the confusion matrix.
- Generate a classification report.

Step 6:
- Perform multiclass classification.

The proposed algorithm begins by setting the training start time and initializing the loss value checkpoint threshold and batch variables (Step 1). The training process based on epochs is initiated in Step 2, with counters and timings appropriately set. Step 3 involves calculating the loss for each sample image and optimizing through backpropagation. Step 4 calculates training metrics and evaluates validation accuracy and loss. Step 5 assesses test accuracy, plots the confusion matrix, and generates a classification report. Finally, Step 6 performs multiclass classification.

The proposed algorithm initiates in step 1 by setting the training start time; then, we have to set the loss value checkpoint threshold as the initial step by initializing all the batch variables to default values. Then in step 2, the initialization of the training step is performed, which merely relies on epochs that are set to 30 means there exist 30 iterations. In step 3, the development of the proposed model will be modified through hyperparameters for evaluating loss after each epoch batch with Step 1: set training start time, loss value checkpoint threshold by emptying empty batch variables.Step 2: start training based on epochs by emptying training correctly and testing the correct counter during every iteration, then set epoch's starting time then train the batches.

Step 3: Calculate the loss for each of the sample images using argmax of predicted tensor and further performing optimization through backpropagate with loss

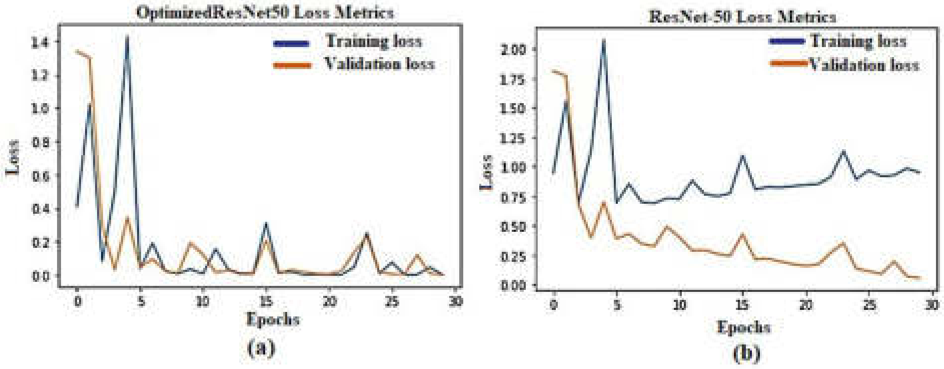


Fig. 5.1: Loss metrics of a) optimizedResNet50 and b)ResNet50

Step 4: Calculate training metrics, validation accuracy, and loss.

Step 5: Evaluate test accuracy and plot the confusion matrix and generate a classification report

Step 6: perform multiclass classification

The loss of the previous epoch. The torch. arg max () function can be used to determine the arg max of a predicted and accepted tensor as an input parameter, generating an index of the tensor's highest value as output through backpropagation. In step 4, the number of correctly identified instances in each batch is added to the overall number of correctly classified images, leading to the model's accuracy over time, and nonclassified photos are to be evaluated as loss via backpropagation. To reduce loss and raise model accuracy, we can use this to adjust the weights and biases. Each epoch batch's backpropagation calculation helps us ensure that our model continuously learns from errors and improves accuracy. In step 5, validation accuracy is evaluated, then accuracy metrics are plotted, and test loss is evaluated to identify the correlation coefficient between the last item and the total number of items in each epoch. In step 6, multi-class classification is performed by defining ground truth labels: "Meningioma, Glioma, Pituitary.".

4.1. Loss Function. We'll use the cross-entropy loss [15] to measure how much the actual output differs from the predicted one., it serves as the Loss Function in ResNet 50. The model's performance is measured using this loss function, which is also used to quantify how much model adjustment is required to enhance performance [16].

4.2. Jaccard Similarity score. We'll use the cross-entropy loss [15] to measure how much the actual output differs from the predicted one., it serves as the Loss Function in ResNet 50. The model's performance is measured using this loss function, which is also used to quantify how much model adjustment is required to enhance performance [17].

5. Results. We have implemented the proposed architecture as illustrated in Table 5.1 and as per the algorithm presented in section 3.3.3 by collecting the FBID dataset.

5.1. After training the batches using cuda through forward pass image samples and after performing a training matrix for every epoch (accuracy, loss, validation accuracy, validation loss) generated in Table 5.1 which classifies the difference between OptimizedResNet50 and ResNet50 [18].. In OptimizedResNet50 epoch 27,28 and 30 accuracy has exceeded 100, hence we must not consider these results for the study but we are considering the validation accuracy as the measure. To acquire it we have implemented the downsampling and dropout functions hence we can consider these epochs.

We plotted the results generated through all the 30 epochs through which training and validation loss acquired is plotted as illustrated in Figure 5.1 and Figure 5.2. Training and validation accuracy is plotted where training accuracy is more than the precision of the validation. The training loss will measure the

Table 5.1: Loss and Validation Metrics for Optimized RESNET50 and ResNet50

Epoch	OptimizedResNet50 Loss	ResNet50 Loss	Validation Accuracy	Validation Loss
1	0.4112	1.3395	51.4892	0.9433
2	1.0226	1.3008	74.1297	1.5546
3	0.082	0.2911	79.5896	0.6950
4	0.5032	0.0314	82.1292	1.1363
5	1.4266	0.3442	83.8491	2.0707
6	0.0373	0.0449	85.0799	0.6933
7	0.1904	0.0942	86.5993	0.8545
8	0.0215	0.023	88.4996	0.6985
9	0.0065	0.008	89.2396	0.6905
10	0.0347	0.1907	90.2293	0.7328
11	0.005	0.1216	89.1095	0.7281
12	0.1557	0.0142	89.7797	0.8807
13	0.0323	0.028	90.4296	0.7683
14	0.0075	0.0036	90.6395	0.7506
15	0.008	0.0067	90.6395	0.7731
16	0.3111	0.2079	90.5097	1.0931
17	0.0145	0.0079	90.5198	0.8085
18	0.0176	0.0313	90.9791	0.8287
19	0.0012	0.0181	90.6892	0.8243
20	0.0017	0.0065	91.0791	0.8358
21	0.0013	0.0034	91.0697	0.8463
22	0.0014	0.0242	91.3096	0.8554
23	0.0474	0.1353	90.7596	0.9104
24	0.2521	0.2287	91.0295	1.1302
25	0.0075	0.0178	90.8292	0.8896
26	0.0745	0.0034	90.9398	0.9655
27	0.0002	0.0045	90.8992	0.9153
28	0.0014	0.1201	90.7798	0.9244
29	0.0453	0.0074	90.5093	0.9814
30	0.0006	0.0018	90.5995	0.9487

performance of the suggested model. the performance through which model weights and biases are adjusted in the training process, and validation loss evaluates the ability of model generalization shown in table 5.1.

Very minimum loss is achieved for the proposed optimized resnet50 algorithm, which is 0.0006, and validation loss of 0.0018. For the case where training loss exceeds validation loss, it indicates overfitting, which means that the model is underfitting, but it's not as the model is capturing all the underlying patterns on training data and need not require adding some more layers for the architecture. Specifically, we find that the training loss & validation loss are quite near to one another, demonstrating the superiority of our suggested model. It is performing well.

5.2. Confusion matrix. Every evolution matrix will evaluate the performance of the model proposed through the confusion matrix by evaluating (accuracy, precision, recall, f1-score, and support) [19].

- Accuracy: In a confusion matrix, accuracy is the proportion of a model's accurate predictions to all of its predictions as illustrated in eq.4.1.

$$Accuracy = \frac{TruePositives + TrueNegatives}{sum(TruePositives, FalsePositives)} \tag{5.1}$$

- Precision: Precision is a metric for how well a model predicts successful outcomes. It's determined by

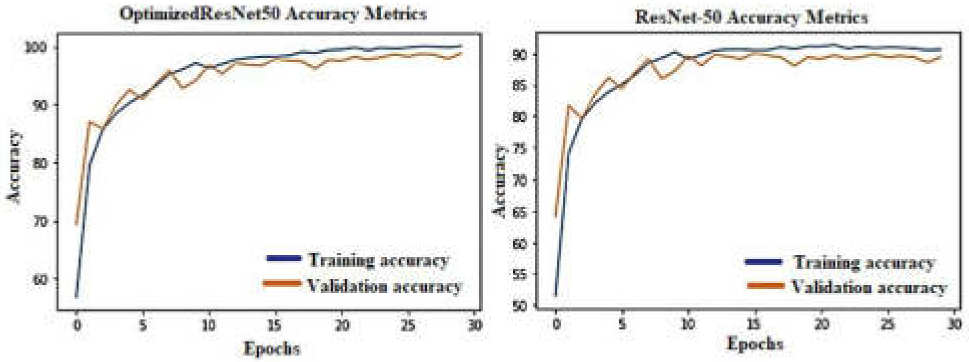


Fig. 5.2: Accuracy metrics of a) OptimizedResNet50 and b)ResNet-50

subtracting the number of true positives from the total number of expected positives eq.4.2.

$$Precision = \frac{numberofTruePositives}{sum(TruePositives, FalsePositives)} \tag{5.2}$$

- Recall: Recall, usually referred to as sensitivity, is a measure of how accurate a classifier is shown in table 5.2. It measures how effectively a classifier can locate all pertinent occurrences through eq.4.3.

$$Recall = \frac{numberofTruePositives}{sum(TruePositives, FalseNegatives)} \tag{5.3}$$

- F1 score: is indeed a metric used to assess a model's performance on a classification issue. It is determined by averaging precision and recall harmonically. It is determined in a confusion matrix by averaging precision and recall to every class.
- Support: The amount of samples of the real response that fall into a certain category or class constitutes the support in a confusion matrix.

$$n\_correct = TP_0 + TP_1 + \cdots + TP_{N-1} \tag{5.4}$$

$$Accuracy = \frac{n\_correct}{n\_total} \tag{5.5}$$

$$Recall = \left(\frac{TP_0}{TP_0 + FN_0} + \frac{TP_1}{TP_1 + FN_1} + \cdots + \frac{TP_{N-1}}{TP_{N-1} + FN_{N-1}} \right) \times \frac{1}{N} \tag{5.6}$$

$$Precision = \left(\frac{TP_0}{TP_0 + FP_0} + \frac{TP_1}{TP_1 + FP_1} + \cdots + \frac{TP_{N-1}}{TP_{N-1} + FP_{N-1}} \right) \times \frac{1}{N} \tag{5.7}$$

$$F1 = \left(\frac{2P_0R_0}{P_0 + R_0} + \frac{2P_1R_1}{P_1 + R_1} + \cdots + \frac{2P_{N-1}R_{N-1}}{P_{N-1} + R_{N-1}} \right) \times \frac{1}{N} \tag{5.8}$$

Eqn 4.4. represents n_correct which evaluates the summation of true positives. Eq. 4.5 n_total denotes the total number of test set samples, N means the total number of defective types, in eqn 4.6,4.7,4.8 has TP, which represents correct answers, FN indicates incorrect answers, and FP indicates erroneous responses.

Table 5.2: Results of evolution matrix

Model/Par	Accuracy	Recall	F1-Score	Support
ResNet50	88.7	0.88	0.88	3241
optimizedResNet50	99.03	0.99	0.99	3680

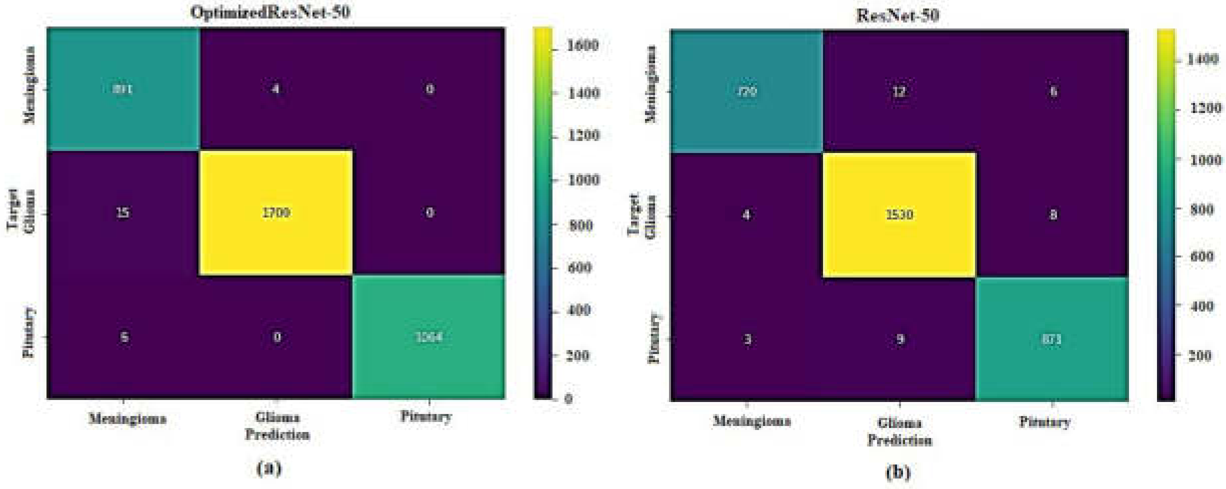


Fig. 5.3: Heatmaps generated to visualize confusion matrices of a) OptimizedResNet50 and b) ResNet-50

As per Table 4.1, we can identify that epochs (22, 24, 25, 26, 27, 28, 30) have acquired more than 98% of validation accuracy with validation loss of (0.0242, 0.2287, 0.0178, 0.0034, 0.0045, 0.1201, 0.0018). We can identify that epoch 30 has acquired maximum validation accuracy with minimum loss.

Table 5.2 denotes that the optimizedResNet50 has acquired more accuracy.

$$J\left(A, B\right) = \tag{5.9}$$

A metric called the Jaccard index or Jaccard similarity coefficient is illustrated in eq.4.9, where J denotes Jacarda distance, A and B are two distinct sets, it is used to assess how similar and diverse sample sets are. The magnitude of The ratio of the intersection to the cardinality of the union of the sample sets and it assesses consistency between finite sample sets. And the Jaccard Index obtained for optimizedResNet50 is: 99.03% where, as the ResNet50 is 88.5%.

Heatmaps were generated through Seaborn library [20] to visualize The utilization of a confusion matrix is a common practice in evaluating the performance of a model, as well as in visualizing the results to identify which classes are being misclassified and how well the model is performing overall and the total frequency of misclassifications is identified in figure 5.3 represents that the optimizedResNet50 has fewer misclassifications concerning ResNet50.

6. Discussion. We show that it is possible to The objective is to effectively categorize the brain tumor image into many classifications, including ”Meningioma, Glioma, Pituitary.” [21]. The challenge is identifying the tumor’s location more precisely and classifying it so that miss classifications can be reduced [22]. We have modified the RESNET50 CNN model by imposing dropouts and downsampling at each layer to avoid overfitting; we have used four layers where each layer has a primary block and identity blocks through which we have implemented multiple hidden layers internally. By using SELU, we have increased the performance of neural networks, which consists of self-normalizing aspects by which the network can automatically adjust

Table 6.1: Comparison of Brain Tumor Detection Techniques

Reference	**No. of Images**	**Dataset Source**	**Technique Used**	**Accuracy (%)**
S. Solanki et al. [1]	1074	BraTS 2018	CNN	95.71
Gu et al. [2]	3064	REMBRANDT	CDLLC on CNN	97.49
Deepak et al. [3]	1426	Figshare	SVM with CNN	96.92
Kumar et al. [4]	3064	Figshare	RNGAP model on CNN	98.18
Rehman et al. [5]	1074	BraTS 2018	3DCNN	93.77
Rajasree et al. [6]	374	BraTS 2015	MSMCNN	97.46
Abd El Kader et al. [7]	3064	Figshare	HSANN	98.43
Bodapati et al. [8]	1074	BraTS 2018	ELM	96.9
Mzoughi et al. [9]	1074	BraTS 2018	3DCNN	97.59
Sajjad et al. [10]	121	Radiopaedia	Deep-CNN	96.68
S. Das et al. [11]	3064	Figshare	CNN	95.49
Sadad et al. [12]	3064	Figshare	ResNet50	96.14
Proposed Methodology	3064	Figshare	OptimizedResNet50	98.4

the parameters by which a stable network is obtained. Additionally, SELU can help minimize overfitting and lead to a regularization effect [23]. Further LogSigmoid activation function is used, which allows more complex relations between training images and testing images and can help minimize overfitting by introducing the nonlinearity aspect into the model [24]. For multiclass classification, we use cross-entropy loss, which is more robust to noisy labels. Moreover, its output is distributed over multiple classes while classifying the nonlinear relationships. And in order to minimize the local convergence, SGDM optimizer is utilized for enhancing the computational efficiency over large datasets to train the deep neural network [25].

In this way, we have implemented the proposed optimized Resnet50 architecture for performing multiclass classification to identify which type of tumor belongs to and which medical facilities can achieve proper treatment. Many models exist in the literature, but we have obtained improved accuracy and a promising strategy for identifying the tumor and performing multiclass classification [26-30].

Table 6.1 represents the % accuracy of the suggested optimisedResNet50 model was found to be 99.3%. which is much better than the existing ones in the literature.

7. Conclusion. One of the most complex difficulties to solve was categorizing brain malignancies. Magnetic resonance imaging (MRI) scans are being employed to examine and classify glioma, meningioma, and pituitary tumors. To prevent the issue of overfitting, we employed a 4-layer model with dropouts and downsampling at each layer. Each layer had primary and identity blocks, which we used to create internal hidden layers. By utilizing SELU, we have improved the performance of neural networks. These networks include self-normalizing features that enable them to modify the parameters necessary to produce stable networks autonomously. To do multiclass classification and help medical facilities determine which type of tumor a patient has, we have built the suggested optimized Resnet50 architecture. Although numerous models exist in the literature, we have improved accuracy and an approach that shows promise for locating tumors and conducting multiclass classification with a 99.03% accuracy rate. The proposed method concentrated the performance efficiency compared to earlier models.

As a part of our future work, we will review the overall performance of our differential 2D ResNet50 model by increasing the network coverage by tweaking the differential filter's parameters. We will enhance deep network architectures by using a multi-channel classifier that improves classification performance more significantly than before.

REFERENCES

[1] Gu, X., Shen, Z., Xue, J., Fan, Y., & Ni, T. (2021) , *Brain tumor MR image classification using convolutional dictionary learning with local constraint.* Frontiers in Neuroscience, 15, 679847.

[2] 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.

[3] Kumar, R. L., Kakarla, J., Isunuri, B. V., & Singh, M. (2021), *Multi-class brain tumor classification using residual network and global average pooling.* Multimedia Tools and Applications, 80, 13429-13438.

[4] Rehman, A., Khan, M. A., Saba, T., Mehmood, Z., Tariq, U., & Ayesha, N. (2021), *Microscopic brain tumor detection and classification using 3D CNN and feature selection architecture.* iMicroscopy Research and Technique, 84(1), 133-149.

[5] Rajasree, R., Columbus, C. C., & Shilaja, C. (2021) , *Multiscale-based multimodal image classification of brain tumor using deep learning method.* Neural Computing and Applications, 33, 5543-5553.

[6] Abd El Kader, I., Xu, G., Shuai, Z., Saminu, S., Javaid, I., & Salim Ahmad, I. (2021), *Differential deep convolutional neural network model for brain tumor classification.* Brain Sciences, 11(3), 352.

[7] Bodapati, J. D., Shaik, N. S., Naralasetti, V., & Mundukur, N. B. (2021), *Joint training of two-channel deep neural network for brain tumor classification.* Signal, Image and Video Processing, 15(4), 753-760.

[8] Mzoughi, H., Njeh, I., Wali, A., Slima, M. B., BenHamida, A., Mhiri, C., & Mahfoudhe, K. B. (2020), *Deep multi-scale 3D convolutional neural network (CNN) for MRI gliomas brain tumor classification.* Journal of Digital Imaging, 33, 903-915.

[9] Sajjad, M., Khan, S., Muhammad, K., Wu, W., Ullah, A., & Baik, S. W. (2019) , *Multi-grade brain tumor classification using deep CNN with extensive data augmentation.* Journal of computational science, 30, 174-182.

[10] Das, S., Aranya, O. R. R., & Labiba, N. N. (2019, May) , *Brain tumor classification using convolutional neural network.* In 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT) (pp. 1-5). IEEE.

[11] Sadad, T., Rehman, A., Munir, A., Saba, T., Tariq, U., Ayesha, N.,& Abbasi, R. (2021) , *Brain tumor detection and multi-classification using advanced deep learning techniques.* Microscopy Research and Technique, 84(6), 1296-1308.

[12] Sadad, T., Rehman, A., Munir, A., Saba, T., Tariq, U., Ayesha, N., & Abbasi, R. (2021) , *Brain tumor detection and multi-classification using advanced deep learning techniques.* Microscopy Research and Technique, 84(6), 1296-1308.

[13] Solanki, S., Singh, U. P., Chouhan, S. S., & Jain, S. (2023) , *Brain Tumor Detection and Classification using Intelligence Techniques: An Overview.* IEEE Access.

[14] Lakshmi, M. J., & Nagaraja Rao, S. (2022) , *Brain tumor magnetic resonance image classification: a deep learning approach.* Soft Computing, 26(13), 6245-6253.

[15] Jun, W., & Liyuan, Z. (2022) , *Brain Tumor Classification Based on Attention Guided Deep Learning Model.* International Journal of Computational Intelligence Systems, 15(1), 35.

[16] HS, S. K., & Karibasappa, K. (2022). , *An effective hybrid deep learning with adaptive search and rescue for brain tumor detection.* Multimedia Tools and Applications, 81(13), 17669-17701.

[17] Lakshmi, M. J., & Nagaraja Rao, S. (2022) , *Brain tumor magnetic resonance image classification: a deep learning approach.* Soft Computing, 26(13), 6245-6253.

[18] Asthana, P., Hanmandlu, M., & Vashisth, S. (2022) , *The proposition of Possibilistic sigmoid features and the Shannon-Hanman transform classifier along with the pervasive learning model for the classification of brain tumor using MRI.* Multimedia Tools and Applications, 81(17), 23913-23939.

[19] Eunice, Jennifer and Popescu, Daniela Elena and Chowdary, M Kalpana and Hemanth, Jude (2022) , *Deep learning-based leaf disease detection in crops using images for agricultural applications.* Agronomy, 10,12, 2395, MDPI.

[20] Lundervold, A. S., & Lundervold, A. (2019) , *An overview of deep learning in medical imaging focusing on MRI.* Zeitschrift für Medizinische Physik, 29(2), 102-127.

[21] Bonte, S., Goethals, I., & Van Holen, R. (2018) , *Machine learning based brain tumour segmentation on limited data using local texture and abnormality.* Computers in biology and medicine, 98, 39-47.

[22] Islam, A. T., Apu, S. M., Sarker, S., Shuvo, S. A., Hasan, I. M., Alam, A., & Dipto, S. M. (2022, December) , *An Efficient deep learning approach to detect Brain Tumor using MRI images.* In 2022 25th International Conference on Computer and Information Technology (ICCIT) (pp. 143-147). IEEE.

[23] Murthy, A and Rani, Paritala Jhansi and Almakassees, Sarah Majeed and Saikumar, K and Saleh, Mohammed and Ettyem, Sajjad Ali (2023) , *A novel classification model for high accuracy detection of Indian currency using image feature extraction process.* AIP Conference Proceedings,2845,1,050028.

[24] Srinivasarao, Gajula and Rajesh, Vullanki and Saikumar, Kayam and Baza, Mohamed and Srivastava, Gautam and Alsabaan, Maazen (2023) , *Cloud-Based LeNet-5 CNN for MRI Brain Tumor Diagnosis and Recognition.* Traitement du Signal, 40,4,pp. 1581–1592.

[25] Sajjad, M., Khan, S., Muhammad, K., Wu, W., Ullah, A., & Baik, S. W. (2019) , *Multi-grade brain tumor classification using deep CNN with extensive data augmentation.* Journal of computational science, 30, 174-182.

Edited by: Polinpapilinho Katina
Special issue on: Scalable Dew Computing for Future Generation IoT Systems
Received: Oct 17, 2023
Accepted: Feb 1, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1681–1701, DOI 10.12694/scpe.v25i3.2713

LOCALIZATION OF DIELECTRIC ANOMALIES WITH MULTI-LEVEL OUTLIER DETECTION THROUGH MEMBERSHIP FUNCTION AND ENSEMBLE CLASSIFICATION FRAMEWORK

MD. NAJUMUNNISA [*], ASCS SASTRY [†], AND B T P MADHAV [‡]

Abstract. This research presents an innovative method for real-time detection of dielectric anomalies, with a primary focus on evaluating apple quality and ripeness using dielectric tomography. The study involves the development of an advanced tomography system within an anechoic chamber, harnessing electromagnetic wave technology and sophisticated antenna systems for data acquisition. The proposed framework encompasses critical stages, including data collection, range bounds computation, threshold determination, class membership assignment, and ensemble classification. By seamlessly integrating statistical methods, density-based clustering, and ensemble learning, this approach significantly enhances precision and reliability in anomaly detection. The integration of available statistical methods, density-based clustering, and ensemble learning may demand substantial computational resources, limiting the scalability and real-time applicability of the proposed framework. Empirical results demonstrate the superior performance of the method, with an accuracy rate of 98.9 %, precision of 0.989, F-measure of 0.989, dielectric anomaly recall rate of 0.99, and a low error rate of 0.18. Overall, this research introduces an advanced approach with the potential to revolutionize apple quality assessment and industrial processes across various sectors.

Key words: Dielectric anomaly; machine learning; statistical analysis; Fast Fourier transform; near-field imaging; spatial smoothing; VSWR, ensemble learning model.

1. Introduction.

Microwave imaging shows promise in detecting breast cancer early due to the differences in dielectric properties of anomaly and non-anomaly exposed to microwave frequencies. This technology offers advantages over X-ray mammography and MRI, including non-ionizing radiation, non-compressive imaging, and cost-effectiveness. Despite these advantages, accurately identifying tumors within the background medium from scattered field data remains a challenge. Further research is needed to develop innovative solutions for precise and efficient breast cancer detection using microwave imaging. The use of microwave imaging has been shown to improve clinical outcomes for breast cancer patients by detecting the disease effectively. However, the nonlinearity and ill-posedness of inverse scattering problems pose significant challenges in this field. Nonetheless, recent advancements have revealed that a small number of observations can detect small scatterers and pinpoint their location with accuracy. But this method requires specific assumptions about the target geometry and data accuracy. Theoretical insights from literature sources [1].can be used to ensure the clear identification of multiple scatterers that remain unidentified. The MUSIC algorithm, a well-established linear technique, has been proven effective in detecting breast cancer using microwave imaging [2]. This method is the top choice for tomography of small inclusions in both full and limited-view inverse scattering problems[3]. It provides quasi-real-time qualitative microwave imaging, which is valuable for those in the field [4]. Additional research and development of these techniques show potential for advancing understanding of inverse scattering problems and their applications [5]. The multiple signal classification (MUSIC) algorithm is a popular technique that employs a well-established method. A precise determination of the nuareer of targets for spectrum computation is necessary to prevent the formation of artificial peaks and to maintain high performance in target localization using pseudo spectrum [6]. While the MUSIC algorithm has demonstrated success in detecting targets of varying sizes, a comprehensive structural analysis is essential to accommodate unforeseen circumstances, such as the emergence of anomalous signals or artifacts that cannot be explained through conventional means [7]. Thus, a detailed structural analysis is strongly advised to gain a more profound comprehension of the system

[*] Department of ECE, Koneru Lakshmaiah Education Foundation, AP, India. (2002040003@kluniversity.in).,
[†] Department of ECE, Koneru Lakshmaiah Education Foundation, AP, India. (ascssastry@kluniversity.in).
[‡] Department of ECE, Koneru Lakshmaiah Education Foundation, AP, India. (madhu.newlook@gmail.com).

and to improve the accuracy of target localization with the application of pseudo spectrum [8]. Gaining a comprehensive understanding of these complex phenomena is crucial for enhancing the overall effectiveness of the methodology [9]. Deriving analytical expressions for the scattered field caused by a slender inclusion is a challenging task that necessitates a deep comprehension of the relationship between the Bessel function of the first kind and the Hankel function of the second kind [10]. Rigorous numerical methods were employed to calculate the eigenvectors and eigenvalues of the MSR matrix to overcome this challenge [11].

The utilization of computers in tomography is referred to as Computerised Tomography, Computer Tomography, or Computer Assisted Tomography (CAT). Tomography has widespread use in medical diagnostics, including MRI, CAT, PET, and ultrasound [12]. Proton Computer Tomography was created specifically for imaging objects the size of a head. Unlike traditional X-rays, which only provide a basic view of an object's outline, tomography offers a comprehensive view of the object's internal composition [13]. The process of reconstruction is utilized to identify the distribution of attenuation in all directions encompassing an object, and CT technology presents the object's cross-sectional data [14]. Industrial tomography has its own set of system requirements for sensor configurations and imaging modalities that differ from those of medical tomography. The X-ray beam intensity is detected in a single direction, which is called a projection [15]. The method of creating an image from projections is used to produce cross-sectional data about an object [16]. Tomography is a diagnostic method that avoids overlapping shadows of structures in front of and behind the region under examination. A tomogram in diagnostic medicine provides a visual representation of the examined area. The exploration of utilizing microwave radiation to capture images of living organisms (at frequencies ranging from 1 to 10 GHz) is a recently developed field. Some people have mistakenly believed that wavelengths in the decimetric range are too lengthy to produce precise spatial resolution or that shorter wavelengths would cause excessive attenuation in the organisms of interest, resulting in limited advancement in this area. Nevertheless, these calculations failed to acknowledge that the wavelength can be reduced when passing through materials with high relative permittivity [17].

Furthermore, enhancing the resolution can be achieved by reducing the aperture size. By using suitable mathematical models, the dispersion exhibited in these measurements can be corrected [18]. The dielectric conductivity of the biosystem is a significant factor in the interaction between microwave energy and biosystems, which is dependent on the complicated permittivity and frequency of the interrogating radiation. Filling the antenna with a higher permittivity material can improve spatial resolution while maintaining its radiating characteristics, by reducing the aperture size of the antenna [19]. The near field of view of an antenna can be utilized to increase the resolution of an imaging system. The utilization of the complete time delay and attenuation measurement through an object can be applied for this purpose at frequencies used in microwaves. X-ray tomography methods rely on the assumption of a straight path approximation, where it is assumed that the ray travels directly to the receiver [20]. The only straight line connecting the transmitter and the observation point is where the ray arriving at each point on the observation plane can be linked to certain properties of the medium. Image reconstruction has utilized various mathematical techniques that have been crucial in medical imaging. There are two main types of reconstruction: analytical reconstruction and iterative reconstruction. Analytical reconstruction involves using precise mathematical solutions to image equations. A common method for this is filtered back projection, which combines all ray sums passing through a point to estimate its density. This technique is widely used in X-ray scanners, but its ability to produce high-resolution images is limited by band-limiting. This means that the image cannot contain spatial frequencies greater than a certain maximum frequency. Iterative techniques can also be used to solve image equations [21]. The Simultaneous Iterative Reconstruction Technique (SIRT), which is a modified version of ART, divides the object into pixels and calculates their values. In situations where there are not enough projections, the samples are noisy, or the projections are taken at limited angles, the iterative method is more efficient than the analytical method [22]. It is becoming less common to conduct mass screening for female breast cancer using X-ray CT due to increased awareness of the risks associated with ionizing radiation. Modern scanners typically acquire data at regular intervals and use analytical reconstruction algorithms that are faster and work well with sufficient sampling. Two types of tomographic methods, diffraction tomography and nonlinear deterministic techniques, are based on the inverse scattering problem. Computed tomography using X-rays has transformed biomedical imaging. The use of waves or low-level microwaves can result in lower quality reconstructions. As a result, alternative energy sources like ultrasound and low-level microwaves have become popular [23]. However, tomographic reconstructions made using electromagnetic or sound waves are of worse quality than those made using X-rays. This is because X-rays move in straight lines and do not diffract, which enables the transmission data to measure the line integral of a particular object parameter along straight line. The Fourier Slice Theorem can be applied in this case. Nonetheless, tomographic imaging using sound waves or low-level microwaves may not result in energy flowing in a straight line, leading to lower quality reconstructions. Imaging objects with large inhomogeneities requires accounting for refraction and various pathways of energy transmission.

An algebraic reconstruction approach with digital ray tracing and ray linking algorithms can address this issue. However, when the object's inhomogeneities match the wavelength, understanding energy transmission requires considering the wavefronts and fields scattered by the inhomogeneities. Therefore, reconstructing an object's constitutive parameter distributions using electromagnetic waves is an inverse scattering problem [24].

2. Related Work. Traditional CT algorithms like ART are inadequate for imaging objects with large inhomogeneities as they do not consider these effects. The analysis of wavefields can be facilitated by using the iterative method. To explain the propagation of waves in a uniform material, a wave equation, which is a second-order linear differential equation, is used. Directly solving the problem of wave propagation in a non-uniform medium is currently not feasible. Nevertheless, approximate methods based on the theory of wave propagation in uniform media can be applied to obtain solutions in the presence of weak inhomogeneities [25]. The iterative method appears to be a viable option when there are inadequate projections, noisy samples, and restricted projection angles for wavefield analysis. Low-dose radiation techniques are utilized for mass screening of female breasts, resulting in a significant reduction of risks associated with traditional X-ray mammography [26]. The analytical technique is widely preferred among other methods due to its speed and effectiveness with adequate sampling. Two types of tomographic methods, diffraction tomography and nonlinear deterministic technique, are based on the inverse scattering problem. X-ray computed tomography has transformed biomedical imaging. Using X-ray CT for imaging is not suitable anymore. Consequently, alternative energy sources like ultrasound and low-level microwaves have become more popular. However, tomographic reconstructions created with electromagnetic or sound waves are not as good as those made with X-rays [27]. This is because X-rays move in straight lines and do not diffract, allowing the transmission data to measure a specific object parameter along these lines. The Fourier Slice Theorem can be used now [28]. Nevertheless, during tomographic imaging, energy does not always flow in a straight line. For this task, sound or microwaves can be utilized. Energy propagation is characterized by refraction and various pathways, particularly when the objects in homogeneities are greater than the wavelength.

The algebraic reconstruction approach, in combination with digital ray tracing and ray linking algorithms, can partially address the bending caused by refraction. However, when the object's inhomogeneities match the wavelength, the energy transmission must be explained in terms of the wave fronts and fields that the inhomogeneities scatter. Consequently, the reconstruction of an object's constitutive parameter distributions using electromagnetic waves is more appropriately considered as an inverse scattering problem [29]. Traditional CT algorithms do not consider these effects. Certain wave fields have characteristics that make the ART method unsuitable. In order to describe wave propagation in a uniform object, a wave equation, which is a second-order linear differential equation, is utilized. Direct methods are currently unable to resolve the issue of wave propagation in a non-uniform medium. However, approximate formalisms can be used to generate solutions in the presence of weak inhomogeneities by utilizing the theory of homogeneous medium wave propagation [29].

The reconstructed images feature dual mesh technology, which is well-suited for systems that use finite element methods. On the other hand, systems that utilize finite difference methods require a thick mesh. Exposure to high-frequency electromagnetic sources is crucial, and displaying the electric fields over a body accurately is necessary due to their frequent and rapid changes. However, the complex wave numbers tend to remain constant in several sub regions, which may result in incomplete sampling of this characteristic in certain body parts. In areas where twin mesh technology was initially implemented, electric fields are calculated using a uniformly dense mesh type. The second mesh, which is less uniform and less dense overall, reveals the complex wavenumber k distribution within the intended region. The Maxwell equation is utilized in these computations. The log-magnitude and phase of the electric signal observed through field measurements were incorporated [30]. This change allows for the detection of phase variation and its spread across multiple

Rieman sheets in the complex plane. The simulation study and microwave imaging tests conducted showed notable enhancements in image quality for significant, high-contrast objects. It was suggested to apply simple visualization and unwinding methods to determine phase values based on the transmitter and receiver positions. A new approach was introduced to restore images of high quality. The objective is to capture images of objects without any prior assumptions about their size and contrast. Several optimization algorithms have been recorded recently to accomplish this task. Nevertheless, identifying targets remains a significant challenge in general. Medical imaging strives to determine an object's size, location, and constitutive parameters, including its refractive index, specifically for intricate objects. For example, when measuring an object from the dispersed field, electromagnetic waves illuminate it successively. In recent times, effective reconstruction methods have been formulated to address this non-linear and ambiguously defined issue. In earlier investigations concerning microwave imaging, the primary focus was on obtaining images of objects. The use of temporal domain techniques helped to simplify the assumptions of wave propagation. These techniques relied heavily on integral equations in the field, which acted as both interior and exterior integral representations of the scattering object [30]. that many of these methods are iterative, with each iteration requiring the solution of the forward problem. The network and location receive dispersed electric field data from an item, resulting in the production of dielectric permittivity output and the cylinder's radius. The evaluation of results considers input and output parameters, as well as various test data sets.

The algorithm's effectiveness suggests that it can be beneficial for real-time remote sensing applications by solving the inverse scattering problem. Additionally, body reconstruction is also a part of this research. Sharp contrasts can be attained by utilizing a genetic algorithm that concentrates on images. This necessitates converting the Inverse Problem into a global nonlinear optimization problem. In [30], a hybrid genetic algorithm was employed to optimize the configurations of dielectrics for tomographic imaging. Another technique was also proposed, which included representing concealed inhomogeneities using multilayer infinite dielectric cylinders with elliptic cross sections. A cost function was formulated, and field terms were generated using Mathieu's series solution. The functionality was minimally reduced through an innovative optimization method. This method was the first to incorporate back projections with linear filters. They utilized wavelet data that was back projected to decompose the projection into one dimension and discovered an alternative basis for the ”Natural pixel” formulation of image reconstruction. The wavelet transform uses modulus sum in the phase-indicated direction. This algorithm utilizes wavelets' localization capability to create a local image reconstruction of the radon transform, which significantly reduces exposure and calculations for X-ray imaging by using nearly local data that corresponds to a region of the body's cross-section. The initial step of the process involves obtaining the quincunx approximation and specific coefficients of a function based on projections. The outcomes of the simulations have indicated an improvement in comparison to the reconstruction that employs discrete wavelets that can be separated. The reconstruction of an image from its sampled projections in the form of Radon Transform Values was accomplished by using a raised-cosine wavelet. However, the projected data slices were found to be noisy when reconstructed with these data due to computed tomography measurements.investigated a new family of regularization techniques for reconstruction that utilizes wavelet and wavelet packet decomposition thresholding. This method is based on the concept that the decomposition almost diagonalizes the inverse and includes prior knowledge of the Radon transformation. In relation to medical images, it was found that these methods had better performance compared to filtered back. Iterative techniques such as OSEM and projection were utilized. Reconstruction of a low dose computed tomographic image leads to instability in its inverse, which then causes problems with noise. This reconstruction technique combines an algorithm with the regularized theory of filtered back projection.

The proposed method, known as the Dual Tree-Complex Wavelet Transform (DT-CWT), can effectively decrease jitter and noise without relying on any specific assumptions about the noise model. Moreover, a relatively simple thresholding technique for recovering functions from noisy data was suggested. In the context of coordinates, the non-linear soft thresholding was subjected to the empirical wavelet coefficient using a method referred to as ”Visu shrink.” This approach provides an optimal estimation of the mean square error for unknown smoothness functions at a particular location, along with a good visual quality estimator. The original data can be restored by utilizing pyramid filtering after flipping it. A system has been established for this purpose. The wavelet-based technique ”SURE shrink” provides minimax rates of confluence across all spaces. The shrink thresholds are adaptively based, and Stein's Unbiased Risk Assessor (SURE) was employed to determine them. Additionally, a practical spatially adaptive method has been developed. The ”Risk shrink” technique is utilized by an adaptive system to decrease the size of empirical wavelet coefficients. This technique involves using different mapping techniques and a discriminative strategy to transform coefficients into functions that produce an open source, noise-free image. used a bilateral denoising filter for images that lacked edge rounding. This technique involves using a spatial non-linear filter to average the data. The approximate (low frequency) sub bands of a signal are filtered bilaterally after being decomposed. The effectiveness of removing noise in real, noisy photographs can be achieved using wavelet filter bank and bilateral multiresolution filter.

A new denoising technique has been developed by measuring the importance of noise-free wavelet coefficients using a local window to define significance, which is known as the ”signal of interest.” Efficient denoising techniques have also been developed by utilizing the rarity and decorrelation features of DWT, with the method relying on an empirical Bayes estimate based on Jeffrey's uninformative prior. Objective wavelet-based Bayesian denoising can be achieved through a simple fixed non-linear shrinkage rule that outperforms alternative methods that require extensive calculations. proposed the use of Bayesian analysis to construct bivariate non-Gaussian distributions and associated non-linear models, which can be used as new shrinkage functions. Assuming that the coefficients of shrinkage functions have no relation to wavelets is not recommended. repaired damaged images of Gaussian noise by combining a bilateral filter with a trivariate shrinkage filter based on wavelets, and they took into account the wavelet coefficients of the trivariate Gaussian distribution in the wavelet domain. The Maximum-a-Posteriori (MAP) estimator was then used to generate a result, considering the statistical correlations between intrascale wavelet coefficients. introduced the trivariate shrinkage function. A new method for denoising medical images in the wavelet domain was proposed, which uses a technique called guided complex shrinking. This method preserves edges and corners while considering features such as orientation. Additionally, an efficient statistical approach for analyzing medical image wavelet coefficients was presented using a combined denoising method. They developed a novel bivariate Laplacian probability density function model with heavy tails to mimic the statistical data of the wavelet coefficients. A straightforward nonlinear shrinkage function was derived to produce noise-free images. A powerful method has been developed to address both noise and blur in reconstructed images. The algorithm includes two steps. The first step involves using a modified version of the Dual Tree-Complex Wavelet Complex Fourier Wavelet Regularized Transform (DT-CWT) called ComForWaRD for pure denoising, along with a generalized Wiener filter and global blur correction. In the second step, a new technique called BiComForWaRD, which is a variation of ComForWaRD, is used specifically for medical imaging. This two-step procedure includes a denoising algorithm corrected using a generalized Wiener filter and a local adaptive bivariate shrinkage function. Overall, this novel algorithm is a powerful tool that can improve the quality of reconstructed images by addressing both noise and blur. The denoising algorithm called dependencies makes use of wavelet coefficients and statistical dependence statistics. The Laplacian probability density function-based pyramid with local dependencies is a denoising technique that can be utilized. Wavelet soft thresholding was suggested by Grace Chang and S. Al as a flexible data-driven approach for picture denoising. Within the Bayesian framework, a prior is frequently utilized to determine the threshold applied to wavelet coefficients. The Generalized Gaussian Distribution (GGD) was proposed as a framework for image processing applications [44]. They introduced a novel statistical signal processing approach in the statistical domain that is wavelet-based, utilizing Hidden Markov Models.

1. Scanning: Scanning is the initial step in the image reconstruction process. It involves acquiring raw data from the object or scene of interest using a suitable imaging device, such as a CT (Computed Tomography) scanner, MRI (Magnetic Resonance Imaging) machine, or a similar medical or industrial imaging system. During scanning, data is collected in the form of signals or measurements that capture the properties of the object being imaged.

2. Discretization: Once the raw data is collected, it often needs to be discretized. Discretization involves converting the continuous data obtained from the scanning process into a discrete format that can be processed and reconstructed digitally. This step typically involves techniques such as sampling and quantization, where the continuous measurements are transformed into discrete data points or pixels, allowing for digital processing.

3. Back-Projection: Back-projection is a crucial step in image reconstruction, especially in the context of tomographic imaging techniques like CT scans. It involves mathematically reconstructing a two-dimensional

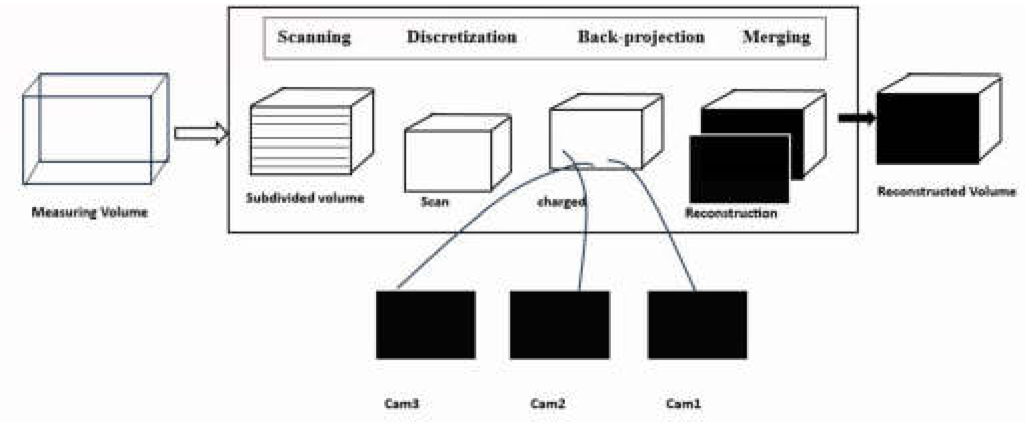


Fig. 3.1: Reconstruction process

or three-dimensional image from the collected data. Back-projection algorithms distribute the discretized data back into a grid or image space, attempting to recreate the distribution of properties within the imaged object. This process is iterative and involves mathematical computations to refine the image reconstruction.

4. Merging: In some cases, multiple sets of data from different scans or imaging angles are merged to create a more comprehensive and accurate image. This merging process combines the reconstructed images from various perspectives to create a single, fused image with improved spatial resolution and reduced artifacts. This is particularly important in techniques like CT, where multiple X-ray projections are acquired from different angles and merged to form a 3D image.

Each of these steps plays a crucial role in the image reconstruction process, allowing us to transform raw data into meaningful and interpretable images. The choice of algorithms, the quality of data acquisition, and the precision of the reconstruction techniques significantly impact the final image's accuracy and clarity, making this process essential in fields such as medical imaging, industrial inspection, and scientific research.

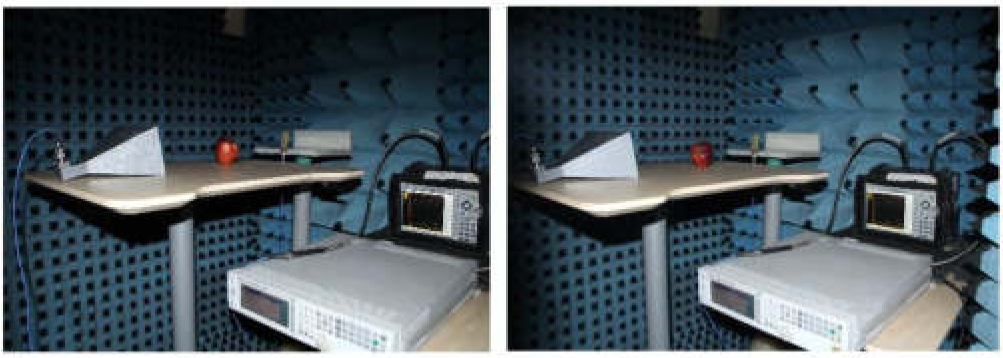
Main contribution of the research:

- Presents a comprehensive approach for dielectric anomaly detection. The study begins by applying the FFT data transformation approach to the raw input data.
- Followed by proposing a gamma quartile distribution-based method for anomaly detection on the transformed data.
- Additionally, a hybrid ensemble learning classifier is introduced to identify dielectric anomalies in real-time test data. As the conventional dielectric anomaly detection models primarily rely on static raw signal data with signal values
- Developed method enhances the overall anomaly detection process of dielectric raw datasets by incorporating statistical range bound analysis and an ensemble classification approach.

The structure of the paper is as follows: Section 3 follows with the proposed framework describing with its system, antenna design, tomographic image reconstruction methodology with their algorithms. Section 4 presents the proposed ensemble learning model. Section 5 illustrates the experimental results and section 6 concludes the research work.

3. Proposed Framework. In this work, primarily an antenna is designed which operates at 0.915GHz and 2.45GHz frequencies. Here apple is taken as the object under test (OUT) and the anomaly on the defected apple has to be reconstructed. The values of dielectric properties like dielectric constant and dielectric loss factor of the object at 0.915 GHz frequency has been considered.

Figure 3.1 illustrates the reconstruction process of an image, which typically involves several key steps, including scanning, discretization, back-projection, and merging. The dielectric constant and dielectric loss factor values of apple at 0.915GHz are 57 and 8. The object is placed in between the proposed antenna and the transmitting antenna, and the signals are collected from the VNA (Vector Network Analyzer).



(a) Healthy Object (b)Defected Object

Fig. 3.2: Experimental setup for tomography reconstruction

3.1. System design. In this section, we provide details about the developed tomography system, which was carried out within an anechoic chamber and, in this context represents a healthy apple and a defective apple. The object under examination is positioned between two antennas: the proposed antenna (acting as the receiving antenna) and the transmitting antenna, which are positioned opposite each other, as illustrated in this section. Specifically, the transmitting antenna is a horn antenna, responsible for emitting signals, while the proposed antenna functions as the receiving antenna.To initiate the experiment, a source generator is connected to the transmitting antenna to provide the necessary excitation, and the receiving antenna is linked to an Agilent Vector Network Analyzer (VNA) to capture the scattered field data. During data collection, the position of the receiving antenna is incrementally shifted in 5-degree phases, and the signals are then routed from the VNA for analysis. It's important to note that this experimental setup takes place within an anechoic chamber, which is equipped with microwave-absorbing materials. These materials serve the crucial role of blocking external radio frequency interference, thus reducing the influence of environmental electromagnetic effects. The use of these absorbers is essential to optimize the performance of the prototype system and achieve accurate and reliable results.

In summary, this section outlines the configuration of the tomography system, its placement within an anechoic chamber, and the equipment and procedures involved in experiments, all aimed at minimizing external electromagnetic interference and ensuring the effectiveness of the developed system which was shown in Figure 3.2.

3.2. Antenna design. In microwave tomography systems, the use of highly directional antennas is common to ensure effective illumination of the object under test (OUT). Figure 3.2 illustrates the design of a monopole antenna created on FR4 material with specific properties, including a dielectric constant of 4.3 and a thickness of 1.6 mm.The antenna's main radiating element, known as the patch, is fabricated from copper material with a thickness of 0.035 mm and incorporates a partial ground plane on the opposite side of the antenna. Additionally, a microstrip feedline is integrated into the design to facilitate the efficient transmission of electromagnetic signals to the patch.To ensure proper impedance matching between the patch and the 50-ohm probe line, careful consideration was given to impedance matching techniques. Furthermore, metamaterials were strategically placed on both sides of the feed line to enhance the antenna's performance.The Computer Simulation Technology (CST) software was employed to fine-tune and optimize the antenna design. The simulated and measured reflection coefficients, often represented as S11 in antenna engineering, were analyzed and compared. The computational model presents the results of this comparison, showcasing how closely the simulated data aligns with the measured data, which is critical for validating the antenna's efficiency. The performance measures, on the other hand, display the antenna patterns. These patterns illustrate how the antenna radiates electromagnetic energy in different directions, highlighting its directional characteristics. Accurate antenna patterns are essential for directing and focusing the emitted signals toward the object under

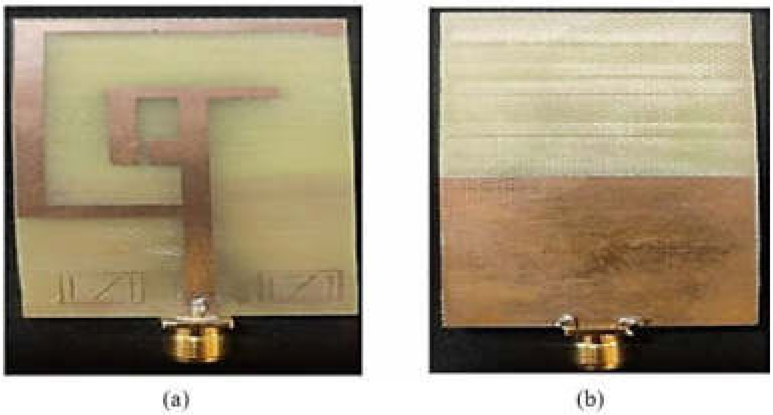


Fig. 3.3: Fabricated antenna designed on FR4. (a) front view, (b) back view

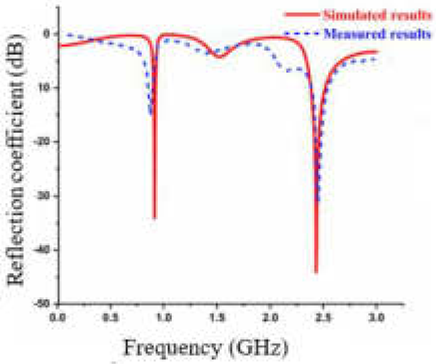
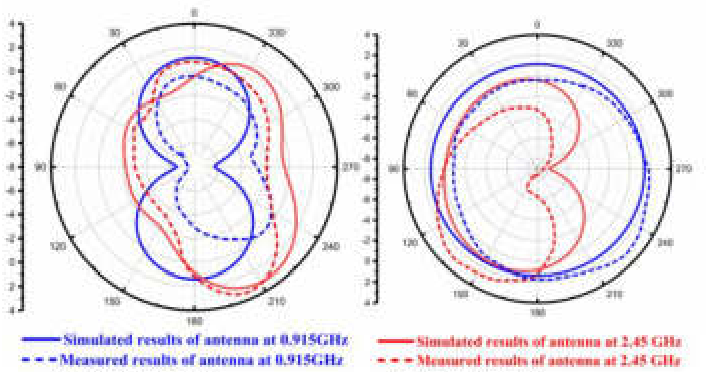


Fig. 3.4: Simulated Measure S11 parameter

test during the tomography process, ensuring optimal data acquisition and image reconstruction.In summary, the described monopole antenna design, constructed with specific material properties, impedance matching techniques, and metamaterial enhancements, underwent rigorous simulation and measurement procedures to validate its performance in terms of reflection coefficients and radiation patterns, all of which are vital for its successful application in microwave tomography systems.

In Figure 3.3 (a), we examine the results of simulated and measured reflection coefficients and VSWR (Voltage Standing Wave Ratio) for the antenna operating at two distinct frequencies: 0.915 GHz and 2.45 GHz. The bandwidth considered for both frequencies is 20 MHz. These parameters are crucial in assessing the antenna's performance and impedance-matching characteristics. Figure 3.3 (b), on the other hand, provides insights into the radiation patterns exhibited by the proposed antenna at the aforementioned operating frequencies, observed along both the E plane (Electric Plane) and H plane (Magnetic Plane). At the lower frequency of 0.915 GHz, the radiation pattern in the E plane indicates a bidirectional nature, meaning that the antenna emits signals in two primary directions.

Simultaneously, along the H plane, the radiation pattern is omnidirectional, implying that the antenna radiates signals in all directions evenly. As the frequency increases to 2.45 GHz, the radiation pattern changes. In the E plane, it exhibits dipole characteristics, suggesting a preference for signal emission in two opposing directions. Along the H plane, it remains nearly omnidirectional, signifying uniform radiation in all directions. These findings highlight the frequency-dependent nature of the antenna's radiation patterns, with its behavior



(a) Along E plane (b) Along Hplane

Fig. 3.5: Simulated and Measured radiation pattern results of proposed antenna

transitioning from bidirectional and omnidirectional at lower frequencies to dipole and omnidirectional at higher frequencies which was explained in Figure 3.4.

These insights are essential for optimizing the antenna's performance and its suitability for specific applications in different frequency bands shown in Figure 3.5

3.3. Tomographic image reconstruction methodology. The provided framework outlines a novel approach for dielectric anomaly detection using real-time sensor data. Let's break down the technical details step by step:

1. Data Collection: Initially, dielectric raw data is collected from real-time sensors. These sensors are responsible for capturing the electrical properties of materials under test, which are crucial for anomaly prediction.

2. Range Bounds Calculation: In this step, a gamma distribution-based method is employed to calculate the range bounds. These bounds are used to establish upper and lower thresholds for anomaly detection. Gamma distribution is a probability distribution often used to model data with positive skewness, which makes it suitable for modeling certain types of dielectric data.

3. Thresholding: For each signal in the raw data, both lower and upper bounds are computed based on the gamma distribution parameters. These bounds serve as reference points to filter out the extreme values in the data, specifically the upper outliers and lower outliers. This filtering process helps in identifying data points that deviate significantly from the expected range.

4. Class Membership Function: The filtered outlier data is then subjected to a class membership function. This function assigns a class label to each data point, categorizing them as either normal or anomalous based on their relationship to the computed bounds. This classification step is crucial for distinguishing between expected and unexpected behavior in the dielectric data.

5. Ensemble Classification: Finally, an ensemble classification framework is proposed for predicting dielectric anomalies in real-time test data. Ensemble methods typically combine the predictions of multiple classifiers to improve overall accuracy and robustness. This ensemble approach leverages the outputs of the class membership function to make predictions about dielectric anomalies, enhancing the reliability of the detection process.

visually represents this comprehensive dielectric anomaly detection framework. It showcases the entire workflow, from data collection and threshold calculation to the final ensemble classification step. This framework is designed to effectively identify anomalies in real-time dielectric data, which is valuable for various applications such as quality control, fault detection, and industrial processes shown in Figure 3.6.

Fundamental steps for the tomographic image reconstruction process. The proposed system involves a series of fundamental steps as follows:

1. Data Acquisition: The initial step revolves around gathering data from an array of sensors, such as microphone arrays or antenna arrays. This data typically comprises a collection of time-domain samples.

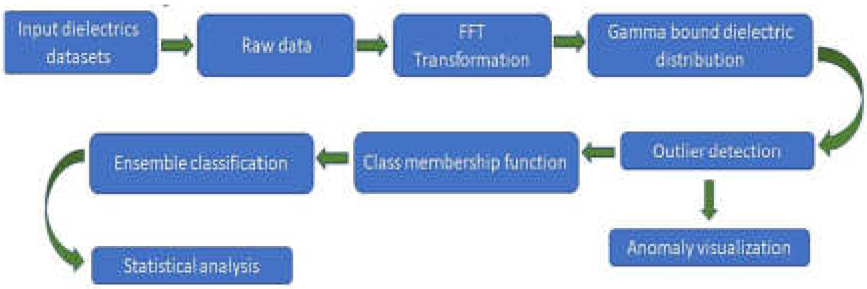


Fig. 3.6: Proposed distributed student dropout prediction model

2. FFT Transformation: Subsequently, the Fast Fourier Transform (FFT) is applied to the acquired data to yield its frequency-domain representation, enabling analysis in the spectral domain.

3. Covariance Matrix Estimation: Estimation of the covariance matrix is carried out using the received signal. This matrix reflects the correlation among various sensor outputs and is instrumental in deriving eigenvalues and eigenvectors.

4. Eigenvalues and Eigenvectors Computation: Eigenvalues and eigenvectors are computed based on the estimated covariance matrix. Eigenvectors signify the directions from which signals arrive, while eigenvalues provide insights into the signal strength in each direction.

5. Spatial Spectrum Construction: Leveraging the obtained eigenvectors and the FFT of the received signal, a spatial spectrum is constructed. This spectrum portrays the signal power across different directions, allowing for directional analysis.

6. Application of Dielectric Anomaly Detection Approach: The system then applies a dielectric anomaly detection approach to assign class labels based on the spatial spectrum and other pertinent information.

7. Validation of Class Labels: The assigned class labels undergo validation through inter and intra similarity distance assessments, ensuring the accuracy and reliability of the labeling process.

8. Data Storage as Training Data: The validated data is stored for subsequent use as training data, facilitating the system's learning and improvement over time.

9. Ensemble Classification Framework Application: An ensemble classification framework is employed to predict anomalies in dielectric test data. This framework incorporates multiple classification techniques for robust prediction.

10. Statistical Metric Analysis: The system performs an in-depth analysis of statistical metrics to evaluate the performance of the ensemble classification framework, providing insights into its accuracy and effectiveness. In essence, these technical steps collectively form a comprehensive system for processing and analyzing data from sensor arrays, with a focus on dielectric anomaly detection, classification, and performance evaluation.

3.4. Algorithm: Dielectric anomaly detection approach. Compute the following hybrid IQR method as

$$DF[x] = features(); \tag{3.1}$$

$$R1 = (V(|DF|/4)); \tag{3.2}$$

$$R2 = (V(|DF|/2) + V(|DF|/2+1))/2; \tag{3.3}$$

$$U\_E[x] = R3 + \eta.log(\Gamma(R3 - R1)) \tag{3.4}$$

$$L\_E[x] = R1 - \eta.log(\Gamma(R3 - R1)) \tag{3.5}$$

$$U\_Outlier = R3 + \eta.min(\chi(R3 + R1, 9)), exp(\Gamma(R3 - R1))) \tag{3.6}$$

$$L\_Outlier = R3 - \eta.min(\chi(R3 + R1, 9)), exp(\Gamma(R3 - R1))) \tag{3.7}$$

Algorithm: Interquartile Range outlines (IqR)

- **Initialization:** Begin with setting the loop index variable i to 1.
- **Loop Start:** Start a loop that iterates from $i = 1$ to the size of the set $|FS|$.
- **Access Element:** Retrieve the i-th element ($F(i)$) from the set F.
- **Check Range:** Examine whether the value of $F(i)$ falls above or below the predefined upper and lower range.
- **Anomaly Check:** If $F(i)$ is found to be above or below the specified range, mark it as an anomaly.
- **Non-Anomaly Check:** If $F(i)$ is within the acceptable range, label it as a non-anomaly.
- **Perform Actions:** Execute any necessary actions or logging related to the anomaly or non-anomaly status of the current element.
- **Increment Index:** Increment the loop index variable i to move on to the next element in the set.
- **Loop Continuation:** Repeat the loop until all elements in the set F have been processed.
- **Process Completion:** Finish the process, having labeled each element in the set F as either an anomaly or a non-anomaly based on its position relative to the upper and lower ranges.

The above algorithm outlines a hybrid method for outlier detection using the interquartile range (IQR) for dielectric data. Here's a step-by-step explanation of the algorithm:

Initialize the array DF[x] as the input features. (Assuming DF[x] contains the dielectric data.) Compute the first quartile range R1, which is the value at the 25th percentile of the absolute values of DF[x]. Compute the second quartile range R2, which is the average of the values at the 25th and 75th percentiles of the absolute values of DF[x]. Compute the third quartile range R3, which is the average of the values at the 75th percentile and 25th percentile from the end of the absolute values of DF[x].Compute the upper outlier threshold U_E[x], defined as R3 plus a parameter eta multiplied by the logarithm of the gamma function evaluated at R3 minus R1.Compute the lower outlier threshold L_E[x], defined as R1 minus a parameter eta multiplied by the logarithm of the gamma function evaluated at R3 minus R1. Compute the upper outlier limit U_Outlier, defined as R3 plus eta times the minimum of the chi function evaluated at R3 plus R1 with 9 degrees of freedom, and the exponential of the gamma function evaluated at R3 minus R1. Compute the lower outlier limit L_Outlier, defined as R3 minus eta times the minimum of the chi function evaluated at R3 plus R1 with 9 degrees of freedom, and the exponential of the gamma function evaluated at R3 minus R1. Perform outlier detection: Iterate through each feature F(i) in the dielectric data FS[x]. If F(i) lies above the upper outlier threshold U_E[x] or below the lower outlier threshold L_E[x], label it as an outlier. Otherwise, do further processing or analysis with the non-outlier data. The algorithm combines the IQR method with additional statistical measures (gamma function, chi function) and thresholds (U_E[x], L_E[x], U_Outlier, L_Outlier) to identify outliers in the dielectric data.

3.5. K-class membership validation approach.

Step 1: To each dielectric labeled partition data PD.
Step 2: To each data point p[i] in PD.
Step 3: Perform weight age density to each data object using the Gaussian transformation function as

$$wd_i = \sum_{j\epsilon, i!=j} exp(-(\frac{d_{ij}}{dc})^2) \tag{3.8}$$

$$d_{i,j} = \sqrt{\sum_{t=1}^{m}(x_i^t - x_j^t)^2} \tag{3.9}$$

Where d_c: Threshold

Step 4: To each object, find the highest density using the following measures

$$hd_j = max\{wd_i\} \tag{3.10}$$

Step 5: Computing the k nearest neighbors by using the k-randomized centres as k initial clusters as. The set of k nearest neighbors of center CPi is defined

$$NP_i^k = \{P_j/min(d_{ij}), i! = j\} \tag{3.11}$$

Step 6: Compute the inter cluster similarity and intra cluster similarity to each k-neighbor initial clusters using the following formula.

$$\lambda 1 = IntraClu(p_c, p_i) = \frac{1}{n_i - 1} hd_c . \sum_{m=1} d(p_c, p_m) \tag{3.12}$$

$$\lambda 2 = IntraClu(p_c, p_i) = min_{1<=m<=k}(\frac{1}{n_m} hd_c . \sum_{r=1} d(p_c, p_r)) \tag{3.13}$$

alpha =Q1, beta =Q2, chi =Q3

$$UppOutlier = \chi + \Gamma max\left\{\lambda 1, \lambda 2\right\}, (\chi - \alpha) \tag{3.14}$$

$$LowerOutlier = \chi - \Gamma max\left\{\lambda 1, \lambda 2\right\}, (\chi - \alpha) \tag{3.15}$$

Step 7: Iterate until all points are assigned to k=clusters or no more changes in clusters.

For each data point in a partition, a weightage density is computed using a Gaussian transformation function. The highest density among neighboring data points is then determined for each object. Initial clusters are formed by randomly selecting k centers and finding their k nearest neighbors. Inter-cluster and intra-cluster similarity measures are calculated for each initial cluster. The algorithm iteratively updates the cluster assignments until convergence or a stopping condition is reached. While specific formulas and details are missing, this algorithm aims to cluster data based on density and proximity, refining the assignments through iterative steps.

4. Proposed Ensemble learning model. Ensemble learning is a robust and effective technique that harnesses the collective intelligence of multiple base models to enhance overall performance and generalization in solving complex problems. When employed in the context of localizing dielectric anomalies, which entails the detection of regions within a material exhibiting atypical electrical properties, an ensemble approach can significantly improve the accuracy and reliability of the results.

1. Base Model Diversity: Ensemble learning starts by creating a set of diverse base models, each trained to recognize dielectric anomalies from different perspectives or with distinct algorithms. These models could include various machine learning techniques like decision trees, support vector machines, or neural networks, each with its own strengths and weaknesses.

2. Aggregated Predictions: Once the base models are trained, their individual predictions are aggregated to make a collective decision. This aggregation can be performed using methods like majority voting, weighted averaging, or stacking, depending on the specific ensemble strategy chosen.

3. Anomaly Identification: The aggregated predictions from the ensemble are then used to identify and localize dielectric anomalies within the material. This involves examining the consensus among the base models. Regions where a significant portion of the models agree on the presence of anomalies are more likely to be actual anomalies.

4. Thresholding: To improve the precision of anomaly localization, thresholding techniques can be applied. This means considering only those regions where the level of agreement among the base models surpasses a predefined threshold. This helps filter out false positives and enhances the robustness of anomaly detection.

5. Visualization: To make the results more interpretable, the localized dielectric anomalies can be visualized. This often involves overlaying the anomaly predictions onto the original dielectric property maps or other relevant visual representations. This allows for a clear visualization of where the anomalies are located within the material.

6. Performance Evaluation: The performance of the ensemble is rigorously evaluated using appropriate metrics like precision, recall, F1-score, and accuracy. This step ensures that the ensemble effectively identifies and localizes dielectric anomalies while minimizing false alarms.

7. Model Refinement: The ensemble's performance can be further improved by refining the base models or adjusting the ensemble strategy based on the evaluation results. This iterative process fine-tunes the ensemble for optimal performance.

Ensemble learning leverages the strengths of multiple models to provide accurate and robust dielectric anomaly localization. It combines their predictions, applies thresholding for precision, and visualizes the results, ultimately enhancing the understanding of regions with unusual electrical properties within the material. This approach is especially valuable in applications such as quality control, defect detection, and materials science.

Initialization of ensemble learning process "Set of proposed classifier and base classifiers are represented as ensemble classifiers as"

$$EC = \{KNN, HDT, Proposedmodel\} \tag{4.1}$$

$$MC = \{\} \,; //Modelclassifier \tag{4.2}$$

$$CO = \{\} \,; //classifierOutput \tag{4.3}$$

Procedure The procedure outlines a sequence of technical steps involved in constructing a Bayesian network using input data and estimating Bayesian network node variables based on two different cases, one for discrete attributes and the other for continuous attributes. Additionally, it introduces a novel feature ranking measure aimed at optimizing the decision tree construction process.

1.Data and Variables: The input data is represented as D. PAi represents the set of input random variables. PIj represents an instance of attribute.

2.Bayesian Network Construction: For each attribute PAi in data PD. Construct the Bayesian network graph using the input data and naïve Bayesian estimations.

3.Estimation of Discrete Attributes: Estimate Bayesian network node variables using the two cases. If the attribute type is discrete, then the Bayesian discrete parameters are estimated using the following measure as

$$P(A_i = I_{k/C_j}) = \frac{N_{ijk}}{N_i} \tag{4.4}$$

where N_ijk is the number of instances of class cj having the value Ik in attribute Ai.

4.Estimation of Continuous Attributes: If the attribute type is continuous type, then the Bayesian continuous parameters are estimated using the following measure as

$$P(A_i = I_k/C_j) = G(I_k, \mu_{ij}, \sigma_{ij}) \tag{4.5}$$

$$G(I_k, \mu_{ij}, \sigma_{ij}) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} e^{\frac{(I_k - \mu_{ij})^2}{2\sigma_{ij}^2}} \tag{4.6}$$

5. Feature Ranking for Decision Tree Optimization: In order to optimize the decision tree construction process, a novel feature ranking measure is proposed to optimize the problem of pruning.

To each feature in DF[]

do

Table 5.1: (Sample data) Dielectric FFT transformed data

S No	Frequency Sample	s1, 1/Rate	s2, 1/Rate	s3, 1/Rate	s4, 1/Rate
1	0.0	0.9943	0.1188	0.2415	0.2723
2	0.001	0.9945	0.1268	0.2412	0.2743
3	0.002	0.9950	0.1519	0.2397	0.2723
4	0.003	0.9957	0.1877	0.2380	0.2711
5	0.004	0.9965	0.2239	0.2365	0.2700

Finding rank of the feature as

$$S = D.log(D); \tag{4.7}$$

$$p1 = -s/((\sqrt{\sum D[i]})^3 * \sqrt{\chi(D) * \sum D[i]}) \tag{4.8}$$

$$p2 = -(s * CE(D))/(\chi(D) * \sum D[i])^3 \tag{4.9}$$

$$Rank(A[i]) = Max\,\{p1, p2\} \tag{4.10}$$

6.Optimizing Pruning: This feature ranking measure is utilized to optimize the decision tree construction process, specifically in the context of pruning. In essence, this technical process involves constructing a Bayesian network, estimating parameters for both discrete and continuous attributes, and introducing a novel feature ranking measure to enhance the decision tree construction process, ultimately optimizing the problem of pruning. These steps are fundamental in various data-driven applications, including machine learning and data analysis.

5. Experimental Results. Dielectric tomography is a non-invasive imaging method that relies on the interaction between electromagnetic waves and matter to create detailed images of an object's internal structure. In the context of apple dielectric tomography, this technique is employed to ascertain the distribution of dielectric properties within an apple's interior. This information is valuable for assessing the apple's quality and degree of ripeness.

The data provided in the question seems to be a table of dielectric tomography measurements conducted on multiple apples. Each row within the table corresponds to a single measurement, while each column represents a different sensor or a group of sensors. The "Frequency Sample" column likely denotes the frequency of the electromagnetic wave employed during the measurement. Meanwhile, columns labeled "S1" through "S8" are associated with various sensors or sensor clusters utilized in the measurements. The "cluster" column appears to categorize each measurement into specific clusters, possibly based on certain criteria or sensor configurations. Lastly, the "Outlier" column serves to indicate whether a particular measurement is considered an outlier or not.

Table 5.1 encapsulates a portion of the collected data, which undergoes processing involving FFT (Fast Fourier Transform) transformation. This transformation is applied to convert the data from the frequency domain into a time-domain series. It's important to note that this data processing step is essential for further analysis and interpretation of the dielectric properties within the apples.In essence, this dataset and the described processing steps are instrumental in leveraging dielectric tomography to gain insights into the internal properties of apples, aiding in quality assessment and ripeness evaluation shown in table 5.1.

Table 5.2 presents the dielectric anomaly prediction values of the ranked attributes computed from the sample data of the defected apple which is placed in between the transmitter and receiver antennas. This table contains values related to the prediction of dielectric anomalies specifically computed for attributes that have been ranked. These attributes likely pertain to the sample data of a defected apple positioned between transmitter and receiver antennas. The prediction values could represent the likelihood or probability of a dielectric anomaly being present in relation to these attributes. These values are crucial for assessing the presence and severity of anomalies in the apple's dielectric properties.

Table 5.2: Ranked features for the dielectric anomaly prediction

Ranked Attributes	**Ranked Attributes**
0.9477	38.S5,5/rate
0.9363	20 S3,3 /rate
0.7971	11 S2,2 /rate
0.4374	65 S8,8 /rate
0.4157	29 S4,4 /rate
0.4071	47 S6,6 /rate
0.4004	56 S7,7 /rate
0.3339	59 S2,8 /rate
0.3339	17 S8,2 /rate
0.2736	45 S4,6 /rate
0.2736	31 $6,4 /rate
0.2665	62 S5,8 /rate
0.2665	41 S8,5 /rate
0.2657	53 S4,7 /rate
0.2622	32 S7,4 /rate
0.2471	35 S2,5 /rate
0.2471	14 $5,2 /rate
0.2408	2 S1,1 /rate
0.233	22 S5,3 /rate
0.233	36 S3,5 /rate
0.2235	51 $2,7 /rate
0.2232	16 S7,2 /rate

Table 5.3: Statistical analysis of proposed model for dielectric anomaly detection on Realtime training data

Correctly classified instances	99.6004%
Incorrectly classified instances	0.3996%
Kappa statistics	0.9713
Mean absolute error	0.006
Root mean square error	0.0629
Relative absolute error	4.2547%
Root relatively squared error	23.8332%

Table 5.3 provides a detailed breakdown of the statistical analysis conducted on our proposed model for detecting dielectric anomalies in real-time training data. Table 3 offers a detailed statistical analysis of a proposed model that has been developed for the purpose of detecting dielectric anomalies. This model has presumably been trained on real-time data.The statistical analysis may encompass various metrics and evaluations that assess the model's performance. It aims to provide a comprehensive understanding of how well the model performs in detecting dielectric anomalies.

Figure 5.1 explicates the plot of feature-based correlation using dielectric anomaly detection process. In this the first rows gives the features of outlier; second row gives the feature of cluster and from third row it gives the features of the sampled data. The red colour depicts the anomaly object features correlation with respect to rows parameters and columns parameters.

In this section explains that the model demonstration at a high accuracy of 99.6 %, indicating that the majority of instances are classified correctly. he Kappa statistic measures the agreement between the model's predictions and the actual classifications. A value of 0.9713 suggests a very high level of agreement. These metrics quantify the average magnitude of errors in the model's predictions. Low values indicate that the model's predictions are close to the actual values on average shown in Table 5.3.

Table 5.4: Detailed accuracy by class

	TP Rate	FP Rate	Precision	Recall	F-call	MCC	ROC	PRC	Class
	0.998	0.026	0.998	0.998	0.998	0.971	0.981	0.997	no
	0.974	0.002	0.974	0.974	0.974	0.971	0.981	0.930	yes
Weighted average	0.996	0.025	0.996	0.996	0.996	0.971	0.981	0.992	

Table 5.5: Confusion matrix

a	b	Classified as
1847	4	a=no
4	147	b=yes

In Table 5.4, explains various key parameters, including precision, recall, frequency measure, accuracy, error rate, and the weighted average of the sampled data. In Table 4, a detailed examination of various critical parameters related to the model's performance is presented.

Precision: Indicates the accuracy of positive predictions.

Recall: Measures the ability to correctly identify positive instances.

Frequency Measure: Represents a harmonic mean of precision and recall.

Accuracy: Reflects the overall correctness of the model's predictions.

Error Rate: Indicates the proportion of incorrect predictions.

Weighted Average: Averages the sampled data with weighted consideration.

Table 5.5 illustrates the confusion matrix generated as a result of the anomaly detection process. Table 5 provides a visualization in the form of a confusion matrix, which is a fundamental tool in evaluating the performance of a classification model. The confusion matrix likely breaks down the model's predictions into true positives, true negatives, false positives, and false negatives. It aids in understanding the model's ability to correctly classify anomalies and non-anomalies.

Figure 5.1 illustrates feature-based correlation within the context of the dielectric anomaly detection process. Let's break down the details of this plot:

1. Feature-Based Correlation: Feature-based correlation refers to the analysis of how different features or attributes within a dataset are related to each other. It's a fundamental step in understanding the relationships and dependencies among various aspects of the data.

2. Row Structure: The figure is organized into rows, with specific information presented in each row. The first row appears to represent the features of outliers. These are data points or instances that deviate significantly from the expected or normal behavior and are often considered anomalies.The second row is dedicated to the features of clusters. Clusters typically group data points with similar characteristics or properties together.

3. Columns Parameters: The columns likely represent various parameters or attributes within the dataset. These parameters could be related to dielectric properties, sensor readings, or other relevant factors.

4. Color Coding: The use of color, specifically the color red, is employed to visually depict the correlation between different features or attributes. The red color often signifies a strong positive correlation or association between variables.

5. Visualization: The figure visually represents how each feature, whether it belongs to outliers, clusters, or sampled data, correlates with the parameters or attributes represented by the columns.

By examining the color coding, one can discern the strength and direction of the correlation. For example, if a feature in the first row (outliers) is strongly red-colored about a specific column parameter, it indicates a notable correlation between outliers and that parameter. Overall, Figure 5.1 serves as a visual aid to help understand the interrelationships between different features and parameters in the context of dielectric anomaly detection. It provides insights into which attributes are more closely related to anomalies, clusters, or the sampled data, which is valuable for further analysis and decision-making in anomaly detection processes.

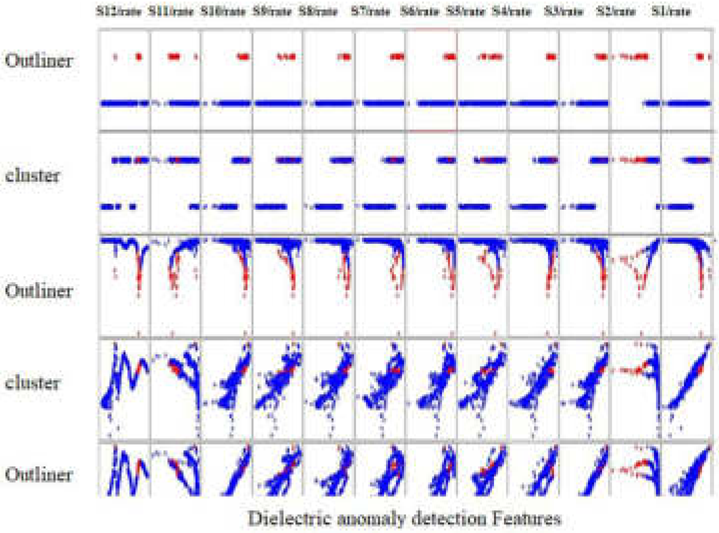


Fig. 5.1: Features-based correlation plot for the dielectric anomaly detection

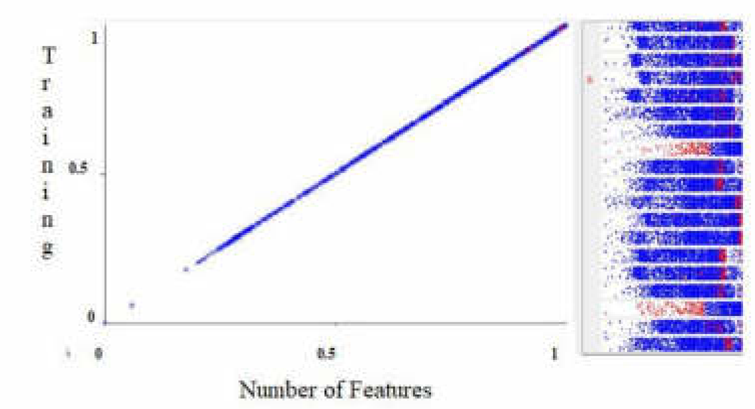


Fig. 5.2: Single feature-based correlation plot for the dielectric anomaly detection.

In Figure 5.2, presents a graphical representation of a single feature and its correlations within the context of our anomaly detection process. The red marks visible in Figure 5.2 represent potential anomaly objects, which have been identified as data points exhibiting unusual behavior or characteristics. These anomalies could signify deviations from expected patterns in the feature space. Conversely, the blue points in figure 5.2 correspond to non-anomaly objects, which are data points displaying normal behavior within the feature space. This visual depiction allows us to observe the distinct separation between potential anomalies (red) and typical data points (blue), demonstrating the effectiveness of our anomaly detection method in identifying aberrations within the single feature and its associated correlations.

In Figure 5.3, we present a comparative analysis of our proposed ensemble dielectric anomaly detection model alongside established models such as KNN, Naive Bayes, logistic classifier, and Random Forest classifier. This analysis encompasses various characteristic metrics including accuracy, precision, Frequency measure, recall, and error rate, all evaluated on the dielectric dataset. The proposed ensemble dielectric anomaly detection approach achieved an impressive accuracy of 98.9%, signifying its ability to correctly classify anomalies and

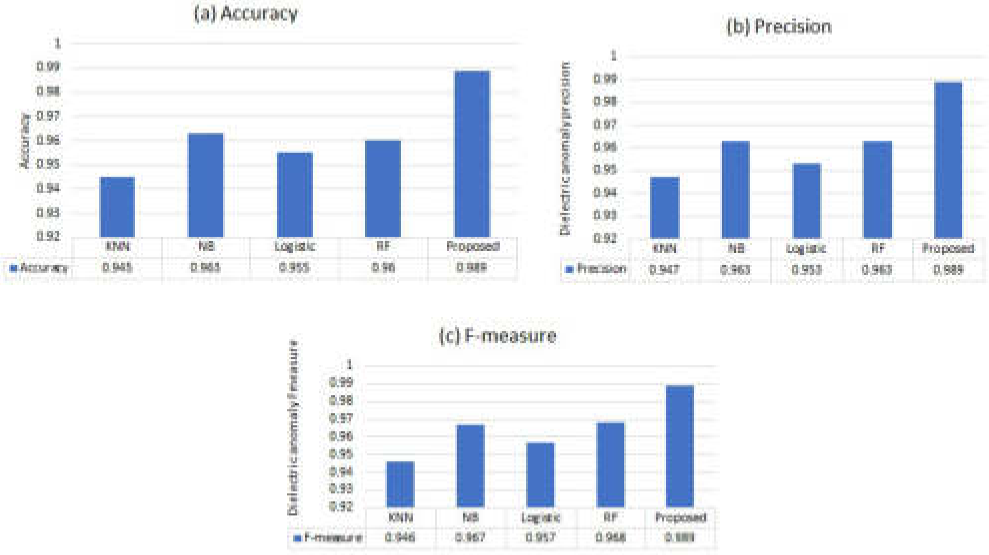


Fig. 5.3: Comparative analysis of proposed ensemble dielectric anomaly detection characteristics to the conventional anomaly detection classifier characteristics on the dielectric training dataset.

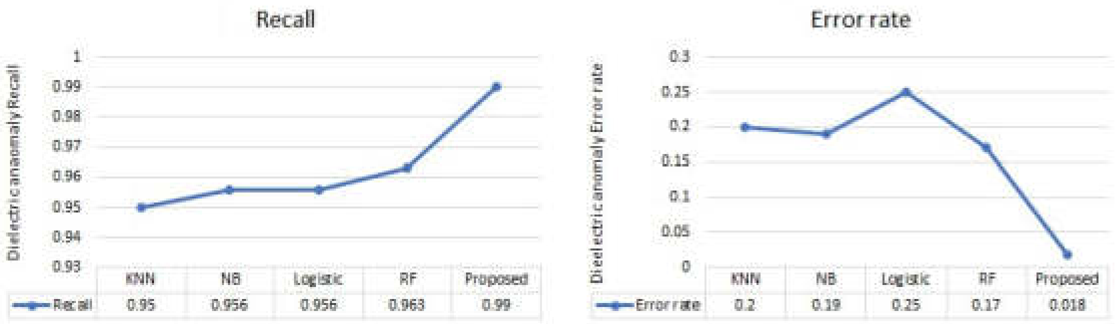


Fig. 5.4: Recall and Error-rate

normal instances within the dataset.

The precision value of 0.989 indicates a high proportion of correctly identified anomalies among all the anomalies detected, minimizing false positives which is shown in Figure 5.3.

The F-measure value of 0.989 demonstrates a balanced harmonic mean of precision and recall, emphasizing the model's effectiveness in handling both precision and recall simultaneously. Dielectric anomaly recall stands at 0.99, indicating a strong ability to capture a significant portion of actual anomalies, which is crucial for comprehensive anomaly detection. Lastly, the error rate, at 0.18, showcases the model's capability to minimize misclassifications and overall classification errors, further underlining its robustness in the context of dielectric anomaly detection. In summary, the proposed ensemble model outperforms traditional classifiers in terms of accuracy, precision, recall, and overall error rate, making it a promising choice for dielectric anomaly detection tasks shown in Figure 5.4.

In Figure 5.5(a, b), we present images of both a selected healthy apple and a defective apple, offering visual representations of these two distinct conditions. Figure 5.5(a) depicts a photograph of a selected healthy apple.

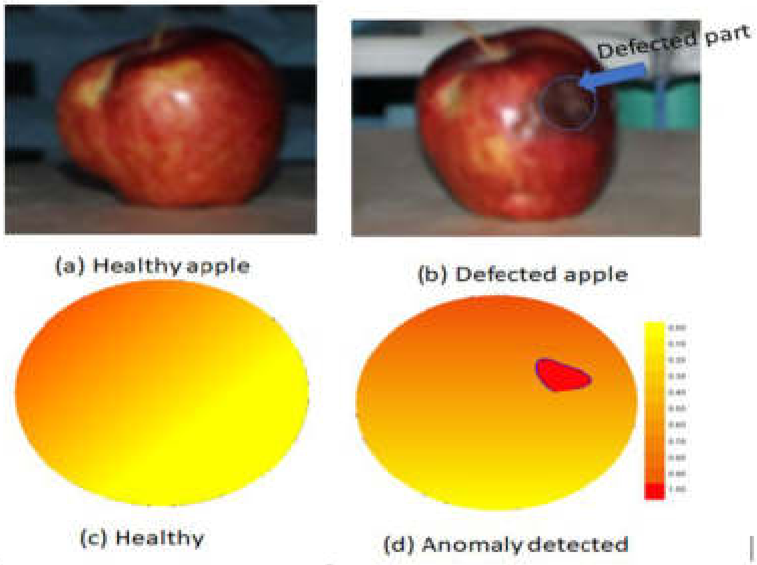


Fig. 5.5: Tomographic image reconstruction, (a) Healthy apple, (b) defected apple, (c, d) image reconstructions of healthy and defective apple.

This image provides a clear and detailed view of a well-formed, undamaged apple, showcasing its external appearance without any apparent defects or blemishes. In contrast, Figure 5.5(b) displays a photograph of a defective apple. In this image, we can observe visible imperfections or irregularities on the surface of the apple, indicating the presence of defects such as bruises, spots, or other forms of damage. Moving on to Figure 5.5 (c, d), we delve into the internal characteristics of these fruits through image reconstruction. Figure 5.5 (c) presents the image reconstruction of the healthy part of the fruit. This reconstruction allows us to visualize the internal structure of the healthy apple, revealing details of its interior composition without defects.In Figure 5.5(d), we focus on the image reconstruction of the defective part of the fruit. This isolated reconstruction provides a closer look at the specific region within the apple that exhibits abnormalities or defects. It enables a detailed examination of the internal damage or irregularities present in the defected portion. These images and reconstructions are instrumental in quality control and assessment processes, aiding in the identification and differentiation of healthy and defective apples, which is vital for ensuring product quality in the agricultural and food industries. This research utilizes sophisticated methodologies, incorporating a multi-level outlier detection strategy and a membership function, for the precise identification of dielectric anomalies. The application of an ensemble classification framework further improves the accuracy and robustness of the process for localizing anomalies. In summary, this study significantly contributes to the advancement of efficient techniques in detecting and pinpointing dielectric anomalies within intricate systems.

6. Conclusion. This study introduces a robust method for dielectric anomaly detection, particularly for assessing apple quality and ripeness using dielectric tomography. The developed tomography system, featuring advanced antennas within an anechoic chamber, enables precise data collection. The framework incorporates statistical analysis, density-based clustering, and ensemble learning to enhance anomaly detection accuracy.The research underscores the importance of accounting for the frequency-dependent characteristics of dielectric properties, which impact antenna radiation patterns and subsequently affect data collection and anomaly detection processes. This insight is critical for optimizing system performance across various frequency bands.Empirical results validate the effectiveness of our approach in accurately identifying dielectric anomalies in real-time sensor data. The ensemble classification framework enhances anomaly detection reliability, making it valuable for quality control, fault detection, and industrial applications.In summary, this research contributes significantly to dielectric tomography and anomaly detection, providing a validated framework. This innovation has the potential to revolutionize non-invasive quality assessment across industries, particularly for perishable goods like apples. In conclusion, our approach can transform quality control, reduce waste, and ensure product consistency, benefiting both consumers and industries.

REFERENCES

[1] G. Kaur,, *"A comparison of two hybrid ensemble techniques for network anomaly detection in spark distributed environment"*,Journal of Information Security and Applications, vol. 55, p. 102601, Dec. 2020, doi: 10.1016/j.jisa.2020.102601.

[2] H. Bhatt and M. Shah , *"A Convolutional Neural Network ensemble model for Pneumonia Detection using chest X-ray images"*, Healthcare Analytics, vol. 3, p. 100176, Nov. 2023, doi: 10.1016/j.health.2023.100176.

[3] J. Jiang et al.,, *"A dynamic ensemble algorithm for anomaly detection in IoT imbalanced data streams,"* Computer Communications, vol. 194, pp. 250–257, Oct. 2022, doi: 10.1016/j.comcom.2022.07.034.

[4] J. Fuentes-Velazquez, E. Beltran, E. Barocio, and C. Angeles-Camacho. , *"A fast automatic detection and classification of voltage magnitude anomalies in distribution network systems using PMU data,"* Measurement, vol. 192, p. 110816, Mar. 2022, doi: 10.1016/j.measurement.2022.110816.

[5] A. K. Dey, G. P. Gupta, and S. P. Sahu, , *"A metaheuristic-based ensemble feature selection framework for cyber threat detection in IoT-enabled networks,"* .Decision Analytics Journal, vol. 7, p. 100206, Jun. 2023, doi: 10.1016/j.dajour.2023.100206.

[6] I. Piekarz et al., , *"A microwave matrix sensor for multipoint label-free Escherichia coli detection,"* Biosensors and Bioelectronics, vol. 147, p. 111784, Jan. 2020, doi: 10.1016/j.bios.2019.111784.

[7] M. Vishwakarma and N. Kesswani,, *"A new two-phase intrusion detection system with Naïve Bayes machine learning for data classification and elliptic envelop method for anomaly detection,"* Decision Analytics Journal, vol. 7, p. 100233, Jun. 2023, doi: 10.1016/j.dajour.2023.100233.

[8] Y. Kayode Saheed, O. Harazeem Abdulganiyu, and T. Ait Tchakoucht, *"A novel hybrid ensemble learning for anomaly detection in industrial sensor networks and SCADA systems for smart city infrastructures,"* Journal of King Saud University - Computer and Information Sciences, vol. 35, no. 5, p. 101532, May 2023, doi: 10.1016/j.jksuci.2023.03.010.

[9] K. B. Sahay, B. Balachander, B. Jagadeesh, G. Anand Kumar, R. Kumar, and L. Rama Parvathy, , *"A real time crime scene intelligent video surveillance systems in violence detection framework using deep learning techniques,"* Computers and Electrical Engineering, vol. 103, p. 108319, Oct. 2022, doi: 10.1016/j.compeleceng.2022.108319.

[10] A. Muhammad and F. Külahci, , *"A semi-supervised total electron content anomaly detection method using LSTM-auto-encoder,"* Journal of Atmospheric and Solar-Terrestrial Physics, vol. 241, p. 105979, Dec. 2022, doi: 10.1016/j.jastp.2022.105979.

[11] E. Mushtaq, A. Zameer, and A. Khan, , *"A two-stage stacked ensemble intrusion detection system using five base classifiers and MLP with optimal feature selection,"* Microprocessors and Microsystems, vol. 94, p. 104660, Oct. 2022, doi: 10.1016/j.micpro.2022.104660.

[12]] J. Jiang et al., , *"AERF: Adaptive ensemble random fuzzy algorithm for anomaly detection in cloud computing,"* Computer Communications, vol. 200, pp. 86–94, Feb. 2023, doi: 10.1016/j.comcom.2023.01.004.

[13] L. A. Souto Arias, C. W. Oosterlee, and P. Cirillo, , *"AIDA: Analytic isolation and distance-based anomaly detection algorithm,"* Pattern Recognition, vol. 141, p. 109607, Sep. 2023, doi: 10.1016/j.patcog.2023.109607.

[14] O. Abu Alghanam, W. Almobaideen, M. Saadeh, and O. Adwan , *"An improved PIO feature selection algorithm for IoT network intrusion detection system based on ensemble learning,"* Expert Systems with Applications, vol. 213, p. 118745, Mar. 2023, doi: 10.1016/j.eswa.2022.118745.

[15] A. Copiaco et al., , *"An innovative deep anomaly detection of building energy consumption using energy time-series images,"* Engineering Applications of Artificial Intelligence, vol. 119, p. 105775, Mar. 2023, doi: 10.1016/j.engappai.2022.105775

[16] W. Khan and M. Haroon,, *"An unsupervised deep learning ensemble model for anomaly detection in static attributed social networks,"*International Journal of Cognitive Computing in Engineering, vol. 3, pp. 153–160, Jun. 2022, doi: 10.1016/j.ijcce.2022.08.002.

[17] Saikumar, K., Arulanantham, D., Rajalakshmi, R. et al, *"Design and Development of Surface Plasmon Polariton Resonance Four-Element Triple-Band Multi-Input Multioutput Systems for LTE/5G Applications"* Plasmonics, vol. 18, p. 1949–1958, 2023, doi: 10.1007/s11468-023-01922-w.

[18] Vasimalla, Y., Pradhan, H.S., Pandya, R.J. et al, *"Titanium Dioxide-2D Nanomaterial Based on the Surface Plasmon Resonance (SPR) Biosensor Performance Signature for Infected Red Cells Detection,"* Plasmonics , vol. 18, p. 1725–1734,2023, doi: 10.1007/s11468-023-01885-y.

[19] P. J. E. Peebles, *"Anomalies in physical cosmology,"* Annals of Physics, vol. 447, p. 169159, Dec. 2022, doi: 10.1016/j.aop.2022.169159.

[20] Revathi, R., Vatambeti, R., Saikumar, K., Alkhafaji, M. A., Khairy, U. R., & Noori, S., *" An advanced online mobile charge calculation using artificial intelligence,"* AIP Conference Proceedings, vol. 2845, p. 050026, Dec. 2023, doi: 10.1063/5.0170422.

[21] Vasimalla, Y. , Pradhan, H.S. , Pandya, R.J. , ... Rashed, A.N.Z. , Hossain, M.A., *"Anomalies in physical cosmology,"* Annals of Physics, vol. 447, p. 169159, Dec. 2022, doi: 10.1016/j.aop.2022.169159.

[22] S.-H. Son, K.-J. Lee, and W.-K. Park, , *"Application and analysis of direct sampling method in real-world microwave imaging,"* Applied Mathematics Letters, vol. 96, pp. 47–53, Oct. 2019, doi: 10.1016/j.aml.2019.04.016.

[23] W.-K. Park, , *“Application of MUSIC algorithm in real-world microwave imaging of unknown anomalies from scattering matrix,”* Mechanical Systems and Signal Processing, vol. 153, p. 107501, May 2021, doi: 10.1016/j.ymssp.2020.107501.

[24] G. Ludeno, I. Catapano, A. Renga, A. R. Vetrella, G. Fasano, and F. Soldovieri, , *“Assessment of a micro-UAV system for microwave tomography radar imaging,”* Remote Sensing of Environment, vol. 212, pp. 90–102, Jun. 2018, doi: 10.1016/j.rse.2018.04.040.

[25] M. S. Benmoussat, M. Guillaume, Y. Caulier, and K. Spinnler, , *“Automatic metal parts inspection: Use of thermographic images and anomaly detection algorithms,”* Infrared Physics & Technology, vol. 61, pp. 68–80, Nov. 2013, doi: 10.1016/j.infrared.2013.07.007.

[26] S. Liu et al., , *“CEUS Versus MRI in Evaluation of the Effect of Microwave Ablation of Breast Cancer,”* Ultrasound in Medicine & Biology, vol. 48, no. 4, pp. 617–625, Apr. 2022, doi: 10.1016/j.ultrasmedbio.2021.11.012.

[27] A. Chohra, P. Shirani, E. B. Karbab, and M. Debbabi, , *“Chameleon: Optimized feature selection using particle swarm optimization and ensemble methods for network anomaly detection,”* Computers & Security, vol. 117, p. 102684, Jun. 2022, doi: 10.1016/j.cose.2022.102684.

[28] J. Yu and J. Kang, *“Clustering ensemble-based novelty score for outlier detection,”* Engineering Applications of Artificial Intelligence, vol. 121, p. 106164, May 2023, doi: 10.1016/j.engappai.2023.106164.

[29] Maji, A.,Choubey, G, *“Improvement of heat transfer through fins: A brief review of recent developments,”* Heat Transfer, vol. 49, p. 1658-1685, Feb 2020, doi: 10.1002/htj.21684.

[30] Baskar, M., Ramkumar, J., Karthikeyan, C., Anbarasu, V., Balaji, A.,Arulananth, T. S, *“Low rate DDoS mitigation using real-time multi threshold traffic monitoring system,”* Journal of Ambient Intelligence and Humanized Computing, p. 1-9, Jan 2021, doi: 10.1007/s12652-020-02744-y.

Edited by: Polinpapilinho Katina
Special issue on: Scalable Dew Computing for Future Generation IoT Systems
Received: Oct 19, 2023
Accepted: Dec 20, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1702–1715, DOI 10.12694/scpe.v25i3.2785

SCHEMOS – SMART COW HEALTH MONITORING SYSTEM: AN IOT BASED COW HOOF DETECTION AND HEALTHCARE ALERT SYSTEM BY USING LSTM NETWORK

DURAIRAJ.K [*] DHILIP KUMAR.V [†] AND KANAGACHIDAMBARESAN.G.R [‡]

Abstract. Human life and existence are intertwined with a few domestic animals. One of the most important animals of this kind is the cow. Cows play a vital role in daily activities. Most of the people in India consume cow's milk as one of their major nutrients. Monitoring the health of a cow's everyday life is quite challenging. After infertility and mastitis, lameness is typically ranked as the third most economically significant health issue in dairy herds.Lameness are caused due to genetics, lack in nutrition i.e. a diet deficient in essential nutrients such as biotin, which can lead to hoof problems. Due to geographical environments like cows kept in wet, muddy conditions are more likely to develop hoof problems. This investigation analyses the typical characteristics of cow behavior, and a Smart Cow Health Monitoring System (ScHeMoS) using IoT is proposed to identify the cow's health through the data obtained from Internet of Things (IoT) sensors, including position, body temperature, stability, acceleration, and animal feed. IoT is combined with Deep learning (DL) technique to monitor and diagnose animal health. We used the Long Short Term Memory (LSTM) network to predict cow lameness by capturing the body temperature and other parameters, which will aid in predicting their illness. The accelerometer values are stored so that it will further help to determine which cow is lame and which is pregnant or regular and could be intimated to the care takers in the farms. We utilised a self collected dataset to perform the investigation. By implementing this system, we achieved 92.45% accuracy and 0.92 as F1 score.

Key words: Animal Behaviour, Cow Hoof Health, IoT in animal monitoring, LSTM, ScHeMo.

1. Introduction.

Cow's milk will be the first choice as one of best nutrient consumed by infant to older people in India or even in most of the countries .The demand for milk in India is raising daily. India's market for dairy products is anticipated to increase dramatically over the next few years due to an increase in consumers, rising incomes, and a growing interest in nutrition. Dairy products that have been pasteurised and packaged are becoming more popular in cities. Numerous national and international brands have joined the market due to increased competition from the private sector, raising consumer expectations for quality. However, a small percentage of people consume these packaged goods. Because of its flavour and perceived freshness, unpackaged, raw milk from a neighbourhood milkman is still preferred in many parts of the nation. [1] Dairy cow productivity is influenced by several factors, one of which is health. A sickness prevents dairy cows from producing milk as efficiently, which lowers milk yield. Dairy cows can have up to 12 to 15 litres of milk per day under normal circumstances, while diseased dairy cows can only generate 3 to 8 litres of milk per day. The incapacity to monitor the ranchers' shared understanding of the disease makes it challenging to identify and treat diseased cows in the early stages.

Lameness in cattle are caused due to genetics, where some cow breeds may be predisposed to hoof problems, lack in nutrition that is ,a diet deficient in essential nutrients, such as biotin, can lead to hoof problems, and due to geographical environments like cows kept in wet, muddy conditions are more likely to develop hoof problems. Fig. 1.1 (a) shows how a Lameness cow walks in a particular motion; (b) the kind of injury a cow would have at the bottom of its hoof, which is something the untrained eyes cannot see,(c) the toes of the cow are crossed because of bilateral damage, which must be cut . A major animal welfare issue in dairy cows is lameness, which causes intense pain and strain, which enervates and decreases milk productivity. For example, in the dairy form, less than 10% of cows are affected by lameness. Lower fertility rate are found among cows

[*] Department of Computer Science and Engineering, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Avadi, Chennai 62, India. (duraiit2011@gmail.com)

[†] Department of Computer Science and Engineering, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Avadi, Chennai 62, India. (vdhilipkumar@veltech.edu.in)

[‡] Department of Computer Science and Engineering, Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Avadi, Chennai 62, India. (kanagachidambaresan@gmail.com)

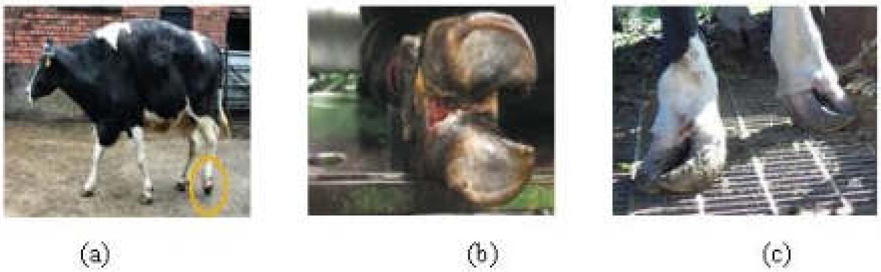


Fig. 1.1: Different forms of Hoof problems (a) Lameness (b) Injured cow hoof root (c) Bilateral crossed toes.

affected with lameness. Foot rot is an infection that causes sudden swelling, heat and inflammation in the foot, resulting in severe lameness [2]. When evaluating structural soundness, one of the critical features to assess is the hoof. Some hoof problems, such as excessive or uneven toe growth, may be caused by hereditary, dietary, or environmental factors, or they may be symptoms of other health issues the animal is experiencing. The ideal hoof will be free of cracks and flaws, with two symmetrical claws facing forward. The heel depth should also be carefully monitored because animals with an excessive tilt to their hocks and pasterns might be exceedingly shallow in the heel. The hoof should be dense and capable of supporting the animal's weight without shattering, as this might cause problems [3].

A corkscrew claw, or screw claw, is another symptom to look for while evaluating the hoof. This toe twisting puts the hoof's side wall in direct touch with the ground. At times, the disease manifests itself with the toes pointing inward rather than outward. The typical symptoms of this illness are present in cattle older than two years old. It can damage all hooves or only one of them [4]. Although the mode of inheritance is not fully known, this condition is thought to be heritable. Due to improper weight distribution inside the toe, the issue can cause lameness. Cattle suffering from this disorder must be discarded and eradicated from the herd immediately.

Animals cannot communicate their health issues to human beings. Hence, this research suggests the prototype of a wearable device for animals with the help of a Smart Cow Health Monitoring System (SCHeMo), which shall be mounted on the animals that produce the alert message related to their health issues to both the forest officer's room and the veterinary doctors [5]. The primary parameter is to monitor the animals regularly. It indicates whether they are suffering from diseases or in dangerous situations caused by natural disasters like floods or wildfires. Therefore, this prototype will be user-friendly for evaluating animals' behavioural monitoring.

We propose SCHeMoS model which can serve as health alert system in cow lameness detection. The research brings out:

- From observing to recognition and healing of the diseased cows, our method continuously monitors and manages using IoT and LSTM model.
- A cutting-edge data analytical method aid in the precise detection of the animal behaviour.
- IoT sensors and actuators assist in detecting animal behaviour for early disease analysis and diagnosis.
- We use a rapid prototype model to monitor the animal and maintain a healthy habitat. To recognirmal cows, SCHeMoS can be used by the farmers in identifying the lame cows
- Our proposed model outperformed the existing approaches in detecting the lameness of cows with the accuracy of 92.45%.

The remainder of this article is structured as follows: Section 2 covers the literature review. Section 3 provides specifics on the proposed approach, while Sections 4 contains the implementation and 5 outline the results that were attained. Section 6 concludes the investigation.

2. Related work. There have been numerous research projects on domestic animal surveillance recently. Cow health monitoring is crucial in today's society since it allows for the prediction of milk production. It is essential because milk production is the primary income source for farmers with many cows. A health observing system that focuses on several options are available. The issue with such systems is that they need help forecasting milk production, which is crucial in determining a cow's health. Video-based monitoring with ResNet 3D and other DL approaches can recognise cow behaviours as resting, walking, and roaming [6]. Another untrained ML model [7] was used to assess cow movement patterns and identify anomalies. It used collar, ankle, or neck accelerometers.

Authors of [8] used gait analysis with accelerometers has been done on both people and animals .It has been established that while accelerometer-based gait metrics are still comparatively immature, latest field in the dairy industry includes the use of sensors that can be worn by cows, such as accelerometers. Sheep behaviour and lameness categorization have both been studied using accelerometers and gyroscopes [9].

Use leg-mounted accelerometers [10] to identify cow lameness. Neck, foot, and throat tri-axial accelerometers are used to predict lameness in sheep and have an overall accuracy of more than 85 percent. Wandering, resting, eating, and sleeping were among the monitored actions. In [11] acceleration signal analysis has been shown to have a reported accuracy of 91.9% when applied to the diagnosis of lameness in cows. Two 400-Hz accelerometers were utilised by the authors of [12] to evaluate bilateral front limb impairment and foot disorders by extracting the full rotation, standing phase, and range of motion. In dairy cows with hoof lesions, the connection between gait features and movement score has been studied.

Reviews the use of accelerometer [13] in various clinical applications, including the assessment of gait disorders, fall detection, and the monitoring of patients with neurological conditions. The authors also discuss the advantages and limitations of accelerometer-based gait analysis, including its non-invasive nature and its ability to provide objective measurements. They also highlight the importance of the development of appropriate algorithms and methods for the analysis of accelerometer data. Finally, the authors suggest future research directions to improve the accuracy and reliability of accelerometer-based gait analysis.

Examines accelerometers [14] which are devices that measure the acceleration of a cow's movement, for the categorization of cattle movement and activities in the dairy. The study used data from accelerometers placed on cows in a commercial dairy farm to classify the cows' behaviours, such as lying down, standing, and walking. The authors used ML algorithms, such as RF and SVM, to classify the behaviours and found that the accelerometer data was able to accurately classify the behaviours with a high accuracy. The authors also discuss the potential of using this technology for cow lameness detection and for monitoring cow welfare in dairy barns.

Authors of [15] explores the use of a combination of locating and accelerating the sensors to detect the differences among cows while feeding that are affected with lameness. The study used data from sensors placed on cows to track the cows' movements and feeding behaviours, and analyzed the data using machine learning algorithms. The authors found that the sensor data accurately distinguishes healthy and unhealthy cows based on their feeding behaviours. The authors also covered the possibility of employing this technology to identify lameness among cattles and the significance of taking feeding behaviour into account as a sign of lameness.

Electroencephalogram (EEG) recordings [16] of patients with a range of poor sleep, including restlessness, snoring, and nerve pain, were employed in the study to collect data. and applied deep learning algorithms to classify the patients into different groups based on their disorder. The authors found that the deep learning algorithms were able to accurately classify the patients into different groups with high accuracy. The authors also discussed the potential of using this technology for the diagnosis and treatment of sleep disorders, and the importance of considering multiple bio signals in the analysis. Few limitations of the literature review are presented in Table 2.1. The goal of this investigation is to create a recurrent neural LSTM model to properly and completely categorise cow behavioural traits, particularly those connected to lameness.

3. Proposed Work. This part contains detailed description of the proposed approach, dataset preparation, data pre-processing and methods and materials used to complete the investigation. The proposed architecture is depicted in Fig. 3.1. Initially the data is pre-processed after acquiring it. The cleaned data is given as input to the LSTM model for classifying lameness. The main advantages of LSTM is that they are much better at managing long-term dependencies. This is due to their capability to remember data for prolonged periods of time. Second, LSTMs are much less vulnerable to the vanishing gradient issues. This also give optimal predictions during machine learning when compared to existing algorithms.

Table 2.1: Overall results of the proposed approach

Work	Methodology	Parameters used	Results	Limitations
Casey et. al [17]	MAPE (Mean Absolute Percentage Error) computation method.	Body Temperature Detection, Heart Rate Detection Minutes	The average MAPE value to identify the body temperature and metric for heart beat rate is 1.254, 2.434 respectively.	Small sample size of cows used in the study, which limits the generalizability of the results to the wider population of dairy cows
Lamb et al [18]	TO1 to TO5 is used to synchronise communication protocols.	Breathing frequency sensing rnit, Monitoring the activity of cow.	Monitoring the cow's behaviour, health, and stress.Accuracy rate is measured as 87.61%	The study only used data from accelerometers placed on the front legs of the cows, which may not provide complete picture of the cows gait and lameness status.
Thorup et al [19]	The milk yield prediction model was created using the MATLAB with ThingSpeak.	Temperature & Humidity	Monitor cow's health, and vets, that quickly identifies and treat minor health issues with accuracy of 85.01%	The study only used data from accelerometers placed on the frontlegs of the cows, which may not provide a complete picture of the cows gait which affects the accuracy.
Haug et al [20]	Used IoT techniques. The MQTT protocol is used for data communication within gateway and nodes. HTTP is used for data transmission within server and gateway	Mastitis, Bloat, PMK, Anthrax, Brucellosis, Leptospirosis, Myiasis, Scabies	IoT and smart systems are integrated in monitoring the cattles.	The study used a simple algorithm for lameness detection and did not use advanced machine learning techniques to improve the accuracy of the lameness detection, which could be a limitation in terms of the performance of the system. IoT systems used MQTT which has higher computational complexities.

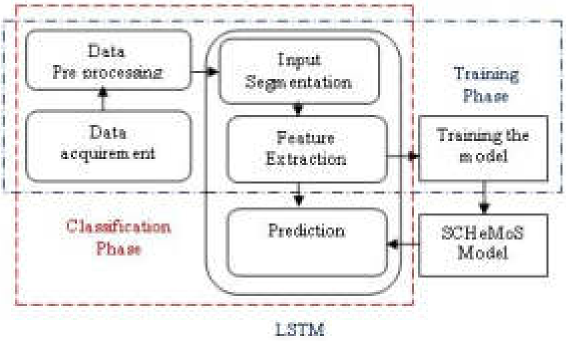


Fig. 3.1: Proposed architecture

3.1. Data pre-processing. To prepare a dataset for cow lameness prediction, we gathered data on cow behavior and physical characteristics, as well as information on their environment and any potential causes of lameness. This may include data on the cow's movement patterns, gait, and posture, as well as information on the flooring and bedding in their living area. Additionally, data on any injuries or conditions that may contribute to lameness specifically hoof infections or joint issues are collected. It is important to have a large

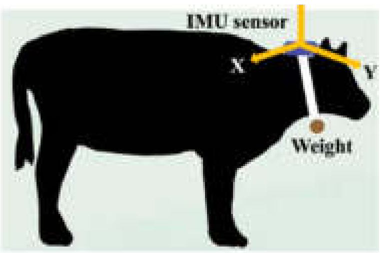


Fig. 3.2: Collar Sensor placed on the cow to collect data

Table 3.1: Category wise cows used for the investigation

Cow breed	Gender	No. of cows used
Angus	Male	12
Angus	Female	11
Brahman	Male	10
Brahman	Female	8
Crossbred	Male	12
Crossbred	Female	14
Charolais	Male	11
Charolais	Female	9
Total		**87**

and diverse dataset, including both healthy and lame cows, to train our model effectively. The data sets include information on the following types of cattle: Angus, Crossbreds and others. Table 3.1 shows statistical information like breeds, gender and number of cows used to collect the data. The data collection was carried out in a farm for 35 days where all the above said breeds are available.

Around four cow breeds are used to prepare the sensor's raw data set, representing different species and ages. Cattle state is divided into eight subcategories, including relaxing, chewing, highly active, moderately active, gasping (heavy breathing), eating, and roaming. The detailed explanation of various conditions of the cow is shown in Table 3.1. The collected dataset contains thorough information for several cattle statuses. Sensors deployed in the neck of the cow as shown in Fig. 3.2 to collect various information Each sensor records the state of the cows for every 30 seconds, and 12,892 data points were gathered for each cow during the time 16:30 IST on December 12,2022.A sample of the collected dataset is shown in Table 3.2.

3.2. Data segmentation. Segmentation is the initial step in the data processing process. Cows belonging to the same breed are grouped together in the statistics. We utilised R programming language to segment the data since the source data were extremely large. The number and breeds of cows used to prepare the dataset is shown in Table 3.1. A large volume of data makes it easier to analyse general characteristics of data and prevent miscalculation brought on by specific and individual data.

We collected data based on various positions of the cow. Table 3.2 shows a few of the cow positions used to collect data for our experiment. The collected data is segmented as per the same breed with a difference in gender. A sample of sensor data from a walking cow of the Brahman breed is shown in Table 3.3. Px, Py, and Pz represent the cow's movement directions. These values are compared to determine the cow's weaning style. To accurately classify hoof foot, the cow's body temperature is also recorded because a change in temperature would also cause a change in the cow's walking style. We collected data and segmented it for all the actions listed in Table 3.2. for all four breeds of cows.

3.3. Data Cleaning. Data cleaning in a cow disease dataset involves a series of steps to ensure that the data is accurate, consistent, and usable for analysis. Removing duplicate records involves identifying and

Table 3.2: Various Positions of Cow used for data collection

Cow ID	Position	Explanation
NC4231	Ideal	The normal cow's position was ideal.
NC4232	Grazing	The average cow here was in a grazing position.
NC4233	Drinking	The normal cow is drinking water.
NC4234	Walking	The normal cow was walking around.
PC6241	Lying Down	The pregnant cow was lying down on the field.
PC6242	Walking	The pregnant cow was walking around.
PC6243	Sleeping	The pregnant cow was sleeping.
LC3351	Ideal	The Lame cow was in an ideal position.
LC3352	Walking	The Lame cow was walking around.
LC3353	Grazing	The Lame cow was in a grazing position.

removing any records that are identical or nearly identical. Handling missing data, involves identifying and addressing missing data, such as by removing records with missing values, imputing missing values, or flagging records with missing data for further investigation. For instance, the relaxed condition of the cow will be marked as zero for specific hour if the cow shows no change in the position. If lot of damage data is collected in that hour; it will affect how the average time stamp is determined. Removing irrelevant data involves identifying and removing data that is not relevant to the analysis. For example Deleting faulty data along with the accompanying time serial number will prevent them from being factored into the average period computation.

In our investigation the sensor delivers very less amount of inaccurate data along with its detection and transmission of the cow's state. Even this less inaccurate data would impact the classification accuracy of the model. Consequently, cleaning up the faulty data is the first step. We used the above methods to clean up the data.

3.4. Prediction using LSTM. Based on the research discussed above, we use the LSTM model in this part to predict the state of a cow lameness whether it is affected with hoof foot.. To be clearer, it is first detailed how to build an LSTM model as well as the characteristics and structure of LSTM. Second, the cow condition is anticipated and simulated using the LSTM model. The model is finally optimised to raise its accuracy. The process of using LSTM to predict lameness in cows would involve the following steps:

Collect data: Collect data related to the cow's behavior, sensor readings, and other relevant factors that may indicate lameness.

Pre-processing: Pre-process the data by cleaning and normalizing it, as well as converting it into a format that can be used by the LSTM model.

Build the LSTM model: Use the pre-processed data to train and build the LSTM model using a suitable deep learning library such as TensorFlow or Keras. Evaluation: Evaluate the model's performance using appropriate metrics such as accuracy, precision, and recall.

Fine-tuning: Fine-tune the model as necessary by adjusting the model parameters, adding additional features, or trying different architectures.

Deployment: Once the model is trained and fine-tuned, it can be deployed for use in predicting the likelihood of lameness in cows.

LSTM (Long Short-Term Memory) is a type of Recurrent Neural Network (RNN) that is particularly well-suited for modelling time-series data and sequences. It can be used to predict the likelihood of lameness in cows by analyzing data from sensor readings, cow behavior, and other related factors. A simple LSTM cell is as shown in Fig. 3.3. Equ. 3.1 to to 3.5 are used by the LSTM network to process the input. The sensor data is given as input through the input gate it ,predicted output is obtained from the output gate o_t and the forget gate f_t is used to analyse the time series data with the help of the memory cell c_t.

$$i_t = \text{sigmoid}\left(W_{ix} \times x_t + W_{ih} \times h\{t - l\} + b_i\right) \tag{3.1}$$

Table 3.3: A sample of male Brahman Cow's walking record

Date	Time	Cow Body Temperature	Px	Py	Px
08 – 12 – 2022	16 : 31 : 02	37.06	2731	7146	9078
08 – 12 – 2022	16 : 31 : 04	37.13	2732	7606	9035
08 – 12 – 2022	16 : 31 : 06	37.19	2730	7466	8872
08 – 12 – 2022	16 : 31 : 07	37.31	2729	7603	8943
08 – 12 – 2022	16 : 31 : 09	37.44	2731	7645	9022
08 – 12 – 2022	16 : 31 : 10	37.5	2732	7630	9021
08 – 12 – 2022	16 : 31 : 12	37.5	2730	7531	8815
08 – 12 – 2022	16 : 31 : 13	37.63	2728	7690	8903
08 – 12 – 2022	16 : 31 : 15	37.75	2729	7738	9062
08 – 12 – 2022	16 : 31 : 17	37.81	2727	7716	9067
08 – 12 – 2022	16 : 31 : 20	37.94	2737	7763	9249
08 – 12 – 2022	16 : 31 : 21	38.06	2731	7591	9140
08 – 12 – 2022	16 : 32 : 14	38.44	2730	7875	9248
08 – 12 – 2022	16 : 32 : 16	38.44	2734	7854	9251
08 – 12 – 2022	16 : 32 : 20	38.5	2734	7913	9292
08 – 12 – 2022	16 : 32 : 22	38.5	2738	7897	9258
08 – 12 – 2022	16 : 32 : 24	38.56	2741	7922	9297
08 – 12 – 2022	16 : 32 : 30	38.63	2739	7811	9226
08 – 12 – 2022	16 : 32 : 59	38.56	2737	7772	9409
08 – 12 – 2022	16 : 33 : 45	38.69	2737	7746	9056
08 – 12 – 2022	16 : 33 : 46	38.75	2737	7779	9044
08 – 12 – 2022	16 : 33 : 49	38.81	2737	7829	9126
08 – 12 – 2022	16 : 33 : 51	38.88	2732	7820	9142
08 – 12 – 2022	16 : 33 : 52	38.88	2734	7829	9183
08 – 12 – 2022	16 : 33 : 54	38.94	2731	7821	9164
08 – 12 – 2022	16 : 33 : 56	39	2733	7831	9190
08 – 12 – 2022	16 : 33 : 57	39	2733	7841	9213
08 – 12 – 2022	16 : 33 : 59	39.06	2733	7837	9241
08 – 12 – 2022	16 : 34 : 00	39.13	2729	7824	9234

$$f_t = \mathrm{sigmoid}\left(W_{f_x} \times x_t + W_{fh} \times h(t-l\} + b_t\right) \tag{3.2}$$

$$o_t = \mathrm{sigmoid}\left(W_{ox} \times x_t + W_{oh} \times h\{t-l\} + b_o\right) \tag{3.3}$$

$$c_t = f_t \times c\{t-l\} + i_t \times \tanh\left(W_{cx} \times x_t + W_{ch} \times h\{t-l\} + b_d\right) \tag{3.4}$$

$$h_t = o_t \times \tanh\left(c_t\right) \tag{3.5}$$

where x_t, h_t, c_t, are the input, hidden state, memory cell for the time stamp t respectively for i_t (input gate), f_t (forget gate), and o_t (output gate) respectively, The weight matrices for the input, output and forget gate are W_i, W_f, W_o and bias terms are b_i, b_o, b_f respectively. We used sigmoid as the sigmoid function; *tanh* is the hyperbolic tangent function [21].

Before applying the created LSTM model for cow status prediction, it is essential to find out the inputs, form of output, time series data. According to the properties of the data sets, predicted output should be the hoof foot affected status of the cow [22]. Hence, one difficult part of this approach is figuring out the input variables. A known fixed periodic function should be the input since periodic changes will be the output. To

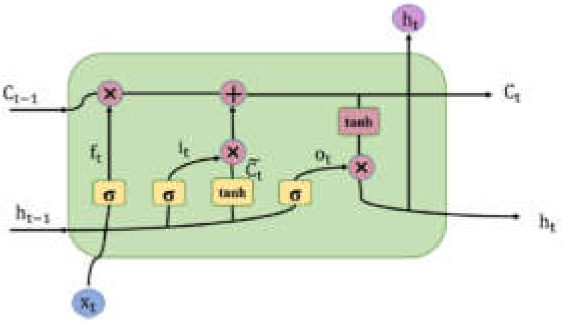


Fig. 3.3: LSTM cell structure

be more precise, it is appropriate to choose total hours as the input variable because the state cycle of cow is one day.

The LSTM model's input and output are periodic. The difference is that while each cycle's output has a different value, this cycle's input has a fixed value[23]. In other words, regardless of time, the same input may produce many results. The matching time series for the input is not the same even though the input is the same [24]. As a result, the LSTM model is capable of handling situations where a single input corresponds to a number of outputs in a time series.

3.5. Feature extraction. In this study, two features RMS (Root Mean Square) and mean were used to extract the features of the sensor data. Equ. 3.6 and Equ. 3.7 are used to perform the feature extraction by the LSTM model.

$$\mathrm{RpX} = \sqrt{\frac{1}{n}\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{PXi}} \tag{3.6}$$

$$\mathrm{m}\left(\mathrm{PX_j}\right) = \frac{1}{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{PX_i} \tag{3.7}$$

where PX is PX-axis data, PX_j is the record j of PX, n are the samples number and fixed as n = 32; PX_i is the ith sample of record PX_j; $m(PX_j)$: mean of PX_j, RPX_j : root means square of PX_j.Hence the formulae for $P_x, P_y, andP_z$ axis are similar.

3.6. Hyperparameter optimisation. To further improve the performance of the proposed model we performed hyper parameter optimisation. The number of hidden layers of the LSTM model is selected to be 7,we started with the number of neurons from 32 and gradually increased to 64.The dropout rate is randomly selected from 0.2 and we obtained most optimised accuracy of 92.4% at the rate of 0.1. The learning rate of the model is selected to be 0.001 with the batch size 0f 128.We achieved highest accuracy at the 120th epoch. The results obtained through hyper parameter optimisation are as shown in Table 3.4.

4. Implementation. This article describes a monitoring system that records a cow's heart rate, rumination rate, relative humidity, and body temperature at regular intervals to predict lameness. The parameters collected will then be sent via NodeMCU to a website called Thing Speak for processing and health evaluation. This study focuses on locating cow hoof wounds and keeping tabs on the well-being and behaviour of the subject. Fig. 4.1 depicts the basic layout of the hardware representation of the proposed system. A 4.7K ohm 1/2 Watt Resistor, AD converter (ADS1115), Temperature Sensor (DS18B20), Accelerometer (ADXL335), NodeMCU (ESP8266), 5V Battery, and Micro SD Card Reader Module are used to build the system. Another method for developing NodeMCU using a well-known IDE, the Arduino IDE. We may also use the Arduino programming

Table 3.4: Hyper Parameter Optimisation.

Parameters	Optimised values
No. of LSTM layers	7 Layers
Neurons	64
Learning rate	0.001
Epoch	120
Batch size	128
Dropout	0.1
Trainingloss	0.4357
Testingloss	4.9821
Error rate	0.167

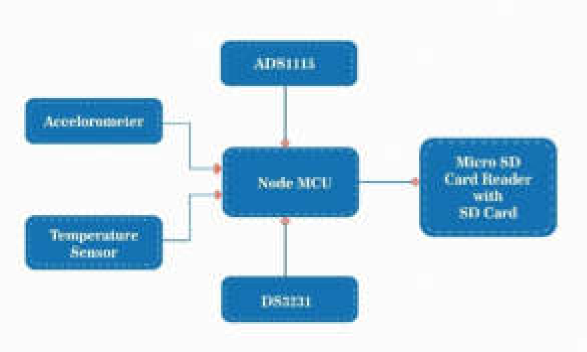


Fig. 4.1: Hardware architecture of SCHeMoS

environment to create applications for NodeMCU. The ADXL335 is used to find the accurate position of the cow. Based on the cow movement, various positions like $P_x, P_y and, P_z$ have been calculated.

4.1. PCB Design of SCHeMoS.

4.2. Product kit (Prototype). Fig. 4.3 (a) shows how all sensors are mounted and fixed by a double-side PCB board, and this process is a preparation for Fig. 4.3 (b), which shows the soldering of all the particulars with a jumper wire. Then, Fig. 4.3 (c) shows how everything is contained in a box to hold the product using a 3D printer. Finally, Fig. 4.3(d) shows the complete design after everything is collected and appropriately adjusted.

Fig. 4.4 (a) shows the ideal position of the cow after applying the device around the neck. Usually, a cow would shake its head in case of finding a burden that affects its liberty of motion, but luckily the device would not cause such a disturbance. As the cow drinks from the bucket, as in Fig. 4.4 (b) 12, it is easy for the cow to bend over and drink without any burden on the device. In Fig. 4.4 (c), the cow can be dragged easily for a walk without harming the device's purpose or the cow's neck. According to the survey, cattle constantly graze for 10- 15 minutes and drink at an average rate for 2-3 minutes. The above three positions have been tested in all three categories of cows. A Timestamp will be generated each minute, and this information will be saved on the MO20's SD card. The comparison after applying the test to the three different categories of cows resulted in various machine-learning algorithms. The system above SCHeMoS will assist the owner in taking preventive and curative steps as soon as possible, preventing catastrophic losses and the owner can take rapid action.

5. Results and discussion. From the results obtained it sounds like the LSTM model used for cow lameness detection is performing well in general, with the exception of the predictions for Brahman males. This could be due to the factors mentioned, such as their continuous rest state and less movement from one

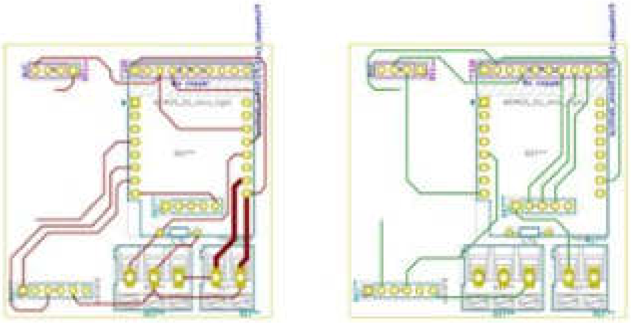


Fig. 4.2: Final PCB board circuit of SCHeMoS using CAD software

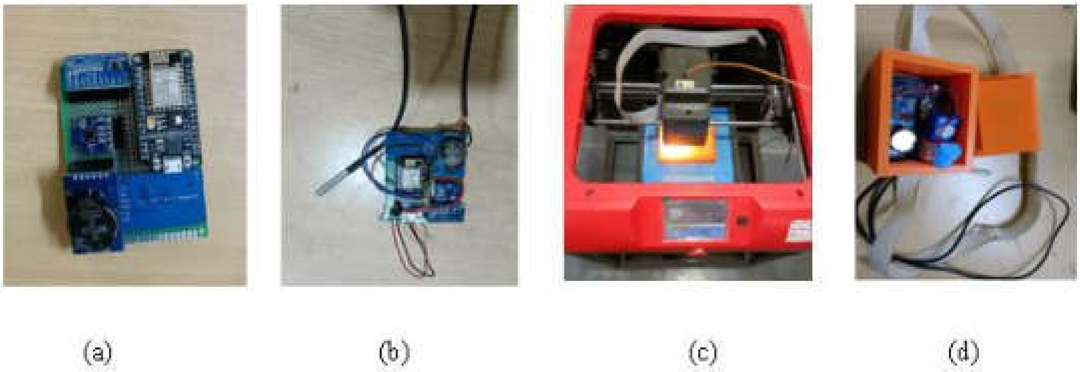


Fig. 4.3: SCHeMoS Hardware design for implementation(a) Sensors with double sided PCB board (b) Circuit of SCHeMoS (c) Preparing box for SCHeMoS in 3D printer (d) Design of SCHeMoS

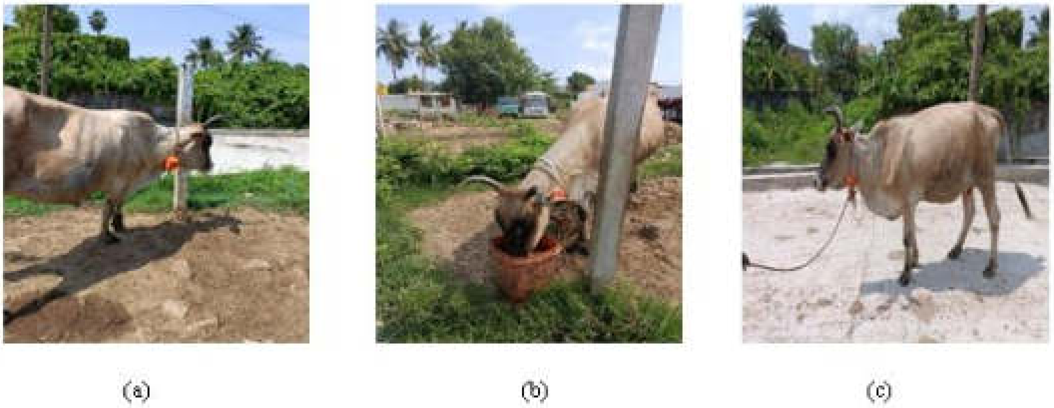


Fig. 4.4: Position of the cow after fixing the device (a) Ideal position of the cow (b) Drinking position (c) Walking position

Table 5.1: Performance of the Proposed System

Model Classification	Accuracy	F1 Score	Specificity	Recall	Precision
Affected with lameness	92.45	0.926	0.931	0.925	0.918
Not Affected with lameness	92.36	0.919	0.92	0.919	0.921

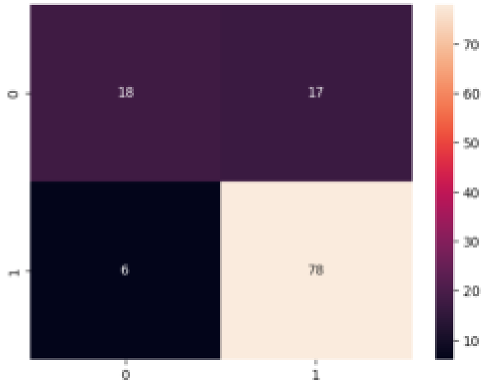


Fig. 5.1: Confusion matrix

place to another. It's important to note that the accuracy of DL models depends on the quality and size of the data used to train them. In this case, the proposed LSTM model is able to precisely calculate the active movement of the next cow, which is a good indication that it is a good fit for predicting cow lameness task. Additionally, the proposed LSTM model could be used for time-series prediction and it is important to pre-process the data accordingly, like normalizing the data, adding time-lag features, etc to obtain more accuracy. We assessed our obtained results by using few of the below ML metrics from Equ. 5.1 to 5.4. tr__pst, are true positive (actually lame and lameness correctly predicted), tr__ngt are true negative (actually not lame predicted as not lame), fp__pst are false positive (actually lame but predicted as not lame),fp__ngt are false negative (actually not lame but predicted as lame). Table 5.1 shows the overall results obtained from the proposed approach.

$$\text{Accuracy} \quad = (\text{tr}\_pst + tr\_ngt) \,/\, (\text{tr}\_pst + tr\_ngt + fs\_pst + fs\_ngt) \tag{5.1}$$

$$\text{Recall} \quad = \text{tr}\_pst / \,(\text{tr}\_pst + fs\_ngt) \tag{5.2}$$

$$\text{Precision} \quad = \text{tr\_pst} / \,(\text{tr\_pst} + \text{fs\_pst}) \tag{5.3}$$

$$\text{F1 Score} \quad = 2 \times (\text{ precision } \times \text{ recall })/(\text{ precision } + \text{ recall }) \tag{5.4}$$

To visualise the results obtained more precisely we represented the classified data in the form of confusion matrix. It summarizes the performance of a classification algorithm. Fig. 5.1 represents the confusion matrix for our approach. A typical confusion matrix for cow lameness detection would have the following structure: We compared our results with few of the existing approaches and the comparison revealed that the proposed approach have outperformed than the existing approaches. The comparison results are shown in Table 5.2. The training and testing accuracy of our SCHeMoS approach is shown in Fig. 5.2. It is observed that the testing accuracy gradually increases when the input data is increased. The training loss of the proposed model reduces to 0.2016 after the 120th epoch. The model stabilises at the learning rate of 0.001.

Table 5.2: Existing versus proposed approach.

Cow Category	SVM	Logistic Regression	Random Forest	K-NN	Proposed SCHeMoS Model
Lameness not affected	79%	83%	86%	90%	92.36%
Lameness affected	80%	85%	90%	90%	92.45%

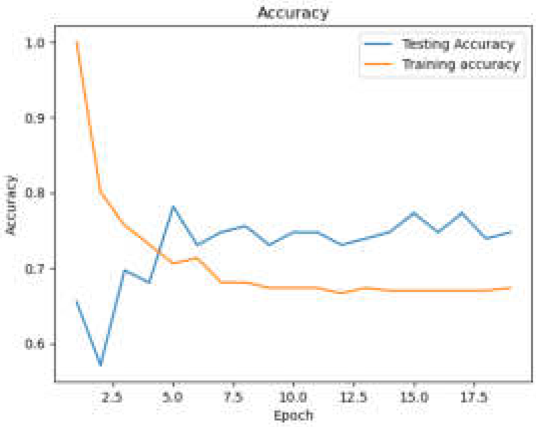


Fig. 5.2: Training and testing accuracy

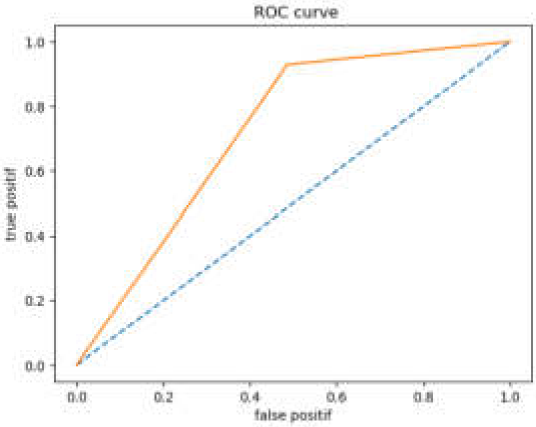


Fig. 5.3: ROC curve for the proposed classifier

A binary classifier's performance is depicted graphically by a Receiver Operating Characteristic (ROC) curve. The true positive rate (tr__pst), often referred to as recall or sensitivity, is represented on the y-axis while the false positive rate (fs__pst) is represented on the x-axis. Sequential and time series data are both well suited for our LSTM-based SCHeMoS model. The time-series data from the cow lameness data set can be analysed and classified using it. The curve line increases as the number of input is increased and stays constant at the 120th epoch with highest accuracy of 92.45%.The ROC of the proposed approach is shown in Fig. 5.3.

6. Conclusion. This research has suggested cow hoof health (lameness) monitoring using the sensor data obtained from the cow's position. We used LSTM network to classify hoof affected or not affected based on the input sensor data of cows performing actions like move, sit,walk,graze and drink. The farmer/diary maintenance person is alerted through IoT devices when the possibility surpasses a specified threshold. As per our result, the SCHeMoS model reduces the computational complexity by classifying in reduced time than few of the existing approaches and also reduced memory storage. In this work, the main goal is to build a cow monitoring model made up of IoT devices on a farm, which is used to collect data of cows in various positions, temperature, grazing habit etc.This collected data is cleaned and given as input to a LSTM network to predict the lameness in cows. After the noise is taken out of the data, a DL based LSTM model for cow lameness detection is built. The model is capable of predicting how the cow's position will transform over the next phase. When the model's predictions are compared to what actually obtained with few of the existing approaches, it shows how accurate and useful it is.The error rate of the model is recorded as 1.201 with the training loss of 0.2016, validation loss as 0.316 and overall accuracy of 92.45%.This model also has its limits. It needs a bunch of input data to learn, and if small amount of data is given as input, the model's prediction accuracy may reduce or sometime may be inaccurate.

REFERENCES

[1] Chao,Ludovico,Korkut,Tokgoz,Jim,Sihan,Hiroyuki, *Integrated Data Augmentation for Accelerometer Time Series in Behavior Recognition: Roles of Sampling, Balancing, and Fourier Surrogates,* in IEEE Sensors Journal, vol. 22, no. 24, (2022), pp. 24230-24241. doi: 10.1109/JSEN.2022.3219594.

[2] Dedi Darwis, Abhishek R Mehta, Novi Eka, Samsugi, Priya, *Digital Smart Collar: Monitoring Cow Health Using Internet of Things* International Symposium on Electronics and Smart Devices (ISESD), (2022), DOI: 10.1109/ISESD56103.2022.9980682.

[3] Boris,Evstatiev,Nikolay,Valov,Seher,Kadirova,Teodo,Nenov *Implementation of a Prototype IoT-Based System for Monitoring the Health, Behavior and Stress of Cows*, 2022 IEEE 9th Electronics System-Integration Technology Conference (ESTC) IEEE,DOI: 10.1109/ESTC55720.2022.9939489.

[4] M. Alsaaod, M. Luternauer, T. Hausegger, R. Kredel and A. Steiner, *The cow pedogram Analysis of gait cycle variables allows the detection of lameness and foot pathologies.* J. Dairy Sci. , vol. 2, no. 100, (2017), pp. 1417–1426, 2017

[5] Akash Trivedi, Pinaki Sankar Chatterjee *CARE: IoT enabled Cow Health Monitoring System*, 2022 International Conference on Intelligent Technologies (CONIT), (2022), IEEE | DOI: 10.1109/CONIT55038.2022.9847701.

[6] Abdul Aziz Chaudhry, Rafia Mumtaz, Syed Mohammad Hassan Zaidi, Muhammad Ali Tahir, Syed Hassan Muzammil School *Internet of Things (IoT) and Machine Learning (ML) enabled Livestock Monitoring*, 2020 IEEE 17th International Conference on Smart Communities: Improving Quality of Life Using ICT, IoT and AI (HONET), (2020) IEEE | DOI: 10.1109/HONET50430.2020.9322666.

[7] G. Suseendran, D. Balaganesh *Cattle Movement Monitoring and Location Prediction System Using Markov Decision Process and IoT Sensors*, 2021 2nd International Conference on Intelligent Engineering and Management (ICIEM).

[8] Varun Mhatre, Vishwesh Vispute, Nitin Mishra, *IoT based Health Monitoring System for Dairy Cows* "IoT based Health Monitoring System for Dairy Cows", IoT based Health Monitoring System for Dairy Cows, (2020).

[9] Achour Brahim, Belkadi Malika, Aoudjit Rachida, Lalam Mustapha, Lalam Mustapha, Laghrouche Mourad *Dairy cows real-time behaviour monitoring by energy-efficient embedded sensor*, IEEE, (2020).

[10] Pratama, Y. P., Kurnia Basuki, D., Sukaridhoto, S., Yusuf, A. A., Yulianus, H., Faruq, F., and Putra, F. B. *Designing of a Smart Collar for Dairy Cow Behavior Monitoring with Application Monitoring in Microservices and Internet of Things-Based Systems.* 2019 International Electronics Symposium (IES), (2019). https://doi.org/10.1109/elecsym.2019.8901676.

[11] S. T. Ahmed, V. V. Kumar and J. Kim *AITel: eHealth Augmented-Intelligence-Based Telemedicine Resource Recommendation Framework for IoT Devices in Smart Cities,* in IEEE Internet of Things Journal, vol. 10, no. 21, (2023) pp. 18461-18468,, doi: 10.1109/JIOT.2023.3243784.

[12] Vannieuwenborg, F., Verbrugge, S., and Colle, D. *Designing and evaluating a smart cow monitoring system from a techno-economic perspective.* 2017 Internet of Things Business Models, Users, and Networks, (2017). https://doi.org/10.1109/ctte.2017.8260982.

[13] Mulla, A. I., Mulik, A. P., Prashant, A., and Gawai, D. D. *Continuous health surveillance system for cattle.* 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), (2017). https://doi.org/10.1109/iccons.2017.8250656

[14] Khanh, P. C. P., Dinh Chinh, N., Cham, T. T., Vui, P. T., and Tan, T. D. *Classification of cow behavior using 3-DOF accelerometer and decision tree algorithm.* 2016 International Conference on Biomedical Engineering (BME-HUST), (2016). https://doi.org/10.1109/bme-hust.2016.7782100.

[15] Smith, K., Martinez, A., Craddolph, R., Erickson, H., Andresen, D., and Warren, S. *An Integrated Cattle Health Monitoring System.* International Conference of the IEEE Engineering in Medicine and Biology Society, (2006).https://doi.org/10.1109/iembs.2006.259693.

[16] Vannieuwenborg, F., Verbrugge, S., and Colle, D. *Designing and evaluating a smart cow monitoring system from a techno-economic perspective.* 2017 Internet of Things Business Models, Users, and Networks, (2017). https://doi.org/10.1109/ctte.2017.8260982.

[17] S. T. Ahmed, V. V. Kumar, K. K. Singh, A. Singh, V. Muthukumaran, and D. Gupta, *6G enabled federated learning for secure IoMT resource recommendation and propagation analysis* Computers and Electrical Engineering, vol. 102, (2022), p. 108210, doi: 10.1016/j.compeleceng.2022.108210.

[18] Khatate, P., Savkar, A., and Patil, C. Y. *Wearable Smart Health Monitoring System for Animals.* 2nd International Conference on Trends in Electronics and Informatics (ICOEI), (2018). https://doi.org/10.1109/icoei.2018.8553844.

[19] Ronghua, G., JingQiu, G., and Jubao, L. *Cow Behavioral Recognition Using Dynamic Analysis.* International Conference on Smart Grid and Electrical Automation (ICSGEA), (2017). https://doi.org/10.1109/icsgea.2017.26.

[20] Junior, R. L. *IoT applications for monitoring companion animals: A systematic literature review.* 14th International Conference on Innovations in Information Technology (IIT), (2020). https://doi.org/10.1109/iit50501.2020.9299045.

[21] 21. Natarajan, R., Lokesh, G. H., Flammini, F., Premkumar, A., Venkatesan, V. K., and Gupta, S. K. *A Novel Framework on Security and Energy Enhancement Based on Internet of Medical Things for Healthcare 5.0.* Infrastructures, 8(2), (2023), 22. https://doi.org/10.3390/infrastructures8020022.

[22] Sindu Divakaran, Lavanya Manukonda, N Sravya, Melinda Morais M, Janani P, Sindu Divakaran, Lavanya Manukonda, N Sravya, Melinda Morais M, Janani P. *IOT clinic-Internet based patient monitoring and diagnosis system.* IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI-2017).

[23] Panda, S., and Panda, G. *Intelligent Classification of IoT Traffic in Healthcare Using Machine Learning Techniques.* 2020 6th International Conference on Control, Automation and Robotics (ICCAR), (2020). https://doi.org/10.1109/iccar49639.2020.9107979.

[24] Ankit Bhavsar, and Harshal Arolkar *ZigBee Based Network Architecture for Animal Health Monitoring*, 2015 1st International Conference on Next Generation Computing Technologies (NGCT-2015).

Edited by: Polinpapilinho Katina
Special issue on: Scalable Dew Computing for Future Generation IoT Systems
Received: Nov 1, 2023
Accepted: Feb 1, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1716–1727, DOI 10.12694/scpe.v25i3.2798

PERFORMANCE ANALYSIS FOR OPTIMIZED LIGHT WEIGHT CNN MODEL FOR LEUKEMIA DETECTION AND CLASSIFICATION USING MICROSCOPIC BLOOD SMEAR IMAGES

MAHMOUD SAED ALKHOULI*AND HIREN JOSHI †

Abstract. The objective of this work is to create a diagnostic tool for the early diagnosis of leukaemia which is a serious type of cancer affecting bones and blood. Acute lymphoblastic leukemia (ALL) is the most dangerous form of leukemia. Doctors diagnose it by blood samples under powerful microscopes with enhanced lenses which can be slow and is sometimes affected by disagreements among experts. Therefore, the purpose of this work was to create a profound diagnostic tool for the early diagnosis of leukaemia.We proposes an Optimized Light Weight CNN to detect ALL at the early stage. Fragmentation and classification based on preprocessing are the two main components of the suggested method. Artificial images are created during the segmentation process and then tamed by chromatic modification. The proposed model is used to extract the best deep features from every blood smear image to predict the presence of ALL. The work was tested by two lymphoblastic leukaemia image databases (ALL_IDB1 and ALL_ IDB2). Deep-learning (DL) models-based segmentation and classification techniques have recently been introduced for detecting ALL; however they still have certain drawbacks. The proposed approach was assessed with few DL parameters like accuracy,F1 score, precision, recall and Area under the curve. In comparison to the most recent research studies already published; the suggested strategy produced exceptional classification accuracy as 99.56%, F1 score as 99.53%.

Key words: Acute lymphoblastic leukaemia, Optimized Light Weight CNN, leukaemia, blood smear image.

1. Introduction. One of the most important elements of the human anatomy is blood. It contains RBCs (Red Blood Cells) and polypropylene or plasma [1]. Plasma makes up the majority of the blood's composition. White blood cells, commonly known as WBCs, platelets each make up less than one percent of the blood's total volume. Red Blood Cells, White blood cells and Platelets [2-4] are the three primary components of blood that may be distinguished by the blood's appearance, colour, size, chemical make-up, and texture, respectively. The RBC is the most significant type of blood sample, and one of its fundamental components is haemoglobin. Hemoglobin is responsible for the red colour of blood and is responsible for carrying oxygen to all regions of the body. When there is a fall in haemoglobin levels, there is also a decrease in oxygen, which causes exhaustion and weakness. There are four million to six million red blood cells in every individual micro liter of blood, which accounts for forty to forty-five percent of the total volume of blood [5]. WBCs protect us from pathogens and provide us antibodies and resilience; the number of WBCs that may be found in one micro liter of blood can range anywhere from 4000 to 10,000 [6]. Platelets have a concentration in the blood that ranges from 1 million to 5 million per micro liter and are responsible for the clotting process [7]. Therefore, alterations in the levels of any of the fundamental components of blood can result in adverse effects on a human's body, including conditions like neutropenia, leukaemia, and opathies. Because it encompasses both RBCs and platelets, a large WBC density is associated with a compromised immune system in the body. ALL), invasive lobular leukaemia, acute myeloblastic leukaemia, and chronic myeloblastic leukaemia are the four subtypes of leukaemia [8] that are recognized by professionals in the medical field. Chronic myeloblastic leukaemia is the most common form of the disease. Of them, acute lymphoblastic leukaemia (ALL) is the most prevalent, accounting for 70% of all instances of leukaemia, and the most deadly [9].

In addition, environmental factors as well as genetic predispositions play an important part in the progression of the disease. The excessive and inappropriate expansion of neutrophils in the bone marrow is the underlying cause of ALL [10]. ALL can be broken down into one of three different structural categories: S1, S2, S3 [9].S1 cells are the tiniest and have the most homogenous population as well as abrasive chromatins.

*Gujarat University, Ahmedabad, India (al-khouli@hotmail.com).
†Gujarat University, Ahmedabad, India (hdjoshi@gujaratuniversity.ac.in).

S2 cells are significantly bigger than S1 cells and have more nuclear variety than the S1 cells. S3 cells are larger than S1 cells and have compartments that protrude into the cell rather than being contained within the cell. Therefore, getting a timely diagnosis of ALL is extremely important for the healing process, especially for young children [10].

The fact that standard and lymphocytes cell types share numerous similarities, on the other hand, creates difficulties for early lymphocyte diagnosis [11]. As a consequence of this, lymphocytes were separated into three distinct categories: normal, unusual, and aggressive. Normal lymphocytes can be identified by their cohesiveness as well as their shaped, tiny, and jagged nuclei; unusual cells can be identified by their enormous diameter and nucleus as well as the fact that their chromatins are doughy; and aggressive cells can be identified by their homogeneity as well as the fact that they are surrounded by red cells. The different types of lymphocytes can be identified through a process called micron - sized investigation, which requires taking samples of bone marrow or blood and submitting them to a pathologist for analysis [12, 13].

However, a proper diagnosis of leukaemia requires the collection and analysis of a sample of bone marrow. As when the analysis is performed individually, it is arduous, time-consuming, and susceptible to divergent expert judgments. Therefore, the accuracy of a human diagnosis rests on the skill of the pathologist, despite the possibility of human mistake. Several studies have developed automated methods for identifying leukaemia by identifying WBC characteristics from micrographs. Thus, the automated classification of blood cell pictures will result in a speedy and accurate diagnosis and will make it possible to examine several cells from every individual. ML ad DL can solve human diagnosis issues. It has been demonstrated that the convolutional neural network (CNN), which has an improved ability to discern between healthy and malignant cells, can assess and resolve many of the deficiencies of manual diagnostic and medical imaging.

In this investigation we used two publicly available dataset ALL-IDB1 and 2 to test support our proposed Optimized Light Weight CNN model. The work has also contributed in:

- Fragmentation and classification based on preprocessing are the two main components of the suggested method.
- Artificial images are created during the segmentation process and then tamed by chromatic modification. The proposed model is used to extract the best deep features from every blood smear image.
- A fine tuned technique was employed amongst CNN models to classify deep features and to achieve promising predictive performance.
- Blood microscopy image analysis systems were developed to aid bone marrow biopsies and professionals in making appropriate diagnostic judgments.

The article contains Literature review in Section 2.The methods and materials of the proposed work are explained in Section 3.Results are analyzed in Section 4 and the work is concluded in Section 6.

2. Related work. During the blood sample image optimization phase of the experiment conducted by authors of [14], the target region was improved, which had an effect on the segmentation outcome .The resampled and misleading edges of the photos created by Razzak et al [15] were eliminated utilizing wiener filter, which allowed the images to be improved. In order to produce high-quality segmented pictures, pre-processing techniques such as k-means, median filtering, and contrast stretching were performed by Manglem et al [16]. The images that were used in the research carried out by the authors were processed by making use of the HSV (hue, saturation, value), CMYK (cyan, magenta, yellow and key) and RGB (red, green, blue) colour spaces.HSV, in order to generate additional images for the purpose of identifying the characteristics that are most significant through the use of principal component analysis to acquire White blood cells nuclei which was proposed by Hellmich et al [17].Authors Wang et al [18] initiated the DL strategy to leukocyte classifying and segmenting during or before and fragmentation steps. They then used a deep convolutional network to optimise the results, and then the DarkNet- and ShuffleNet models were used to extract deep features.

Authors of [19] improved images by reducing their luminance when transforming from RGB to HSV. They then used fcm method to separate the cores from the rest of the image (the hydrologic process breaks the relation between the clusters and the image context and then pulls out the largest significant design and analytical features), presented an intuitive method for improving images organized by ML algorithms. Scotti et al [20] Used DL techniques for cell radius convergence, image sharpening and shape etc.

Using segmentation with obscured C-means clustering and then classification with five different classifiers, Satpatti et al [21] presented a statistical infinitesimal technique for discerning pernicious lymphocytes from normal capillary images and symptomatic lymphocytes. This method was accomplished by first using shadowed C-means clustering for segmentation. When evaluating leukaemia, Yaakob et al [22] described many phenotypic and ecological characteristics, and how these features were fed into four classification techniques; all of the algorithms obtained superior diagnosis outcomes across all age demographics. Alrefai et al [23] Performed feature selection that used the algorithm for particle swarm optimization using the ensemble learning approach, and then used five classification algorithms to score the features that were picked; the techniques produced favorable outcomes.

Mandal et al [24] Developed a technique for screening cancer cells that involves the extraction of critical traits (such as an enlarged nuclei and neighboring nuclei, both of which are indicative of cancer cells) through the utilization of various learning algorithms. Shah J.H et al [25] proposed an efficient method for analyzing the blood database for the purpose of treating leukocytes. This system consisted of the following phases: enhancing pictures, assembling wavelets, retraining the dataset, and categorizing the supplied classes through using CNN model.

Using DeepLabv3 and ResNet-50, Ameer et al [26] proposed a saliency detection method for extracting leukocytes from the remaining portion of the image in order to obtain deep feature maps; the system obtained good WBC recognition accuracy. Marr et al [27] performed an analysis using a CNN model as a component of training. They developed the model, and when the number of images increased while it was being trained, they came to the conclusion that the model was even more successful when it comes to the majority of training sets.

Pooja et al., [28] used T cells, or T lymphocytes, that are responsible for cell-mediated immunity. These cells mature in the thymus gland, from which they derive their name. There are various subtypes of T cells, each serving specific functions. Helper T cells (CD4+) assist other immune cells, activating both B cells and cytotoxic T cells. Jayachitra et al., [29] suggested Cytotoxic T cells (CD8+) in combating infections by attacking and destroying infected or damaged cells. On the other hand, Gupta et al.,[30] used B cells for conducting their investigation, where it contributes to immunity by generating antibodies that target specific pathogens.

3. Methods and materials. This part of the article contains the methods and materials used for the entire investigation Fig. 3.1 represents the proposed architecture.

3.1. Dataset. In this study, the publicly accessible ALL-IDB dataset, which contains specimens of blood samples, was used to test neural networks and other hybrids of DL models. The most deadly type of leukaemia, ALL, is the focus of the dataset. Experts in lymphoma have detected and categorized every lymphoma in each image. All of the photographs were captured with a fluorescence microscopy in a JPG format with a G5 PowerShot Canon camera with a high resolution of 2592 × 1944.

ALL-IDB datasets come in two varieties: ALL IDB1 (subset1) and ALL IDB2 (subset2). Around 108 total images, that has 49 lymphoma and 59 photographs of normal individuals, are included in the ALL-IDB1 collection. Each image shows over 39,000 blood components that have been categorized by lymphoma specialists. On the other hand, the ALL-IDB2 dataset has 260 and 130 of lymphomas and normal cells respectively. The blasted cells and healthy tissues from the ALL-IDB1 dataset were used to crop the ALLIDB2 dataset.

Samples from both datasets are shown in Fig. 3.2. The dataset is collected from https: //www.kaggle.com/nikhilsharma00/leukemia-dataset.

3.2. Image preprocessing. The initial stage in imaging techniques is preprocessing. When examining blood samples underneath a microscope, the illumination of the microscope is altered to capture photographs of the samples. As a result, AI imaging approaches perform worse as a result of fluctuations in the luminance of the microscope and variations in reflections caused by light. So, using noise reduction algorithms can help you get better photographs. The photos in this study were improved by finding the parameters of the RGB colour streams, and the colors were then stabilized using cropping. The noise was then eliminated, and the contrast of the edges was improved using Laplacian filters until the image was smooth, the filter is gradually positioned around the image. Each central pixel was swapped out for an average of 35 nearby pixels to lessen the disparities between the pixels.Equ. 3.1 to Equ. 3.3 represents the average work of image preprocessing. y(m) represents the input, ω^2 d represents the differential equation.

$$Improvedimage = y(m) - \mho^2 d \tag{3.1}$$

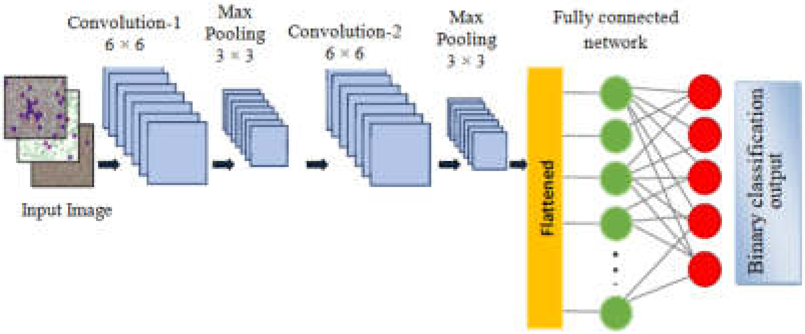


Fig. 3.1: Proposed architecture for Optimized Light Weight CNN

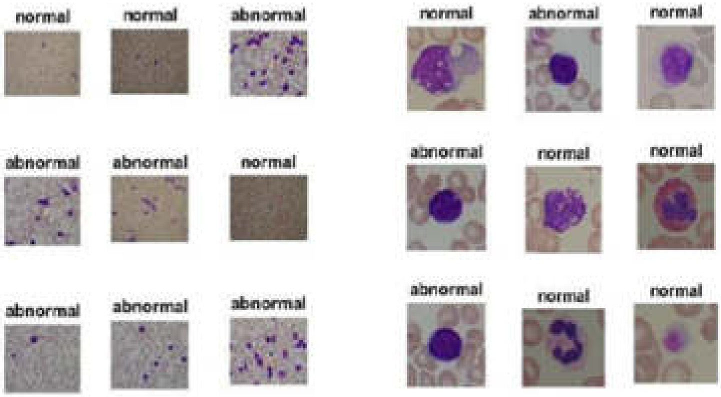


Fig. 3.2: Samples of Dataset (a) Subset1-ALL-IDB1 (b) Subset2-ALLIDB2

$$y(m) = m - 1 + \sum_{i=1}^{m}(y(m-1)) \tag{3.2}$$

$$\mho^2 d = \frac{f(nx)}{(y)} + \frac{f(ny)}{(x)} \tag{3.3}$$

y(m) is the input image,(m-1) represents the preceding input, and m represents the number of pixels present in the image. A Laplacian filter is used to reduce the noise in the image, which is represented in Equ. 3.3,in which x and y are the matrix coordinates. Finally an improved image is acquired by subtracting the value of the filtered image with the average y(m) obtained from Equ. 3.2. Few samples of the enhanced image is as shown in Fig. 3.3.

3.3. Optimized Light Weight CNN. Our proposed Optimized Light Weight CNN consists of Convolution, pooling, flattening layers, and multilevel hidden neurons made up the majority of the CNN architecture. Fully connected neural networks (FCNNs) used automatic extraction of characteristics from the input photos to classify them. Pooling layer and convolution layer were used to extract features. The features were obtained after binarizing the image in these layers, and the classification process was then initiated. Feature Learning, Receptive Fields, sharing information among parameters, structural arrangement, shift variance in pooling, non linear activations are the specific reasons for following a layered approach in the proposed method. The

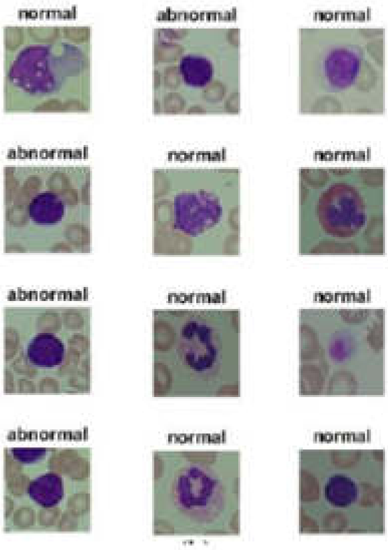


Fig. 3.3: Samples of improved image

specifics of the architecture we created are shown in Fig. 3.3.The phase wise explanation of the proposed model is explained as follows.

Layer: 1-Convolution Layer This layer is in charge of utilizing various attributes to investigate various filters on the source images. We utilized a CNN with a 64 feature map that was 6 × 6 in size. The image was subjected to convolution filters through sliding. The settings for the filter were chosen at random. To prevent over fitting, we employed three convolutional layers.

Layer: 2-Max Pooling Layer The Max pooling layer is in charge of reducing the size of the segmented image to help it focus on any significant feature, region, or entity. We utilized a max-pooling layer in our network with a size of 2×2. We increased this layer's quantity by two.

Layer: 3-Flattening Layer A two-dimensional max-pooled matrix was converted into a one-dimensional array by this layer to ensure that each cell may serve as an input node for the entire connected network.

Layer: 4-Fully connected Layer With one input layer (in our example, the flattened layer), an output and hidden layer, this component was a naive linked complete forward network. The hidden layer in our model has 128 nodes, Dropout of 10 percent, and Batch-Normalization to reduce over fitting. The ReLU activation function had been applied for a straightforward computation. Two types of optimization techniques stochastic gradient descent and ADAM optimizers are set up at the output layer, each optimizer at a time. Each of the five output nodes we added, one for every leukaemia class and normal class samples which was managed by a Soft Max activation mechanism. Since this configuration was more appropriate for the sample size of dataset we employed, our Optimized Light Weight CNN model was trained with 30 epochs and 64 batch sizes. To get the optimal performance, different epoch counts were tested. Although we attempted to raise the number of epochs to 100, the additional running time was not significantly faster.

Algorithm 1: Optimized Lightweight CNN for Leukemia Detection

Input: Smear segmented images

Output: Leukemia classified images

Build the Optimized CNN model

1. Convolutional Layers

```
model.add (layers.Conv2D (64,(6,6), activation='relu',
input_shape=(image_height, image_width, channels)))
model.add (layers.Conv2D (64, (6, 6), activation='relu'))
model.add (layers.Conv2D (64, (6, 6), activation='relu'))
```

2. Max Pooling Layers

```
model.add (layers.MaxPooling2D ((2, 2)))
model.add (layers.MaxPooling2D ((2, 2)))
```

3. Flattening Layer

```
model.add (layers. Flatten ())
```

4. Fully Connected Layer

```
model.add (layers.Dense (128, activation='relu'))
model.add (layers. Dropout (0.1))
model.add (layers.BatchNormalization ())
```

5. Output Layer

```
model.add (layers.Dense (num_classes, activation='softmax'))
return model
```

6. Model training and evaluation

```
train_model (model, train_data, validation_data, epochs, batch_size):
model. Compile (optimizer='adam', loss='categorical_crossentropy',
metrics=['accuracy'])
model.fit (train_data, epochs=epochs, batch_size=batch_size,
validation_data=validation_data)
cnn_model = build_cnn_model ()
train_model (cnn_model, train_data, validation_data, epochs=30, batch_size=64)
```

3.4. Experimental Analysis. The convolutional layer is one of the most important layers. Equ. 3.4 describes how this layer conducts a linear operation known as convolution in between filter w(t) and the image x(t). The convolution layer is governed by three variables: p.step, filter size and, zero padding. The enveloping around the images increases as the filter size increases. Every filter is intended to find particular elements throughout the input image. For instance, some filters are made to recognize edges, others to recognize geometric characteristics, and still others to recognize patterns and shades. Translation invariance is the term used to describe this characteristic of CNNs. To preserve the size of the original input, zero padding utilized. The convolutional filter and original input sizes are used to calculate the size of the zero pad. The number of steps the filter applies to the image at once is determined by the p-step parameter.

$$In(t) = (k^{*}w)\,(r) = \int k(n)w(u-v)\cdot da \tag{3.4}$$

'In(t)' represents the input and 'u' and 'v' are the integer values. Using 'u' and 'v'the same two classes of normal and healthy, 'k(n)' represents the preceding input ,'w' as weights.we investigated developing a classification model in this experiment, where image transformation techniques were applied to ALL afected and not affected data. As a result, we used 275 samples for testing and 1000 samples for training for both classes. In addition, we used 5-fold cross validation in this study.

The dimensionality and color information of an image should be prioritised when implementing our Optimized Light Weight CNN model; as a result, convolutional filters are customised for the input images. Two-dimensional images are processed using the convolutional layer of the filter kf, using image as the input, as illustrated in Equ. 3.5. The convolutional layers in the case of the RGB input images operate on separate two dimensional convolutions for every colour: R, G, and B. A rectified linear unit (ReLU) layer is added after a number of convolutional layers for additional processing. The negative input is inhibited and turned into 0 by this layer, which transmits the positive input. The ReLU layer only transmits positive characteristics and transforms negative values to zero, as shown in Equ. 3.6.

$$s(i,j) = (I * K)(\mathrm{i,j}) = \sum_{m}\sum_{n} I(m,n)K(i-m,j-n) \tag{3.5}$$

$$ReLU(\mathrm{x}) = max(0,x) = \begin{cases} x, x \geq 0 \\ 0, x < 0 \end{cases} \tag{3.6}$$

Table 3.1: Data split for training and testing the proposed approach

Dataset	Phase	No. of samples
ALL_IDB1 (subset-1)	Training-80%	Leukemia Affected ....42 Leukemia Not Affected.....39
	Validation-10%	Leukemia Affected ....09 Leukemia Not Affected.....10
	Testing -10%	Leukemia Affected ....12 Leukemia Not Affected.....13
ALL_IDB2 (subset-1)	Training-80%	Leukemia Affected ....78 Leukemia Not Affected.....69
	Validation-10%	Leukemia Affected ....28 Leukemia Not Affected.....26
	Testing -10%	Leukemia Affected ....16 Leukemia Not Affected.....16

Before using any imagery transformation algorithms, we split the dataset in half, using 30 percent of the overall for testing and 70 percent for training. Then, in order to enhance the amount of data samples, we applied the image transformations to both portions. In each cross-validation iteration, we made sure that each dataset component had the same quantity of images with various folds. As a result, we ran our trials for each fold separately, after which we determined the accuracy and loss metrics for each fold. The final performance was expressed as a five-fold average. The data split for training and testing is shown in Table. 3.1.

Convolutional layers high amount of parameters, which leads to an over fitting issue. Our Optimized Light Weight CNN model offers a remedy for this issue by utilizing a dropout layer. Each time the dropout layer is used, 50 percent of the neurons are stopped and 50 percent are passed. The dropout layer in this investigation was adjusted at 50%. The training time is doubled by this layer. The training process is slowed down by the high-dimensional feature maps produced by convolutional layers. Thus, Our Optimized Light Weight CNN model offers a remedy for this issue by utilizing a dropout layer. Each time the dropout layer offer pooling layers to minimize the dimensionality in order to accelerate the training phase. The same convolutional layer technique powers interactions between pooling layers within CNNs.

$$P(i;j) = \max_{m,n=1\ldots k} A[(i-1)p+m;(j-1)p+n] \tag{3.7}$$

$$P(i;j) = \frac{1}{k^2} \sum_{m,n=1\ldots\ldots k} A[(i-1)p+m;(j-1)p+n] \tag{3.8}$$

The max and average-pooling layers are two different categories of pooling layers. Equ. 3.7 illustrates how the upper limit is selected when employing the max-pooling layer work. On the other hand, the average specified values are calculated and replaced by the average value in the average-pooling layer work mechanism, as indicated in Equ. 3.8.

In our Optimized Light Weight CNN the layer in charge of classification is the FCL. All neurons are connected to one another within the FCL. Two dimensional feature maps are transformed into one-dimensional maps using the FCL layer. Different CNNs have different numbers of FCLs; some networks have more than one FCL, which assigns each image to the appropriate class. Finally, the softmax activation function receives the FCL result and generates neurons with the same amount of classes supplied. Softmax developed two types of neurons for this study: leukaemia affected and normal.

The proposed investigation was conducted as follows: TensorFlow and OpenCV were utilized for image preprocessing. Python was used for this deep learning scripting. Development of the model was facilitated by employing an Integrated Development Environment (IDE) Jupyter Notebooks and effective collaboration was maintained through version control using Git. Model evaluation tools, including metrics such as accuracy and F1-score, were integrated, along with data annotation tool Labelbox for preparing datasets. A GPU,

Table 4.1: Results obtained from the proposed approach

Model Classification	Training Accuracy	Validation Accuracy	Training Loss	Validation Loss
Leukaemia Affected	99.46 %	94.93 %	0.00362	1.2462
Leukaemia not Affected	99.67 %	92.45 %	0.00594	0.8357
Average Assessment	**99.56 %**	**93.68 %**	**0.00478**	**1.04095**

from the NVIDIA GeForce RTX series was used to expedite the computationally intensive training process, while general computation and data handling were necessitated by a multi-core CPU, 16GB RAM. Docker for containerization and optimization libraries Intel's MKL are also used. The model was experimented in 80:20 ratio where 80 per cent of the input images are used for testing and remaining 20 per cent for testing the model.

3.5. Model Evaluation. We chose 2 key criteria for measuring the performance of our proposed Optimized Light Weight CNN in order to evaluate our model. The first measure of accuracy was the proportion of consistently categorized input images among all samples. We evaluated our model by 1. Training accuracy, which assesses how well the model performs during training, and validation accuracy, which demonstrates how well the model performs while classifying unknown data. 2. The loss metric concentrates on figuring out the prediction error and is utilized to modify the weights of nodes. Additionally, the training loss and validation loss is also computed.

4. Results and Discussion. The overall summary of the findings of our investigation trial which is assessed with few DL parameters like accuracy,F1 score,recall,precision and AUC that are as represented in Table 4.1. Using the input samples, we were able to classify leukaemia affected and not affected in binary form with the best performance which records the recognition accuracy as 99.56%. The model identified more complexity to distinguish between the classes as there are more classes adhered to and included in a classification process. In terms of accuracy and loss measures, we found that SGD optimizer performs significantly better than ADAM optimizer.

Additionally, we found that extended epoch iterations had little effect on the model's performance. One of the benefits of our proposed approach is that they are resource-flexible because they just need computers with moderate requirements, as opposed to other existing models, which need machines with high specifications. In addition, our proposed approach differs from CNN models in that they train more quickly. CNN models require a lot of training time whereas our approach utilized less amount of time for classification and recognition. The time concern of the proposed method is noted by the Receiver operating curve values obtained which is discussed in further part of this section.

The results of each fold's performance for our proposed approach are shown in Table 4.2. In binary classification, fold 4 received the highest score. We can draw the conclusion that the performance results may vary greatly depending on the samples used for the test and training set. We assessed our strategy using 5-fold cross-validation since several fold cross validation continues to produce more trustworthy outcomes in evaluation. Table 4.3 provides the final overall performance metrics assessed with respect to ML parameters

The proposed Optimized Light Weight CNN for the early detection of leukaemia using two named datasets is depicted in Fig. 4.1by the confusion matrix. All dataset samples that were wrongly identified in the auxiliary diameter (True_Positive and True_Negative) but correctly labelled in the prime diameter (True_ positive and True_ negatives) are included in the confusion matrix (False_Positive and False_Negative).

Cross-entropy, a measurement of the error rate between the observed and projected outputs, is one way to assess how well the deep learning models. Fig. 4.2 displays the training, testing and validation results of our approach. The performance is shown in three different colors: red when testing, blue for training, and green throughout validation. The intersecting lines were used to obtain the optimum performance. As the number of epochs progressed, the error rate between the actual and expected outputs reduced; the training ended when

Table 4.2: Folds wise results of the proposed approach

Model Classification	Training Accuracy	Validation Accuracy	Training Loss	Validation Loss
Fold-1	99.46%	94.93%	0.00362	1.2462
Fold-2	99.82%	93.98%	0.10478	0.0098
Fold-3	99.29%	93.79%	0.00982	1.1062
Fold-4	99.73%	94.68%	0.00327	0.9834
Fold-5	99.58%	94.82%	1.00387	0.7263
Average	**99.56%**	**93.68%**	**0.003997**	**0.9978**

Table 4.3: Overall results of the proposed approach

Model Classification	Accuracy	Precision	Recall	Sensitivity	F1 Score	Error rate
Leukaemia Affected	99.46 %	98.91 %	97.11 %	98.62 %	99.15 %	0.0026
Leukaemia not Affected	99.67 %	99.43 %	99.38 %	98.79 %	98.93 %	0.1935

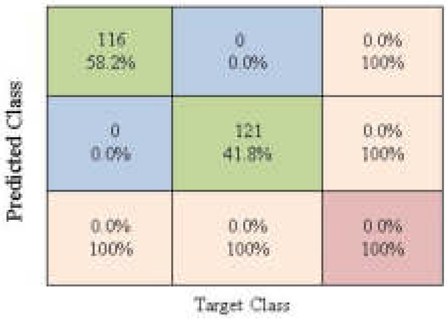


Fig. 4.1: Confusion Matrix

the method's least error value was reached. The proposed model's best validation result was 0.034817 during the fifth epoch , and other phases showed the same performance.

The execution time of our approach is recorded and assessed in terms of Receiver Operating Curve (ROC).This curve exhibits the graphical performance of the model.Fig. 4.3 depicts the ROC of our proposed approach.

The performance of our suggested pertinent systems is compared to the evaluation findings of earlier systems in Table 4.4. The accuracy of the prior systems ranged from 89% to maximum of 93% that of our proposed system was 99.5%. The existing systems' precision levels were 90% and 94% respectively, but our suggested system was 98%.

The sensitivity of the compared existing systems systems ranged from 89% to 93%, but the recall of the system we proposed was 99.38%. In terms of specificity, the earlier methods achieved 89% and 94%, but our suggested solution achieved 99.03%. The prior systems' AUCs were 83% and 97%, whereas our approach attained 99.21%.

5. Conclusion. In the medical field, AI technologies have surfaced to offer analytical ability and diagnosing tools with maximum reliability. Techniques utilizing ML and DL algorithms solve issues with conventional diagnosis limitations, expert disagreements, time-consuming monitoring of blood samples. These methods are

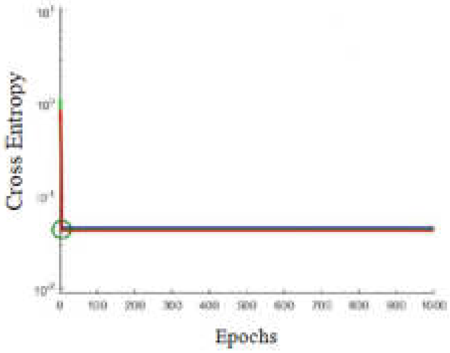


Fig. 4.2: Performance measure of the proposed approach

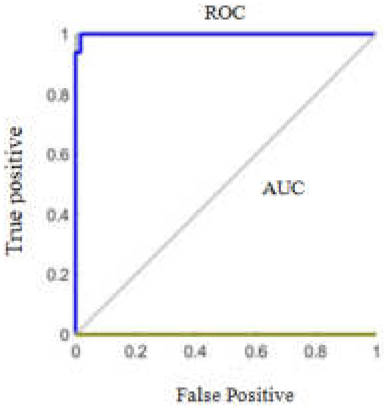


Fig. 4.3: ROC and AUC values obtained for the proposed model

Table 4.4: Existing versus proposed approach

Model	Accuracy (%)	F1 Score (%)	Specificity (%)	Recall (%)	Precision (%)	Error rate	AUC (%)
Inception V3	89.32	90.32	89.60	93.44	90.61	0.426	97.43
DCNN	92.43	89.85	93.21	89.42	91.92	0.392	83.82
AlexNet	93.15	91.37	94.84	89.28	94.76	0.562	92.46
Proposed approach	**99.56**	**99.53**	**98.79**	**99.38**	**98.91**	**0.01**	**99.21**

essential for the early identification of leukaemia. It affects white blood cells, bone marrow, and the immune system. Blood smears are the frequent diagnostic tool. In this paper, we described a new method for diagnosing leukaemia from microscopic blood pictures utilizing an Optimized Light Weight CNN. Our model established its ability by employing data augmentation strategies to overcome the over fitting problem. It surpassed previous machine learning methods by reaching 99.56% accuracy for binary classification of one leukaemia type as affected and not affected. All experiments included cross-validation. Medical picture categorization takes a long time to execute, but it's crucial to make sure the model is stable throughout. In the upcoming phases, there are plans to broaden this research by delving into quantum computing algorithms to enhance the precision of White Blood Cell (WBC) detection. Subsequent efforts will include the implementation of a hybrid approach, integrating features derived from pre trained CNN models with those obtained from Spatial Pattern Analysis, Gray Tone Spatial Relationships, and Self-Similarity Encoding algorithms in a comprehensive feature vector. This combined feature vector will then be fed to Feed forward Neural Networks, for image classification.

REFERENCES

[1] Huang, N.T et al, "A micro fluidic device for simultaneous extraction of plasma, red blood cells, and on-chip white blood cell trapping". *Scientific Report, 2018, 8, 1–9.*

[2] Dhiman, G., Vinoth Kumar, V., Kaur, A., and Sharma, A. "Don: deep learning and optimization-based framework for detection of novel coronavirus disease using x-ray images." Interdisciplinary Sciences: Computational Life Sciences, 2021 , 13, 260-272.

[3] Zadeh, H et al. "Automatic recognition of five types of white blood cells in peripheral blood". *Computerized Medical Imaging and Graphics 2011, 35, 333–343*

[4] Nabavi, S.M.and Nabavi et al, "Flavonoids and platelet aggregation: A brief review." European Journal of Pharmacology, 2017, 807, 91–101

[5] Al-Megren and S. Kurdi et al, "Red blood cell segmentation by thresholding and canny detector". *Procedia Computer Science, 2018, 141, 327–334.*

[6] Natarajan, R., Lokesh, G. H., Flammini, F., Premkumar, A., Venkatesan, V. K., and Gupta, S. K. "A Novel Framework on Security and Energy Enhancement Based on Internet of Medical Things for Healthcare 5.0. " Infrastructures, 2023, 8(2), 22.

[7] Heemskerk P.E and, J.Wen et al. "Platelet biology and functions: New concepts and clinical perspectives". *Nature Reviews Cardiology, 2019, 16, 166–179.*

[8] Sawyers and Denny, "Leukemia and the disruption of normal hematopoiesis." Cell 1991, 64, 337–350.

[9] Messmore and H.L, "Wintrobes Atlas of Clinical Hematology." *JAMA 2007, 297, 2641–2645.*

[10] S. T. Ahmed, V. V. Kumar and J. Kim, "AITel: eHealth Augmented-Intelligence-Based Telemedicine Resource Recommendation Framework for IoT Devices in Smart Cities," in IEEE Internet of Things Journal, 2023, vol. 10, no. 21, pp. 18461-18468.

[11] Fati S.M et al." Score and Correlation Coefficient-Based Feature Selection for Predicting Heart Failure Diagnosis by Using Machine Learning Algorithms". *Computational and Mathematical Methods in Medicine, 2021, 85003.*

[12] Nam, Y. and Raza, M. "3D semantic deep learning networks for leukemia detection." Computers, Materials & Continua 2021, 69, 785–799.

[13] Muthukumaran, V., Joseph, R. B., and Uday, A. K. "Intelligent medical data analytics using classifiers and clusters in machine learning." *In Handbook of Research on Innovations and Applications of AI, IoT, and Cognitive Technologies, 2021, pp. 321-335.*

[14] Ramaneswaran, and Srinivasan, K et al "Hybrid Inception v3 XG Boost Model for Acute Lymphoblastic Leukemia Classification." Computational and Mathematical Methods in Medicine 2021, 2021, 2577375.

[15] Razzak et al, "M.I. Efficient leukocyte segmentation and recognition in peripheral blood image Technology," *Health Care 2016, 24, 335–347.*

[16] Dhanachandra, N and Manglem, K. "Image Segmentation Using K -means Clustering Algorithm and Subtractive Clustering Algorithm". Proceedings Computer Science. 2015, 54, 764–777.

[17] Hellmich, H.L et al. "Principal component analysis of blood micro RNA datasets facilitates diagnosis of diverse diseases." *PLoS ONE, 2020, 15, e0234185.*

[18] Iqbal, M and Wang, S.H.et al "A deep network designed for segmentation and classification of leukemia using fusion of the transfer learning models." Complex and Intelligent Systems 2021, 1, 1–16.

[19] Mirmohammadi, P et al, "A Recognition of acute lymphoblastic leukemia and lymphocytes cell Subtypes in microscopic images using random forest classifier". *Physical and Engineering Sciences in Medicine, 2021, 44, 433–444.*

[20] Scotti, F et al. "ALL Detection Based on Adaptive Unsharpening and DL". In Proceedings of the ICASSP 2021-2021 IEEE Internl. Conf. on Acoustics Speech and Signal Processing Toronto, ON, Canada, 6–11 June 2021, pp. 1205–1209.

[21] Satpathy, S et al. "An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images". *Neural Computing and applications, 2014, 24, 1887–1904*

[22] Yaqoob, M. et al," Identification of significant risks in pediatric acute lymphoblastic leukemia through machine learning (ML) approach." Medical & Biological Engineering & Computing, 2020, 58, 2631–2640

[23] Alrefai, N. "Ensemble ML for Leukemia Cancer Diagnosis based on Microarray Datasets." *International Journal of Applied Engineering Research, 2019, 14, 4077–4084.*

[24] Rajagopalan, V and Mandal et al. "ML based system for automatic detection of leukemia cancer cell". In Proc. of the 2019 IEEE 16th India Council Internl. Conf. (INDICON), Rajkot, India, 13–15 December 2019; pp. 1–4.

[25] Shah, J.H. and Fernandes, S.L." Robust discrimination of leukocytes protuberant types for early diagnosis of leukemia". *Journal of Mechanics in Medicine and Biology 2019, 19, 1950055.*

[26] Ameer, P.M et al. "Segmentation of leukocyte by semantic segmentation model: A DL approach". Biomedical Signal Processing and Control,2021, 65, 102385.

[27] Bosna.D.and Marr, C et al." Tens of images can suffice to train neural networks for malignant leukocyte detection." *Scientific Reports 2021, 11, 1–8.*

[28] Pooja, S. and Megha, G.S. "Using CNN to Detect Cancerous Cells in White Blood Cells from Bone Marrow through

Microscopic Images." Journal of Advancement in Parallel Computing, 2021, 4(2).

[29] Jayachitra, J. and Umarkathaf, N. March. "Blood Cancer Identification using Hybrid Ensemble Deep Learning Technique." *In 2023 Second International Conference on Electronics and Renewable Systems (ICEARS) , 2023, pp. 1194-1198. IEEE.*

*[30] Gupta, R., Gehlot, S. and Gupta, A. "C-NMC: B-lineage acute lymphoblastic leukaemia: A blood cancer dataset. "*Medical Engineering and Physics,2022, 103, p.103793.

[31] Available: https://www.kaggle.com/nikhilsharma00/leukemia-dataset.

Edited by: Polinpapilinho Katina
Special issue on: Scalable Dew Computing for Future Generation IoT Systems
Received: Nov 4, 2023
Accepted: Jan 8, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1728–1747, DOI 10.12694/scpe.v25i3.2664

KNOWLEDGE GRAPH ANALYSIS FOR CHRONIC DISEASES NURSING BASED ON VISUALIZATION TECHNOLOGY AND LITERATURE BIG DATA

SIYU DUAN AND YANG ZHAO*

Abstract. The use of knowledge graph analysis for chronic disease nursing based on visualization technology and literature big data is an unexplored area of research in this field of study. To uncover research hotspots and developmental trends in the field of chronic disease nursing, and to provide a scholarly reference, we employed mathematical and statistical methods along with CiteSpace literature visualization analysis software for quantitative analysis of extensive literature data from the Web of Science Core Collection. We examined aspects such as publication trends, journals, author collaborations, research institutions, national and regional distributions, keyword co-occurrence, clustering, time zones, emergence, literature co-citations, and more. These analyses identified the current hotspots and future directions for research. Notably, scholars' interest in chronic disease nursing exhibited a consistent upward trajectory. In particular, the field of artificial intelligence technology application in nursing yielded $3,610$ published papers in 141 journals with more than or equal to 10 published papers on the topic, accounting for 58.41% of the total number of published papers in this field of study. Furthermore, the top three publishers were the "Journal of Clinical Nursing," "Journal of Advanced Nursing," and "BMC Health Services Research." Among authors, Hu, Frank B., Willett, Walter C., and Rimm, Eric B., ranked as the top three, and 12 authors had more than 10 publications. The most active research institutions included Harvard University, Harvard Medical School, Brigham & Women's Hospital, University of California System, University of London, US Department of Veterans Affairs, Veterans Health Administration (VHA), Harvard T. H. Chan School of Public Health, University of Sydney, and the University of Toronto. The United States, Australia, England, China, Canada, Netherlands, Spain, Italy, Sweden, and Germany emerged as the leading countries in terms of research output, while emerging hotspots encompassed topics such as incidence, rheumatoid arthritis, qualitative research, burnout, kidney transplantation, critical illness, COVID-19, Sars-COV-2, public health, and the well-being of medical staff. These findings present valuable insights for prospective research endeavors.

Key words: Chronic diseases, nursing, literature big data, bibliometric analysis, trends, hotspots

1. Introduction. Chronic non-communicable diseases (NCDs) stand as the world's foremost cause of death and disability [16], constituting a staggering 73.6% of chronic disease-related deaths, as reported in the "World Health Statistics 2021" by the World Health Organization (WHO) [20]. In 2019, chronic diseases were responsible for nearly 70% of the global disease burden, and they accounted for a staggering 88.5% of deaths in China, with 80.7% attributed to cardiovascular diseases, cancer, and chronic respiratory diseases. Presently, the field of chronic disease prevention and control confronts challenges, underscoring the pressing need to fortify chronic disease nursing practices.

Bibliometrics employs mathematical and statistical methods to scrutinize vast repositories of literature within specific domains and databases, unearthing the current state of research in these fields. This approach predominantly measures documents, authors, and word counts, harnessing mathematical and statistical techniques to conduct a quantitative examination of the knowledge contained within these documents. The insights derived from bibliometrics are further visualized through knowledge graphs, offering a more objective portrayal of the research landscape. Scholars rely on bibliometrics and related literature analysis software to comprehensively dissect the research status and identify hotspots within nursing, ultimately providing invaluable reference points for fellow researchers. For instance, Juan-Jose delved into bibliometric and gender-based analyses of scientific publications within Scopus and Web of Science, offering insights into annual article production, prominent authors, top-cited articles, and thematic keyword analyses [2]. Similarly, Cant conducted a bibliometric exploration of highly-cited virtual simulation nursing education articles, revealing rankings, topic diversity, and authorship patterns [4].

*School of Health and Wellness, Panzhihua University, Panzhihua 617000, China (zhaoyang@pzhu. edu. cn). This work was sponsored by China Panxi Health Industry Research Center (No. PXKY-YB-202213).

Besides, Hahn engaged in quantitative and statistical analyses of publication trends, prolific authors, highly-cited documents, and keywords about clinical reasoning in nursing [9]. In the same vein, Su, through bibliometrics, explored the trends in high-impact international nursing core competencies research [13]. While Wang employed literature visualization analysis to unveil research hotspots and future directions in Traditional Chinese Medicine (TCM) nursing for insomnia [19]. That said, Yesilbas conducted a literature data analysis to investigate the knowledge structure and developmental process in nursing empowerment [21]. Building on the findings of past studies in this area, Zhang utilized VOSviewer and CiteSpace to scrutinize COVID-19-related nursing research, uncovering the current state and hot topics within this realm [22]. Likewise, Zhao employed CiteSpace and VOSviewer to assess the application of virtual reality technology in nursing studies [23].

Contemporaneously, De Oliveira conducted a bibliometric analysis to discern trends in burnout research among nursing professionals, comparing the contributions of various countries, institutions, journals, authors, keywords, and citations [6]. Intriguingly, Ghamgosar also employed bibliometric analysis to offer insights into global research output on geriatric nursing [7]. Whereas, Huang evaluated the literature on family nursing to identify development trends and research focal points [10]. Correspondingly, Molassiotis conducted a bibliometric exploration of disaster nursing, unveiling global development and trends [12]. Just as Blazun adopted an automated, electronic approach to scrutinize the nursing informatics literature, tracing its historical origins, and analyzing the evolution of topics and themes contribute to the understanding of knowledge development within nursing informatics [17].

By the same token, Guo harnessed literature mining and information visualization technologies to examine the bibliometric characteristics of cirrhosis nursing articles in the Web of Science spanning from 1986 to 2020. This endeavor aimed to comprehensively depict the present state of this field and furnish essential evidence for enhancing research in nursing and clinical liver cirrhosis within Mainland China [8].

To unveil the current landscape and pressing concerns within chronic disease nursing, this study sought to provide a reference point for further research endeavors by domestic scholars. Employing mathematical and statistical techniques, along with the CiteSpace literature visualization analysis software, we conducted a quantitative exploration of literature big data about chronic disease nursing. This analysis encompassed ten years from 2013 to 2022 and involved key facets such as publication trends, journals, author collaborations, research institutions, national and regional distributions, keyword co-occurrence, clustering, time zone mapping, emergence, literature co-citation, and more. By scrutinizing the interplay and internal correlations within this wealth of information, we aimed to unearth research hotpots and development trends to guide the field of chronic disease nursing.

The initial section of this paper provides an introduction to the research context, status, innovations, and primary contributions. The remaining sections of the paper are arranged as follows: The subsequent section focuses on elucidating the objectives, design, sample, search strategy, inclusion criteria, and statistical methods. The third part delves into a comprehensive discussion based on the results derived from CiteSpace literature visualization analysis, offering insights into pertinent research hotpots and evolving trends. The closing section of the paper presents the conclusions drawn from our research endeavors.

2. Materials and methods.

2.1. Objectives. The objectives of this study encompass the following key aspects: (1) Identifying the principal contributors in the realm of nursing research linked to Virtual Simulation (VS), including countries, institutions, journals, authors, and articles. (2) Analyzing collaborative relationships within this field. (3) Constructing a knowledge network and pinpointing the frontier topics, thus elucidating future directions in this domain.

2.2. Study design. A descriptive bibliometric analysis was conducted on publications within the domain of nursing related to chronic disease research, retrieved from a comprehensive literature database.

2.3. Data source. The data for this research were obtained from the Web of Science™ (WOS) database, specifically the Web of Science™ Core Collection, encompassing SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI, CCR, EXPAND, and related indices.

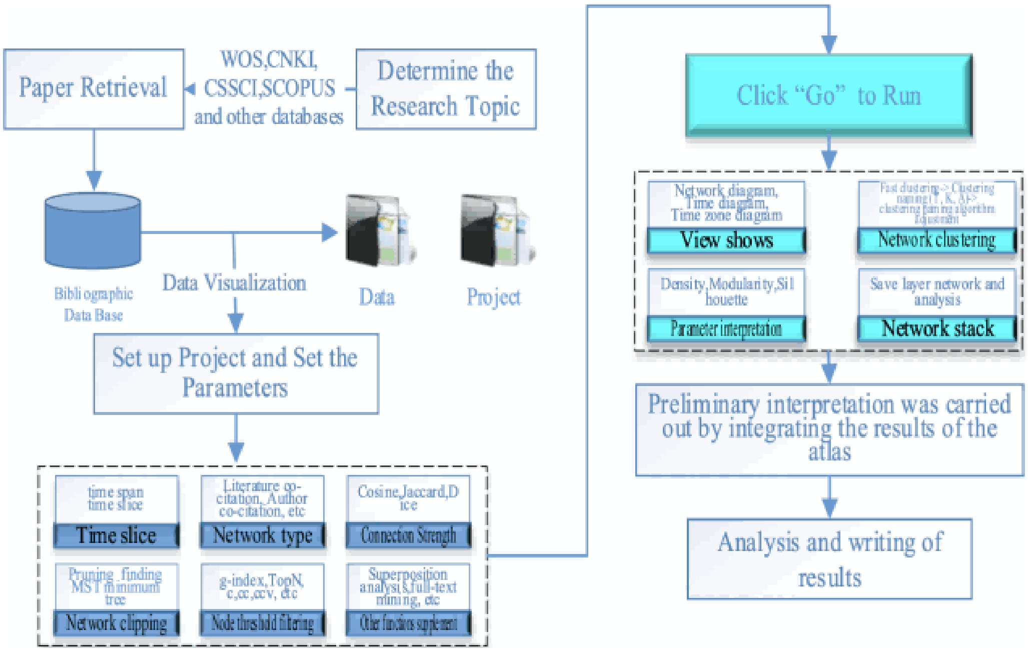

Paper Retrieval
WOS,CNKI,
CSSCI,SCOPUS
and other databases
Determine the Research Topic
Bibliographic Data Base
Data Visualization
Data
Project
Set up Project and Set the Parameters
time span
time slice
Time slice
Literature co-citation, Author co-citation, etc
Network type
Cosine,Jaccard,Dice
Connection Strength
Pruning, finding MST minimum tree
Network clipping
g-index,TopN, c,cc,ccv, etc
Node threshold filtering
Superposition analysis,full-text mining, etc
Other functions supplement
Click "Go" to Run
Network diagram, Time diagram, Time zone diagram
View shows
Network clustering
Density,Modularity,Silhouette
Parameter interpretation
Save layer network and analysis
Network stack
Preliminary interpretation was carried out by integrating the results of the atlas
Analysis and writing of results

Fig. 2.1: Process of CiteSpace visual analysis

2.4. Search method. The search was conducted on April 23^{rd}, 2023, utilizing Web of Science. The search query utilized the following formula: TS = (chronic disease OR chronic non-communicable disease OR chronic illness or chronic non-communicable illness) AND (nursing OR nurse), AND DOP = $(2013-01-01/2022-12-31)$) AND DT =(Article OR Review)) was used to screen out publications associated with on chronic diseases nursing. This method efficiently filtered publications relating to chronic diseases nursing, specifying that they were either full papers or review articles in the field of nursing. The search yielded a total of 6,180 literature records.

2.5. Inclusion criteria. The following criteria were employed for inclusion:

(1) Peer-reviewed articles involving VS related to nursing
(2) Original articles and review articles
(3) Web of Science core collection (WoSCC) literature big database
(4) The language of the document is English

2.6. Statistical analysis. Bibliometric methods and CiteSpace visualization techniques were employed to analyze the annual volume of articles, authors, institutions, countries or regions, journals, keywords, and literature citations within the scope of Chronic disease nursing research. This approach was undertaken to gauge the influence and attention accorded to each country or region in this field. The process of literature visualization analysis using CiteSpace is depicted in Figure 2.1 [14, 5].

The primary procedural steps, as outlined in reference [5], are as follows:

Step 1: Define the research focus, which, in this instance, pertains to chronic disease nursing.

Step 2: Gather literature data by formulating a tailored search strategy aligned with the research focus established in Step 1. This strategy may include keyword and topic searches. Additionally, specify the sources of literature, such as WOS, CNKI, CSSCI, SCOPUS, etc., and execute the literature search in the respective databases by the established search strategy. It is crucial to preprocess the retrieved literature, with a particular note that only data from WOS can be directly utilized and analyzed within CiteSpace. Literature data collected from sources like CNKI, CSSCI, SCOPUS, or others necessitate conversion into WOS format for compatibility with CiteSpace.

Step 3: Create an analysis project within CiteSpace. Configure analytical parameters such as time segmentation, network type, and correlation strength. Afterward, initiate the analysis process using CiteSpace.

Step 4: Visualize the results. Review the analysis outcomes (detailed analysis content can be found in reference [24], and, as needed, adjust clustering algorithms and relevant parameters. Generate visual displays for various analysis types, including network diagrams, timeline graphs, and temporal zone charts.

Step 5: Conduct a visual analysis. Utilize the insights from the analysis results and the designated view types to undertake a comprehensive analysis, integrating domain-specific knowledge to produce the analysis report.

2.7. Calculation algorithms for key literature analysis indices. (1) The Ziff's Law of Co-word Analysis can be expressed as:

$$\ln C = \ln f + \ln r, 0.1 < C < 1 \tag{2.1}$$

where f signifies the frequency of literature occurrence, r denotes the rank number of literature frequency, and C represents a constant within the range of $0.1 < C < 1$.

(2) Betweenness Centrality calculation algorithm. The calculation algorithm for Betweenness Centrality is given as:

$$BC_i = \sum_{s \neq i \neq t} \frac{n_{st}^i}{g_{st}} \tag{2.2}$$

where g_{st} denotes the number of shortest paths from node s to node t, and n_{st}^i symbolizes the number of shortest paths that traverse nodes within the shortest path between node s and node t.

(3) Network density calculation algorithm. The Network Density is computed using equation (2.3) below:

$$Density = \frac{m}{C_n^2} = \frac{2m}{n(n-1)} \tag{2.3}$$

With m representing the number of actual network relations, and n being the number of network nodes. Additionally, equation (2.4) defines the associations between nodes i and j in the network.

$$Q = \frac{1}{2m} \sum_{i,j} \left(a_{ij} - p_{ij}\right) \sigma \left(C_i, C_j\right) \tag{2.4}$$

where $A = a_{ij}$ is the adjacency matrix of the actual network, p_{ij} is the expected value of the number of connecting edges between nodes i and j, and C represents the associations between nodes i and j in the network. If C_i, C_j are part of the same club, $\sigma \left(C_i, C_j\right) = 1$. Otherwise $\sigma \left(C_i, C_j\right) = 0$. The size of Q relates to node density, and typically, a Q value greater than 0.3 is preferred.

(4) Silhouette calculation algorithm. The Silhouette Calculation Algorithm distinguishes between classes, assigning a value of 0 if $a(i)$ equals $b(i)$, 1 if $a(i)$ exceeds $b(i)$, and 0 if $a(i)$ is less than $b(i)$ according to equation (2.5).

$$S_i = \begin{cases} 1 - a(i)/b(i) & \text{if } a(i) < b(i) \\ 0 & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1 & \text{if } a(i) > b(i) \end{cases} \tag{2.5}$$

Here, $a(i)$ signifies the average distance between point i and other points in the class, while $b(i)$ denotes the average distance between point i and all points in the class of the nearest point i. The average silhouette value S can be used to measure the cluster's homogeneity, with a higher S value indicating greater homogeneity within the network. Generally, an S value greater than 0.5 is indicative of a highly reliable cluster.

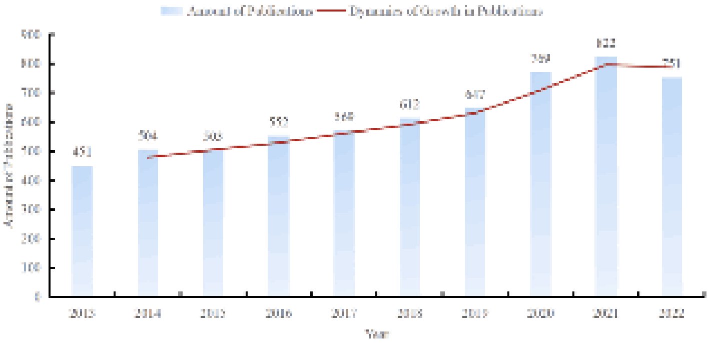


Fig. 3.1: Number of published papers from 2013 to 2022

3. Results and discussion. In this section, we configured analysis parameters in CiteSpace, including time segmentation, network type, and correlation strength, and subsequently conducted the analysis. We imported a total of $6,180$ literature records from the WoSCC database into the CiteSpace software, employing keywords as analysis nodes. The visualization analysis was performed using CiteSpace $6.2r^2$, with the software set to run from 2013 to 2022. A one-year time slice was applied, and thresholds were established for authors, institutions, countries, keywords $K = 15$, and co-citation $N = 50$. Pruning methods included Pathfinder, Annual pruning, and overall network pruning. We utilized the LLR algorithm to integrate and analyze the data, as well as to visually present the results.

3.1. Annual number of published papers. Analyzing the temporal evolution of the number of published papers in the WoSCC database provides a macroscopic perspective on the research hotspots within the field. Therefore, we documented the annual number of papers related to chronic disease nursing from 2013 to 2022, as illustrated in Figure 3.1.

As depicted in Figure 3.1, the number of papers addressing chronic disease nursing exhibited a consistent uptrend from 2018 to 2021, underscoring the sustained global interest in this subject across various countries and regions. However, it's important to note that the data for 2022 might not fully represent the total number of papers published on the topic of chronic disease nursing for that year, as certain papers from 2022 may remain unpublished or are yet to be included in the WoSCC.

3.2. Publication journals. A total of 141 journals featured ten or more published papers on the topic of chronic diseases nursing, collectively accounting for 58.41% of the total published papers. Notably, the top three journals with the highest publication counts were "The Journal of Clinical Nursing," with 218 papers, followed by "The Journal of Advanced Nursing," with 194 papers, and "The BMC Health Services Research," with 132 papers. Other notable contributors to the literature included "BMJ Open" with 128 papers, "International Journal of Environmental Research And Public Health" with 99 papers, "Plos One" with 98 papers, "International Journal of Nursing Studies" with 66 papers, "Cochrane Database of Systematic Reviews" with 64 papers, "Journal of The American Association of Nurse Practitioners" with 56 papers. More so, journals publishing over 40 articles are highlighted in Figure 3.2.

3.3. Geographic distribution. The analysis of papers published in the field of chronic disease nursing within the WOS literature database by different countries or regions was conducted using CiteSpace, with the node type set as "Country." The visual representation of papers published through collaborative efforts between countries or regions is illustrated in Figure 3.3.

Within Figure 3.3, the size of nodes signifies the volume of papers published in the respective country or region, while the connecting lines between nodes indicate the level of collaboration between different countries or regions. Analogously, the line thickness corresponds to the strength of cooperation.

Furthermore, this analysis encompasses a total of 138 nodes and 386 connections, resulting in an overall network density of 0.0408. This density value suggests a significant presence of countries or regions within this

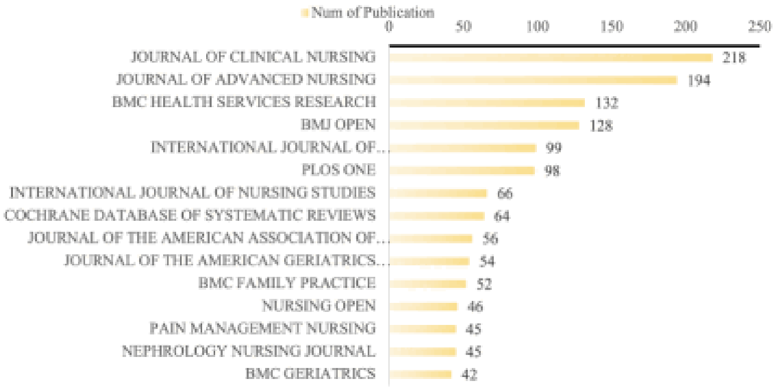


Fig. 3.2: Journals publishing over 40 articles

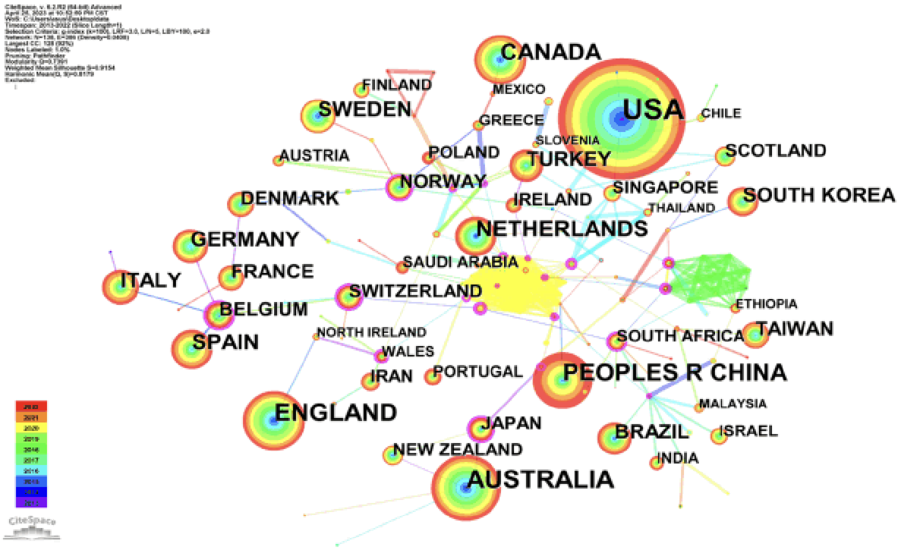


Fig. 3.3: Spectrum of papers published jointly among countries or regions

field, underlining the closeness of collaboration between them. Expectedly, the United States emerged as the leading country in terms of publication count, closely followed by Australia and England. The top ten countries or regions with the highest number of published papers can be identified by examining the publication counts across different nations or regions, as presented in Table 3.1. With a frequency of 5, 491, these top contributors account for 46.28% of the total. Notably, the United States outpaces all others in terms of published papers, with a significant lead. Besides, it's worth noting that the number of papers published in the United States surpasses the combined output of the following four countries or regions. From a centrality perspective, countries or regions with higher publication counts generally exhibit more pronounced centrality.

Nevertheless, it is apparent that the publication numbers in countries like China and Canada, despite their volume, do not correlate proportionally with their centrality, indicating less-than-ideal cooperative relationships with other countries or regions.

3.4. Institutional distribution. We harnessed CiteSpace software to visually analyze the dataset within the WOS literature database. Our analysis encompassed the time range from 2013 to 2022, with a yearly breakdown. The k value was set to 25, and Node Types were designated as "Institution." Pruning was executed

Table 3.1: Top ten countries or regions of published papers

Ranking	Country	Frequency	Centrality	Began Year
1	USA	2089	0. 16	2013
2	AUSTRALIA	677	0. 13	2013
3	ENGLAND	559	0. 23	2013
4	PEOPLES R CHINA	481	0. 01	2013
5	CANADA	443	0. 04	2013
6	NETHERLANDS	301	0. 06	2013
7	SPAIN	276	0. 03	2013
8	ITALY	235	0. 07	2013
9	SWEDEN	216	0. 03	2013
10	GERMANY	214	0. 04	2013

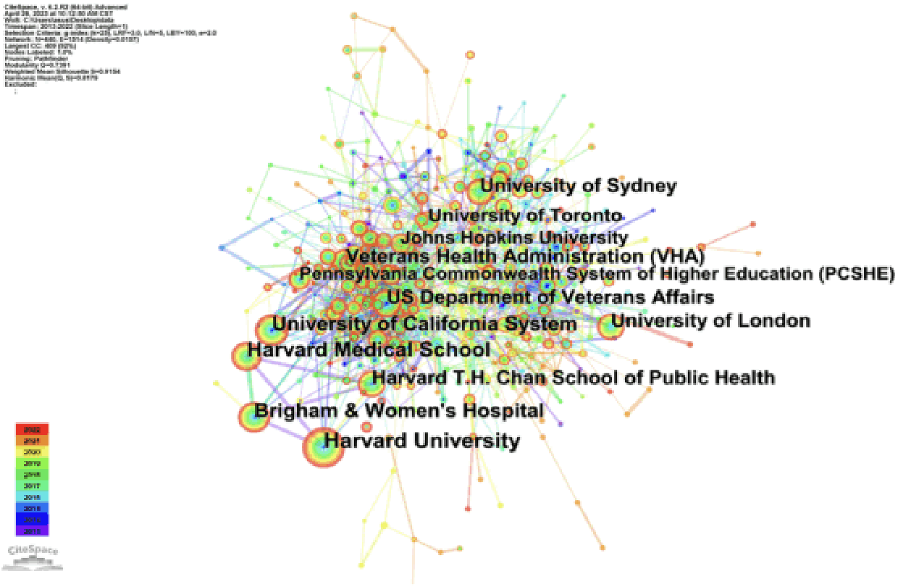

University of Sydney
University of Toronto
Johns Hopkins University
Veterans Health Administration (VHA)
Pennsylvania Commonwealth System of Higher Education (PCSHE)
US Department of Veterans Affairs
University of California System
University of London
Harvard Medical School
Harvard T.H. Chan School of Public Health
Brigham & Women's Hospital
Harvard University
CiteSpace

Fig. 3.4: Distribution network map of research institutions

using the Pathfinder mode on an annual basis and for the entire network, while other options remained in their default configurations. This process culminated in a visual depiction of research institutions' distribution, as displayed in Figure 3.4. Moreover, in Figure 3.4, the size of nodes corresponds to the number of papers published by each research institution, while the connecting lines delineate the degree of collaboration between various institutions. Additionally, the color of the lines indicates the collaborative relationships across different periods.

As depicted in Figure 3.4, the network comprises 440 nodes and 1, 514 connections, with a network density of 0.0157. This indicates a substantial presence of research institutions, with notable connections among key institutions. The primary institutional collaboration network is notably centered around Harvard University and Harvard Medical School. To gain deeper insights into the accomplishments and collaborative dynamics of these research institutions, we conducted further data analysis in Figure 3.4. This analysis revealed the top ten research institutions with the highest number of publications, as presented in Table 3.2, with a combined frequency of 1, 718 times, accounting for 19.29% of the total. Harvard University, Harvard Medical School, and Brigham & Women's Hospital were the most prolific contributors. Figure 3.4 and Table 3.2 demonstrate the close collaborative network between Harvard University, Harvard Medical School, and Brigham & Women's

Table 3.2: Top 10 research institutions with the number of publications

Ranking	Institutions	Year	Papers	Cooperation Degree
1	Harvard University	2013	288	33
2	Harvard Medical School	2013	203	32
3	Brigham & Women's Hospital	2013	188	29
4	University of California System	2013	168	38
5	University of London	2013	161	38
6	US Department of Veterans Affairs	2013	156	54
7	Veterans Health Administration (VHA)	2013	154	53
8	Harvard T. H. Chan School of Public Health	2013	150	26
9	University of Sydney	2013	140	27
10	University of Toronto	2013	110	29

Hospital. Equally, in terms of inter-institutional cooperation, the U.S. Department of Veterans Affairs and Veterans Health Administration (VHA) exhibited a relatively high cooperation density, indicating a concentration of foreign scholars' research efforts in major institutions. The elevated density of collaboration between these major institutions underscores the maturity of research cooperation within the international community.

Likewise, the Timezone function was applied to assess cooperative institutions through a time series perspective, with the analysis findings presented in Figure 3.5. Over and above that, Figure 3.5 illustrates that the node size corresponds to the number of papers published by each research institution, the connecting lines delineate the intensity of collaboration between different institutions, and the color of the lines signifies the collaborative relationships across distinct periods. Institutions such as Harvard University, Harvard Medical School, and Brigham & Women's Hospital have a longer history of collaboration.

The peak productivity of institutions was primarily concentrated in the 2013-2014 period, with most of the research institutions represented in yellow, indicating rapid emergence and relatively shorter research duration. These findings align with the shifts in publication trends observed during this timeframe. More recent research institutions in this field include Central South University, Zhejiang University, Yonsei University Health System, and others.

3.5. Authorship distribution. The total number of papers authored by an individual in a journal to some extent signifies the academic standing of that author within the field. The author collaboration network provides a clear depiction of the core author groups and their cooperative relationships in the research domain. For this paper analysis, the analysis node within CiteSpace software was configured as "author," and the amassed literature data underwent visual analysis. The knowledge graph portraying authors and their collaborative networks is presented in Figure 3.6. Additionally, in Figure 3.6, the font and node size correspond to the number of papers published by each author, while the connections between nodes delineate the cooperative relationships between different authors. Similarly, the line thickness indicates the extent of collaboration.

As observed in Figure 3.6, the network encompasses 492 nodes and 566 connections, with an overall network density of 0.0047, signifying robust collaborative ties among authors in the research domain. The most extensive collaborative author network within this field includes figures such as Hu, Frank B., Willett, Walter C., Rimm, Eric B., Rexrode, Kathryn M., and others. Notably, the outer circle of authors, including Willett, Walter C., Rexrode, Kathryn M., and Chan, Andrew T., is depicted in red, indicating recent article contributions. In terms of the number of articles published by these authors, Hu, Frank B., Willett, Walter C., and Rimm, Eric B. secured the top three positions. There were also 12 authors with ten or more articles published. The top ten authors with the highest number of publications are detailed in Table 3.3, accounting for 7.7% of the total. Considering the cooperative degree among research authors, it is evident that the primary authors maintain a relatively high level of cooperation. This signifies the establishment of a close-knit and mature cooperative network within this field. Generally, a substantial correlation exists between highly prolific authors and the density and intensity of their collaborative networks, thereby fostering denser cooperation networks.

To explore author relationships from a time series perspective, this paper employs the Timezone (Time

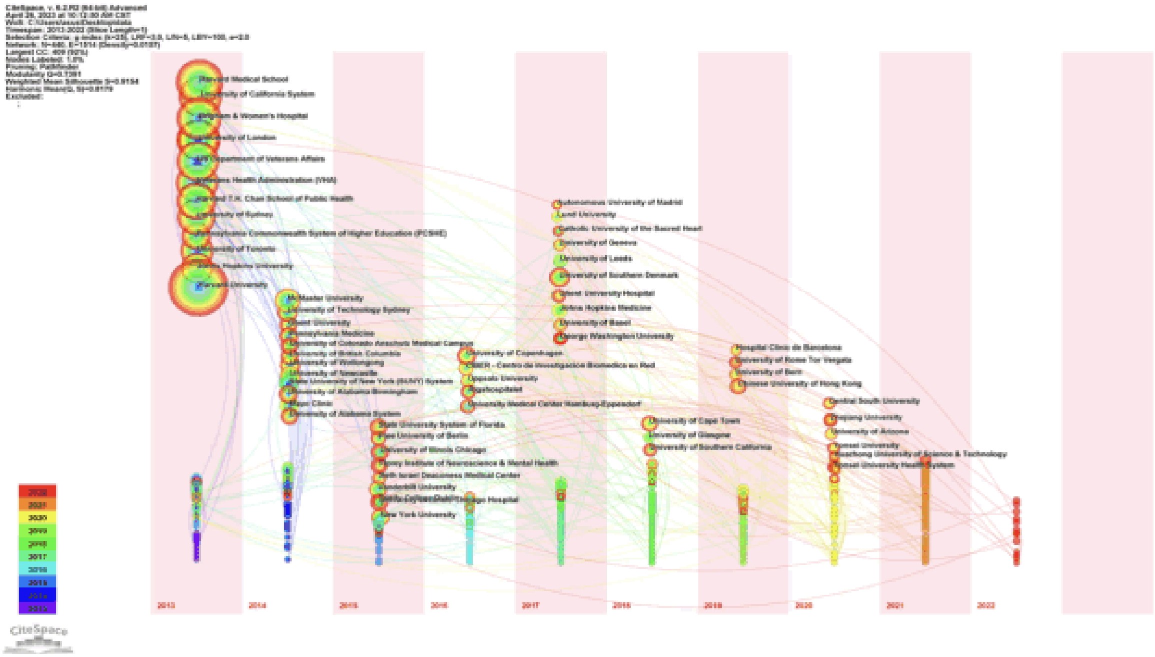

CiteSpace, v. 6.2.R2 (64-bit) Advanced
Harvard Medical School
University of California System
Brigham & Women's Hospital
University of London
US Department of Veterans Affairs
Veterans Health Administration (VHA)
Harvard T.H. Chan School of Public Health
University of Sydney
Pennsylvania Commonwealth System of Higher Education (PCSHE)
University of Toronto
Johns Hopkins University
Harvard University
McMaster University
University of Technology Sydney
Ghent University
Pennsylvania Medicine
University of Colorado Anschutz Medical Campus
University of British Columbia
University of Wollongong
University of Newcastle
State University of New York (SUNY) System
University of Alabama Birmingham
Mayo Clinic
University of Alabama System
State University System of Florida
Free University of Berlin
University of Illinois Chicago
Florey Institute of Neuroscience & Mental Health
Beth Israel Deaconess Medical Center
Vanderbilt University
New York University
University of Copenhagen
CIBER - Centro de Investigacion Biomedica en Red
Uppsala University
Rigshospitalet
University Medical Center Hamburg-Eppendorf
Autonomous University of Madrid
Lund University
Catholic University of the Sacred Heart
University of Geneva
University of Leeds
University of Southern Denmark
Ghent University Hospital
Johns Hopkins Medicine
University of Basel
George Washington University
University of Cape Town
University of Glasgow
University of Southern California
Hospital Clinic de Barcelona
University of Rome Tor Vergata
University of Bern
Chinese University of Hong Kong
Central South University
Zhejiang University
University of Arizona
Yonsei University
Huazhong University of Science & Technology
Yonsei University Health System
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
CiteSpace

Fig. 3.5: Time zone diagram of Institutions

Table 3.3: Top ten authors with several publications

Ranking	Author	Year	Numbers	Connectivity
1	Hu, Frank B	2013	39	26
2	Willett, Walter C	2013	36	29
3	Rimm, Eric B	2015	23	28
4	Bonner, Ann	2016	17	3
5	Rexrode, Kathryn M	2017	14	23
6	Chan, Andrew T	2016	14	17
7	Kubzansky, Laura D	2017	12	14
8	Halcomb, Elizabeth	2015	12	4
9	Manson, JoAnn E	2014	11	15
10	Missmer, Stacey A	2016	11	9

zone diagram) function in CiteSpace, depicting author relationships along a coordinate axis with time as the horizontal parameter, as displayed in Figure 3.7. In this time zone diagram, the node size represents the frequency of an author's presence, the year associated with each node denotes the author's initial appearance, and the color of the lines connecting nodes signifies the timing of the author's co-appearances.

As depicted in Figure 3.7, Hu and Frank B emerge as nodes with the highest number of publications in the related literature, commencing their publication year in 2013. These nodes exhibit extensive connections and a prolonged timespan, highlighting the significant academic status and reference value associated with this author and their work in this field. Over time, the number of authors contributing to related studies increased, and other prolific authors have an extended publication history, indicating the field's sustainability. Remarkably, authors such as Willett, Walter C., Chan, Andrew T., Bonner, Ann, Kubzansky, and Laura D., among others, have both a high publication output and recent contributions, making the exploration of their research trajectories potentially valuable to this field.

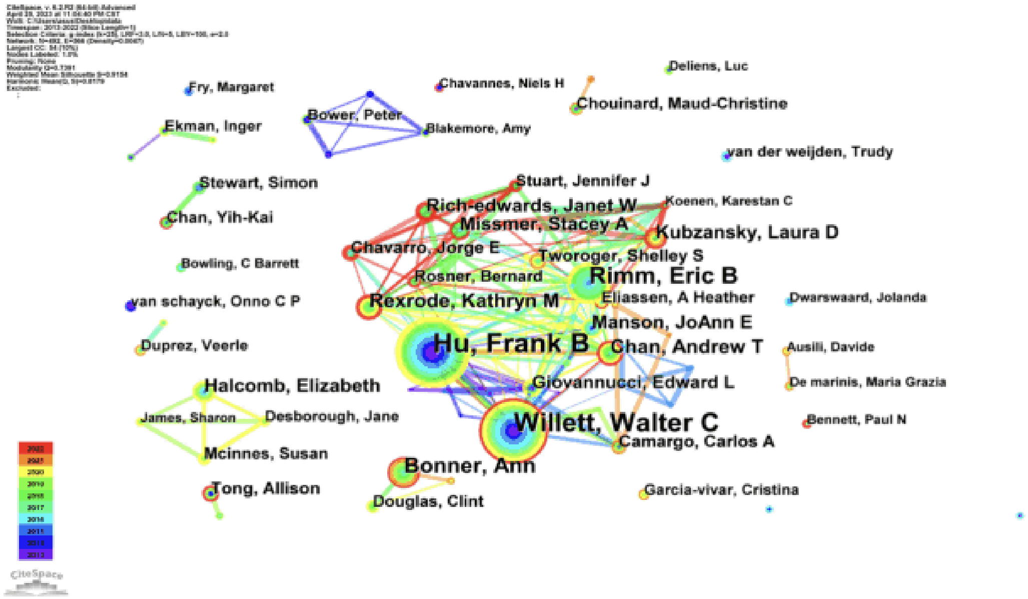


Fig. 3.6: Knowledge graph of authors and their cooperation network

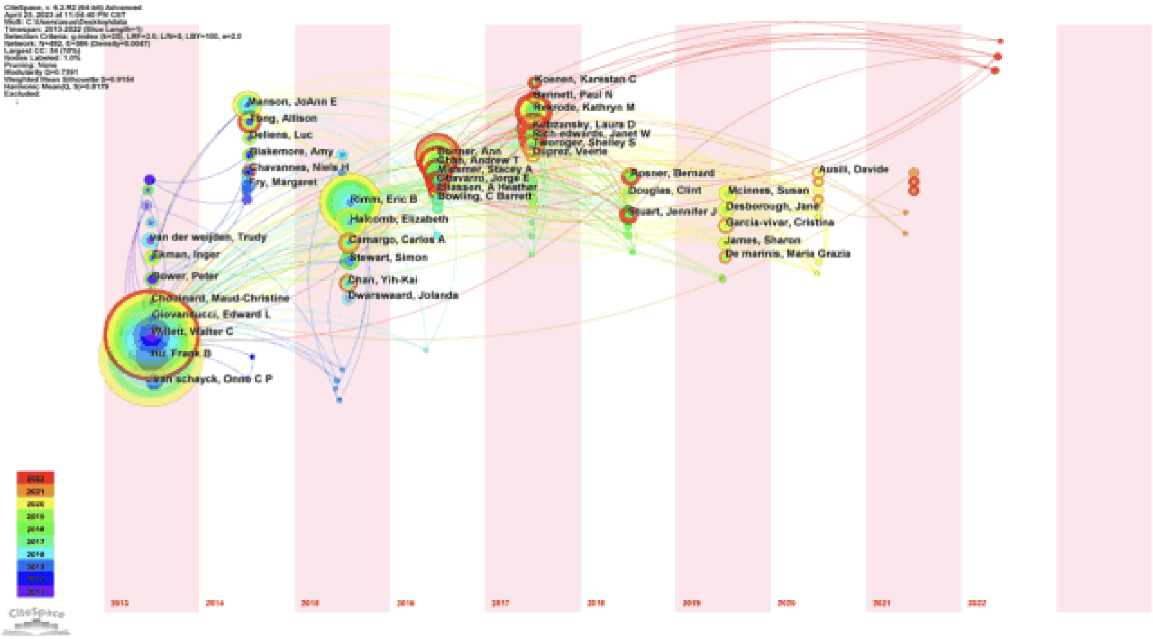


Fig. 3.7: Relationship between authors in the coordinate with time

3.6. High-frequency keywords. Co-word analysis primarily involves extracting title information, such as keywords and abstracts, from citations and forming an informative knowledge map through statistical analysis. Research into high-frequency keywords can elucidate the prevailing trends in the field of chronic disease nursing over a specific period. The software's operating timeframe was set as "2013-2022," with a threshold of $K = 15$, YearPerSlice configured as "1," and pruning performed annually and across the entire network. This

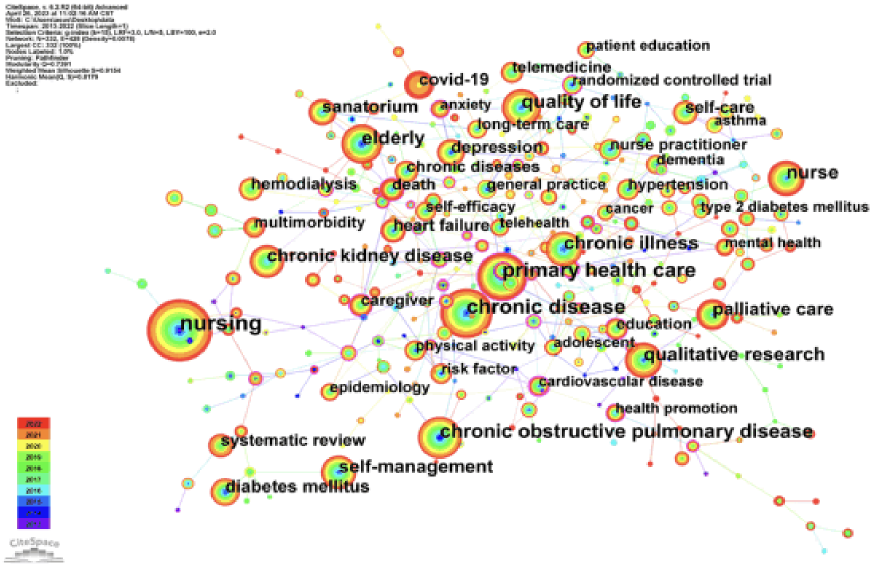


Fig. 3.8: Spectrum of High-frequency keywords

facilitated visual analysis, leading to the creation of a co-occurrence map of frequently used keywords in the literature, showcased in Figure 3.7. Within Figure 3.8, 332 high-frequency keywords were identified, forming 428 connections. The node size and text denote keyword frequencies, while the lines connecting nodes signify associations established during various periods. The thickness and density of these lines reflect the intensity of keyword co-occurrence. Notably, “nursing” emerges as the largest node, followed by “chronic disease” and “primary health care.” In terms of historical presence, keywords like nursing, chronic disease, primary health care, and chronic obstructive pulmonary disease have appeared early. More recently, terms like burnout, kidney transplantation, critical illness, quality of health care, and coronavirus have surfaced, potentially signifying new research directions in chronic disease nursing.

On top of that, the mediating centrality of keywords serves as a pivotal metric for evaluating research hotspots and scholars’ primary areas of interest in this field. Analyzing the mediation centrality index, which represents nodes’ facilitating influence (as shown in Table 3.4), reveals that “incidence,” “rheumatoid arthritis,” and “qualitative” exhibit strong connectivity with other prominent keywords. This suggests that these keywords often lie within the communication path with other terms, actively contributing to the mutual citation relationships among literature.

Methodologically, the use of keywords encapsulates the essential content of scholarly work, and by conducting a co-occurrence analysis of high-frequency keywords, we can pinpoint research hotspots within the domain of chronic disease nursing. The inter-mediation centrality value offers insight into the significance and impact of keywords, with higher values indicating greater mediating influence. Table 3.4 showcases the occurrence frequency and inter-mediation centrality values (Centrality $\geq$ 0)of keywords in the field of chronic disease nursing. As revealed by the centrality values in Table 3.4, ”incidence” boasts the highest centrality value (Centrality ≥ 0.34) and exhibits the closest associations with other keywords. Notably, “rheumatoid arthritis,” “qualitative,” and other keywords also display substantial intermediation centrality values (Centrality $\geq$ 0.3). Considering both keyword occurrence frequency and centrality values, it becomes evident that the primary research foci in chronic disease nursing revolve around “incidence,” “rheumatoid arthritis,” and “qualitative.”

3.7. Keywords clustering. To intuitively visualize the research hot topics within the papers found in the WOS literature big database, we employed CiteSpace software along with the LLR algorithm for keyword co-occurrence cluster analysis. The resulting keyword clustering view is depicted in Figure 3.9, where color blocks delineate distinct clusters, each containing associated keywords. The analysis encompasses $N = 332$ keywords,

Table 3.4: Top ten centrality of keywords

Ranking	Keywords	Frequence	Centrality
1	Incidence	15	0. 34
2	Rheumatoid arthritis	10	0. 32
3	Qualitative	55	0. 3
4	Chronic heart failure	43	0. 28
5	Cardiovascular disease	67	0. 25
6	Adherence	48	0. 24
7	Nurse practitioners	27	0. 21
8	Chronic	23	0. 2
9	Prevalence	31	0. 18
10	Treatment	16	0. 18

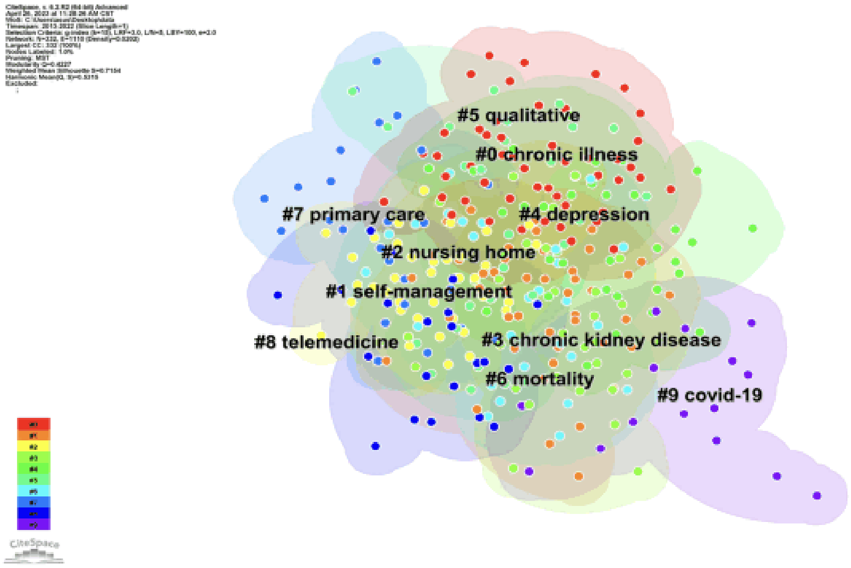


Fig. 3.9: Spectrum of keywords clustering

$E = 1110$ connections, and a network density of 0.0202. Also, the size of module Q, a measure related to node density, plays a crucial role in scientific cluster analysis, with a larger Q indicating a more effective clustering result. The average silhouette value S gauges the homogeneity of clusters, with higher values indicating greater credibility. In this context, Figure 3.9 reveals a Q of 0.54227, signifying a well-structured network with a favorable clustering effect. The associated S value of 0.7154 underscores the high homogeneity of clusters, showcasing a clear distinction between different clusters. This figure showcases ten clusters, spearheaded by "Chronic Illness," "self-management," and "nursing home." That said, the primary clusters have an average inception around 2014-2016, signifying a period of maturity in related studies. The largest cluster, "Chronic Illness," with an initiation year of 2013, comprises 50 keywords, with key terms such as nursing, chronic disease, qualitative research, and caregivers, among others. The main keywords for each cluster are summarized in Table 3.5.

3.8. Keywords time zone analysis. To further explore the development and evolution of research over time, the analysis of keywords within the papers from the WOS literature big database was carried out using the time zone map feature in CiteSpace. As illustrated in Figure 3.8, the size of each node signifies the

Table 3.5: Keywords time zone analysis

Ranking	Clustering name	Numbers	S	Year	Main keywords
0	Chronic illness	50	0. 673	2016	Chronic illness (171.43, 1.0E-4); nursing (156. 72, 1. 0E-4); chronic disease (141. 72, 1. 0E-4); qualitative research (97. 66, 1. 0E-4); caregivers (87. 04, 1. 0E-4)
1	Self-management	46	0. 699	2015	Self-management (144. 95, 1. 0E-4); self-care (113. 15, 1. 0E-4); self-efficacy (99. 87, 1. 0E-4); patient education (67. 18, 1. 0E-4); chronic heart failure (61. 24, 1. 0E-4)
2	Nursing home	39	0. 674	2014	Nursing home (129. 62, 1. 0E-4); dementia (128. 91, 1. 0E-4); elderly (105. 91, 1. 0E-4); nursing homes (103. 84, 1. 0E-4); long-term care (102. 59, 1. 0E-4)
3	Chronic kidney disease	33	0. 704	2016	Chronic kidney disease (264. 35, 1. 0E-4); hemodialysis (139. 24, 1. 0E-4); dialysis (89. 86, 1. 0E-4); peritoneal dialysis (85. 15, 1. 0E-4); education (82. 7, 1. 0E-4)
4	Depression	33	0. 663	2016	Depression (227.75, 1.0E-4); anxiety (154. 97, 1. 0E-4); quality of life (95. 92, 1. 0E-4); physical activity (80. 33, 1. 0E-4); mental health (68. 92, 1. 0E-4)
5	Qualitive	27	0. 711	2016	Qualitative (60. 29, 1.0E-4); chronic (37. 67, 1. 0E-4); health care (27. 02, 1. 0E-4); implementation (24. 85, 1. 0E-4); end of life (21.17, 1. 0E-4)
6	Mortality	27	0. 758	2016	Mortality (90.25, 1.0E-4); risk factors (87. 43, 1. 0E-4); epidemiology (85. 53, 1. 0E-4); pulmonary rehabilitation (78. 92, 1. 0E-4); copd (74.92, 1.0E-4)
7	Primary care	26	0. 733	2016	Primary care (238.24, 1.0E-4); general practice (174. 01, 1. 0E-4); nurse practitioner (75. 35, 1. 0E-4); primary health care (72. 55, 1. 0E-4); nurse practitioners (67. 09, 1. 0E-4)
8	Telemedicine	21	0. 817	2015	Telemedicine (171. 13, 1. 0E-4); telehealth (132. 73, 1. 0E-4); mhealth (72. 38, 1. 0E-4); hypertension (56. 84, 1. 0E-4); ehealth (52. 95, 1. 0E-4)
9	COVID-19	15	0. 824	2020	COVID-19 (173. 91, 1. 0E-4); Sars-COV-2 (111. 21, 1. 0E-4); pandemic (55. 38, 1. 0E-4); healthcare workers (49. 17, 1. 0E-4); Coronavirus (49. 03, 1. 0E-4)

keyword's frequency, and the year assigned to each node marks the keyword's initial appearance. Connecting lines between nodes represent instances where different keywords appear in the same articles simultaneously, revealing relationships of inheritance and evolution across different periods. This approach not only helps identify primary areas of research focus during hot periods but also elucidates the stages of development within the field. As seen in Figure 3.10, the most significant node is "nursing," introduced in 2013. In the early studies, high-frequency keywords included chronic disease, primary health care, chronic obstructive pulmonary disease, and elderly care, among others. These related concepts have spanned a significant timeframe and had a substantial impact. Research has continued to this day, introducing fresh concepts such as burnout, kidney transplantation, and critical illness in recent studies.

3.9. Timeline of keywords analysis. A two-dimensional timeline, referred to as the "Timeline graph," was utilized to display literature keyword clustering, providing researchers with insights into the evolving processes and cutting-edge trends within topic clusters. The Timezone function in CiteSpace was used to analyze the keywords, shedding light on the development and evolution of research within the WOS literature big database. The size of each node in Figure 3.11 reflects the frequency of the keyword, while the year associated with each node indicates when the keyword was first used. The connecting lines between nodes reveal instances where different keywords appeared together in the same articles, representing relationships of inheritance and

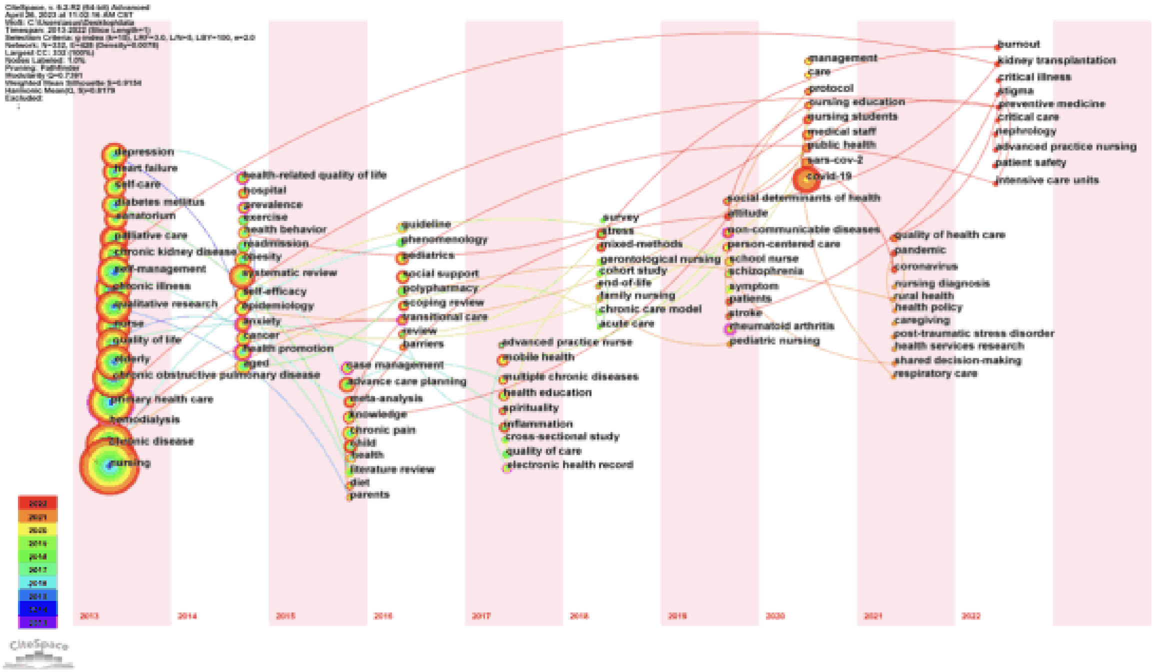


Fig. 3.10: Time zone diagram of keywords

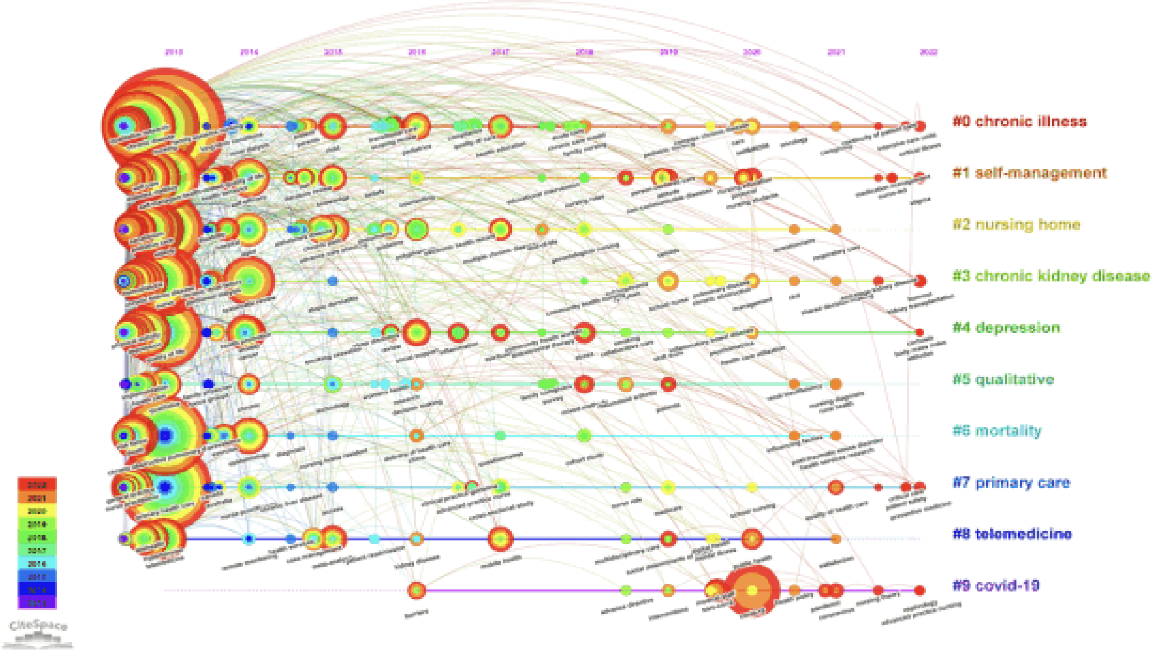


Fig. 3.11: Timeline diagram of keywords

Table 3.6: Keywords burstness

Keywords	**Year**	**Strength**	**Begin**	**End**	**2013–2022**
Randomized controlled trial	2013	5. 6	**2013**	2017	🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦
Chronic disease management	2013	5. 31	**2013**	2016	🟥🟥🟥🟥🟦🟦🟦🟦🟦🟦
Malnutrition	2013	3. 92	**2013**	2014	🟥🟥🟦🟦🟦🟦🟦🟦🟦🟦
Hepatitis c	2013	3. 9	**2013**	2015	🟥🟥🟥🟦🟦🟦🟦🟦🟦🟦
Disease management	2013	3. 88	**2013**	2015	🟥🟥🟥🟦🟦🟦🟦🟦🟦🟦
Grounded theory	2013	3. 36	**2013**	2014	🟥🟥🟦🟦🟦🟦🟦🟦🟦🟦
Screening	2013	3. 26	**2013**	2017	🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦
Motivational interviewing	2013	2. 84	**2013**	2016	🟥🟥🟥🟥🟦🟦🟦🟦🟦🟦
Rehabilitation	2013	2. 7	**2013**	2014	🟥🟥🟦🟦🟦🟦🟦🟦🟦🟦
General practice	2013	2. 66	**2013**	2015	🟥🟥🟥🟦🟦🟦🟦🟦🟦🟦
Evidence-based practice	2013	2. 54	**2013**	2017	🟥🟥🟥🟥🟥🟦🟦🟦🟦🟦
Empowerment	2013	2. 43	**2013**	2016	🟥🟥🟥🟥🟦🟦🟦🟦🟦🟦
Frail elderly	2014	6. 11	**2014**	2016	⬜🟥🟥🟥🟦🟦🟦🟦🟦🟦
Long-term conditions	2014	5. 52	**2014**	2015	⬜🟥🟥🟦🟦🟦🟦🟦🟦🟦
Chronic illness	2013	4. 85	**2014**	2015	🟦🟥🟥🟦🟦🟦🟦🟦🟦🟦
Nurse practitioner	2013	4. 58	**2014**	2017	🟦🟥🟥🟥🟥🟦🟦🟦🟦🟦
Readmission	2014	4. 35	**2014**	2018	⬜🟥🟥🟥🟥🟥🟦🟦🟦🟦
Prevalence	2014	4. 16	**2014**	2016	⬜🟥🟥🟥🟦🟦🟦🟦🟦🟦
Content analysis	2014	3. 76	**2014**	2016	⬜🟥🟥🟥🟦🟦🟦🟦🟦🟦
Remote monitoring	2014	3. 76	**2014**	2016	⬜🟥🟥🟥🟦🟦🟦🟦🟦🟦
Multiple sclerosis	2014	3. 56	**2014**	2015	⬜🟥🟥🟦🟦🟦🟦🟦🟦🟦
Nurse practitioners	2014	3. 37	**2014**	2017	⬜🟥🟥🟥🟥🟦🟦🟦🟦🟦
Canada	2014	3. 31	**2014**	2015	⬜🟥🟥🟦🟦🟦🟦🟦🟦🟦
Australia	2014	2. 75	**2014**	2015	⬜🟥🟥🟦🟦🟦🟦🟦🟦🟦

evolution over time. This approach combines the number of publications over the years, providing insights into primary research areas during hot periods and illustrating the field's developmental stages. Figure 3.11 reveals that the most prominent node is "nursing," introduced in 2013. Early studies featured high-frequency keywords such as chronic disease, primary health care, chronic obstructive pulmonary disease, and elderly care, among others. These concepts have exerted significant and lasting influence. Recent studies have introduced novel concepts like burnout, kidney transplantation, and critical illness, signifying new directions in the field.

Additionally, the Timeline graph, featured in Figure 3.11, presents the clustering of literature keywords on a two-dimensional timeline. This provides researchers with valuable insights into the evolution and cutting-edge trends of specific topic clusters and the mutual relationships between these hot topics. The graph displays different color-coded clusters, each representing a set of important keywords within the same cluster. The top 10 clusters included #0 chronic illness, #1 self-management, #2 nursing home, #3 chronic kidney disease, #4 depression, #5 qualitative, #6 mortality, #7 primary care, #8 telemedicine, and #9 COVID-19. Figure 3.11 shows that the largest cluster in related literature was "chronic illness," consisting of 50 keywords, with an average year of 2016. Key terms in this cluster included nursing, chronic disease, qualitative research, caregivers, and more. Over time, additional keywords such as medication management and intensive care units made their appearance. The cluster report generated by the system highlighted Facchinetti, G. (2019) and their study titled "Discharge of Older Patients with Chronic Diseases: What Nurses Do and What They Record. An observational study."

3.10. Keywords burstness. The table in section 3.10 (Table 3.6) reveals emergent keywords in the research field over the past decade. The "beginning year" indicates when a particular keyword's frequency began to surge, and the "end year" represents when that keyword's frequency started stabilizing. The intensity of emergence reflects the degree of a sudden increase in a keyword's frequency during its emergence period, often correlated with its research popularity. Keywords with red bars signify their relevance in specific durations.

Keywords	**Year**	**Strength**	**Begin**	**End**	**2013–2022**
Self-management support	2014	2. 71	**2014**	2016	─🟥🟥🟥▬▬▬▬▬▬
Cognitive impairment	2014	2. 67	**2014**	2017	─🟥🟥🟥🟥▬▬▬▬▬
Coping	2014	2. 41	**2014**	2016	─🟥🟥🟥▬▬▬▬▬▬
Humans	2013	6. 95	**2015**	2017	▬▬🟥🟥🟥▬▬▬▬▬
Technology	2015	4. 78	**2015**	2016	──🟥🟥▬▬▬▬▬▬
Technology	2015	4. 78	**2015**	2016	──🟥🟥▬▬▬▬▬▬
Emergency department	2015	4. 62	**2015**	2016	──🟥🟥▬▬▬▬▬▬
Concept analysis	2015	4. 06	**2015**	2017	──🟥🟥🟥▬▬▬▬▬
Alzheimers disease	2015	3. 87	**2015**	2019	──🟥🟥🟥🟥🟥▬▬▬
Health	2015	3. 7	**2015**	2016	──🟥🟥▬▬▬▬▬▬
Health promotion	2014	3. 22	**2015**	2016	─▬🟥🟥▬▬▬▬▬▬
Care management	2015	3. 16	**2015**	2017	──🟥🟥🟥▬▬▬▬▬
Phenomenology	2016	7. 65	**2016**	2018	───🟥🟥🟥▬▬▬▬
Polypharmacy	2016	6. 35	**2016**	2019	───🟥🟥🟥🟥▬▬▬
Exercise	2014	5. 96	**2016**	2018	─▬▬🟥🟥🟥▬▬▬▬
Randomized controlled trials as topic	2016	5. 76	**2016**	2017	───🟥🟥▬▬▬▬▬
Pneumonia	2016	3. 82	**2016**	2018	───🟥🟥🟥▬▬▬▬
Review	2016	3. 18	**2016**	2017	───🟥🟥▬▬▬▬▬
Guideline	2016	2. 82	**2016**	2017	───🟥🟥▬▬▬▬▬
Adult	2013	2. 68	**2016**	2019	▬▬▬🟥🟥🟥🟥▬▬▬
Quality of care	2017	6. 08	**2017**	2018	────🟥🟥▬▬▬▬
Cross-sectional study	2017	4. 32	**2017**	2018	────🟥🟥▬▬▬▬
End-stage renal disease	2013	4. 17	**2017**	2019	▬▬▬▬🟥🟥🟥▬▬▬
Inflammation	2017	3. 95	**2017**	2018	────🟥🟥▬▬▬▬
Health education	2017	3. 67	**2017**	2019	────🟥🟥🟥▬▬▬
Length of stay	2017	3. 17	**2017**	2019	────🟥🟥🟥▬▬▬
Literature review	2015	3. 09	**2017**	2020	──▬▬🟥🟥🟥🟥▬▬
End of life	2013	3. 01	**2017**	2018	▬▬▬▬🟥🟥▬▬▬▬
Compliance	2017	2. 77	**2017**	2019	────🟥🟥🟥▬▬▬
Electronic health record	2017	2. 66	**2017**	2020	────🟥🟥🟥🟥▬▬
Cohort study	2018	4. 36	**2018**	2019	─────🟥🟥▬▬▬
Gerontological nursing	2018	4. 26	**2018**	2020	─────🟥🟥🟥▬▬
Chronic care model	2018	3. 39	**2018**	2019	─────🟥🟥▬▬▬
Stress	2018	3. 24	**2018**	2019	─────🟥🟥▬▬▬
Acute care	2018	2. 91	**2018**	2019	─────🟥🟥▬▬▬
Treatment	2014	2. 9	**2018**	2019	─▬▬▬▬🟥🟥▬▬▬
Outcomes	2013	2. 9	**2018**	2019	▬▬▬▬▬🟥🟥▬▬▬
Symptom	2019	4. 99	**2019**	2020	──────🟥🟥▬▬
School nurse	2019	4. 85	**2019**	2022	──────🟥🟥🟥🟥
Person-centered care	2019	4. 49	**2019**	2020	──────🟥🟥▬▬
Medicare	2019	4. 08	**2019**	2020	──────🟥🟥▬▬
Asthma	2013	3. 4	**2019**	2020	▬▬▬▬▬▬🟥🟥▬▬
Rheumatoid arthritis	2019	3. 03	**2019**	2022	──────🟥🟥🟥🟥

The table encompasses 30 emergent keywords, and by considering their commencement times, “randomized controlled trial,” “chronic disease management,” and “malnutrition” emerge as early research focal points. From a duration perspective, keywords like “randomized controlled trial,” “screening,” “evidence-based practice,” and “readmission” have maintained their relevance over an extended period, indicating their prolonged status as research hotspots. Regarding the strength of emergent keywords, “COVID-19” (Strength = 35.66), “phenomenology” (Strength = 7.65), “Sars-COV-2” (Strength = 6.98), and “humans” (Strength = 6.95) exhibit substantial sudden intensity, signifying significant changes in their frequency of occurrence. In summary, “COVID-19,” “Sars-COV-2,” “public health,” and “medical staff” not only boast high emergence intensity but

Keywords	Year	Strength	Begin	End	2013–2022
Pediatric nursing	2019	3. 03	**2019**	2022	⬜⬜⬜⬜⬜⬜🟥🟥🟥🟥
Counselling	2016	2. 91	**2019**	2020	⬜⬜⬜🟦🟦🟦🟥🟥🟦🟦
Dyspnea	2014	2. 91	**2019**	2020	⬜🟦🟦🟦🟦🟦🟥🟥🟦🟦
Assessment	2013	2. 56	**2019**	2020	🟦🟦🟦🟦🟦🟦🟥🟥🟦🟦
Mobile health	2017	2. 45	**2019**	2022	⬜⬜⬜⬜🟦🟦🟥🟥🟥🟥
Interventions	2019	2. 42	**2019**	2022	⬜⬜⬜⬜⬜⬜🟥🟥🟥🟥
COVID-19	2020	35. 66	**2020**	2022	⬜⬜⬜⬜⬜⬜⬜🟥🟥🟥
Sars-COV-2	2020	6. 98	**2020**	2022	⬜⬜⬜⬜⬜⬜⬜🟥🟥🟥
Public health	2020	6. 14	**2020**	2022	⬜⬜⬜⬜⬜⬜⬜🟥🟥🟥
Medical staff	2020	5. 3	**2020**	2022	⬜⬜⬜⬜⬜⬜⬜🟥🟥🟥
Nursing students	2020	5. 3	**2020**	2022	⬜⬜⬜⬜⬜⬜⬜🟥🟥🟥
Management	2020	4. 96	**2020**	2022	⬜⬜⬜⬜⬜⬜⬜🟥🟥🟥
Protocol	2020	4. 74	**2020**	2022	⬜⬜⬜⬜⬜⬜⬜🟥🟥🟥
Nursing education	2020	4. 74	**2020**	2022	⬜⬜⬜⬜⬜⬜⬜🟥🟥🟥
Care	2020	4. 13	**2020**	2022	⬜⬜⬜⬜⬜⬜⬜🟥🟥🟥
Pregnancy	2020	3. 72	**2020**	2022	⬜⬜⬜⬜⬜⬜⬜🟥🟥🟥
Health care utilization	2020	3. 31	**2020**	2022	⬜⬜⬜⬜⬜⬜⬜🟥🟥🟥
Transitional care	2016	3. 13	**2020**	2022	⬜⬜⬜🟦🟦🟦🟦🟥🟥🟥
Patient	2015	3. 01	**2020**	2022	⬜⬜🟦🟦🟦🟦🟦🟥🟥🟥
Fatigue	2013	2. 67	**2020**	2022	🟦🟦🟦🟦🟦🟦🟦🟥🟥🟥

also closely align with the current timeline, suggesting that they represent the latest emerging research hotspots.

In general, as time progresses and society evolves, along with shifts in the external environment, the research content and hotspots within Chronic disease nursing continue to change. This dynamic landscape underscores the enduring research value of Chronic disease nursing from a different perspective.

3.11. Literature co-citation analysis. The analysis of literature co-citation serves to identify the interconnections between co-cited works within a specific research field, shedding light on influential literature that has a substantial impact on both the field itself and related disciplines. The quantity of co-citations directly correlates with the strength of associations between works and the significance of high-level literature. Illustrated in Figure 3.12, this co-citation network features 346 nodes, 455 connections, and a network density of 0.0076, highlighting several prominent co-citation relationships. Notably, the works by Braun V. (2006), Tong A. (2007), and Wagner E.H. (2001) emerge as key figures, boasting relatively high citation frequencies. For instance, Tong, A.; Sainsbury, P.; Craig, J.'s extensive search of various sources for existing checklists used to assess qualitative studies, including systematic reviews and major medical journals, has garnered a remarkable $15,637$ citations [15]. Braun, V.; and Clarke, V., known for their outline of thematic analysis and its relation to other qualitative analytic methods, have accumulated $4,406$ citations [3]. Wagner E.H., whose work focuses on the challenges faced by those with chronic illnesses in accessing appropriate medical care, particularly in systems designed for acute illnesses, has amassed $4,127$ citations [18].

Furthermore, Bodenheimer T., in their exploration of the potential of the computer revolution in improving primary care, particularly through systems aimed at enhancing physician performance and patient outcomes, has garnered 243 citations [1]. Meanwhile, Lorig, K.R.; Holman, H.R.'s work on self-management tasks and skills has accumulated $2,296$ citations, emphasizing its profound influence within the field [11]. In addition, the centrality of literature co-citation identifies works like Barlow J. (2002), Higgins J.P.T. (2003), and Bodenheimer T. (2002) as frequently cited classical references within the literature. A comprehensive list of the top ten cited papers is available in Table 3.7.

4. Conclusions. To gain insights into research hotspots and development trends within the domain of chronic disease nursing, this study employed visual analysis techniques on literature within the extensive WOS literature database, examining parameters such as the annual number of published papers, geographical and institutional distributions, authorship, high-frequency keywords, keyword clustering, keyword time zone analysis, keyword burstness, and co-citation relationships. Notably, several papers related to chronic disease nursing ex-

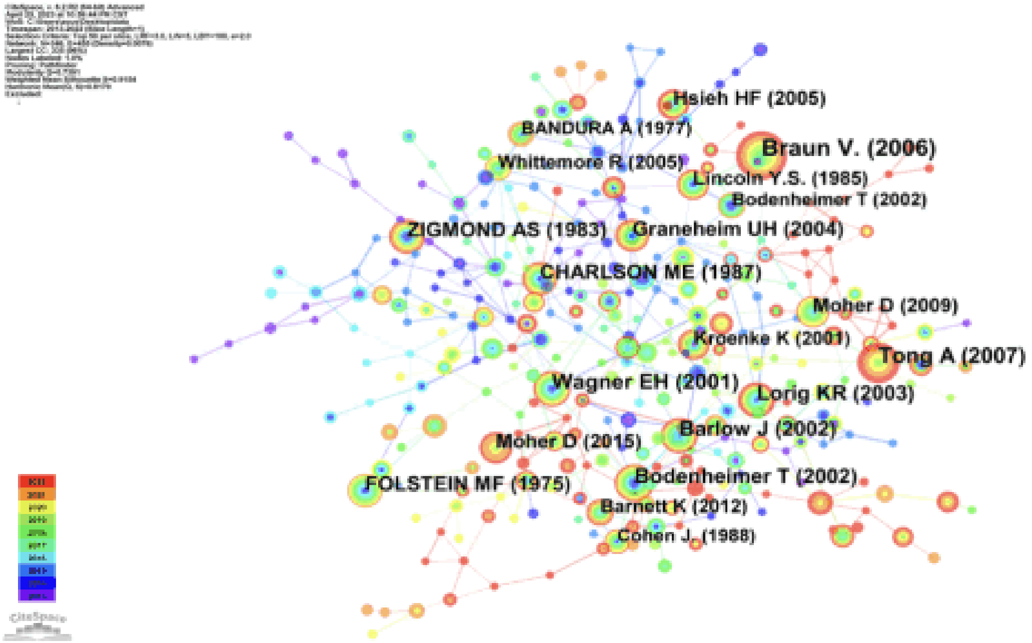


Fig. 3.12: Literature co-citation relationship

Table 3.7: Ten top of cited papers

Ranking	Frequency	Author	year	Journals Resource
1	170	Braun V.	2006	Qualitative Res Psyc
2	122	Tong A	2007	Int J Qual Health C
3	108	Wagner EH	2001	Health Affair
4	99	Bodenheimer T	2002	Jama-J AM Med Assoc
5	98	Lorig KR	2003	Ann Behav Med
6	93	Barlow J	2002	Patient Educ Couns
7	90	Charlson ME	1987	J Chron Dis
8	86	Folstein MF	1975	J Psychiat Res
9	85	Zigmond AS	1983	Acta Psychiat Scand
10	81	Graneheim UH	2004	Nurs EducC Today

hibited a steady increase from 2013 to 2021, signaling sustained global scholarly interest. Within this research, 141 journals published a significant number of papers on the application of artificial intelligence technology in nursing, totaling 3, 610 published papers, constituting 58.41% of all publications. The top three journals with the highest publication volume were “Journal of Clinical Nursing,” “Journal of Advanced Nursing,” and “BMC Health Services Research.” Among the authors, twelve individuals authored ten or more articles, with Hu, Frank B., Willett, Walter C., and Rimm, Eric B. leading the way, collectively contributing to 110 papers or 7.7% of the total. Their prolific output suggests a strong cooperative network among key researchers in this field.

Additionally, the research uncovered the top ten research institutions by publication frequency, including

Harvard University, Harvard Medical School, Brigham & Women's Hospital, and others, amounting to 19.29% of publications. Harvard University, Harvard Medical School, and Brigham & Women's Hospital featured prominently, indicating a closely-knit cooperation network among these core institutions. Furthermore, the analysis of countries or regions revealed the USA, Australia, and England as the leading contributors, publishing $5,491$ papers and accounting for 46.28% of the total. The centrality of these publications correlated positively with their quantity, while countries such as China and Canada exhibited disproportionality between their publication volumes and centrality, suggesting room for improved international collaboration. Similarly, cooperation degrees in the USA, Australia, and England stood at $0.16, 0.13$, and 0.23 respectively. Interestingly, certain papers such as those authored by Braun V. (2006), Tong A. (2007), Wagner E.H. (2001), Bodenheimer T. (2002), Lorig K.R. (2003), Barlow J. (2002), Charlson M.E. (1987), Folstein M.F. (1975), Zigmond A.S. (1983), Graneheim U.H. (2004) garnered widespread citations, underscoring their substantial impact on the field. Current research hotspots encompass incidence, rheumatoid arthritis, and qualitative aspects, as well as emerging areas like burnout, kidney transplantation, critical illness, COVID-19, Sars-COV-2, public health, and medical staff. These findings reveal the evolving nature of research in chronic disease nursing, underscoring its enduring research value in a dynamically changing landscape. Data supporting this study's findings are available from the corresponding author upon request. The authors declare no known financial or personal conflicts of interest that could have influenced the reported work.

In conclusion, this study has provided a comprehensive analysis of the research landscape within the field of chronic disease nursing, revealing key trends and hotspots. The increasing number of papers published over the years underscores the enduring relevance of this field to scholars across the globe. The substantial prevalence of publications related to artificial intelligence applications in nursing further highlights the field's dynamism. Prolific authors and core institutions with high cooperation degrees exemplify a closely-knit scholarly community actively contributing to this domain. Additionally, influential papers with widespread citations illustrate the pivotal role of certain research in shaping the discourse. As research evolves, new emerging hotspots, such as burnout, kidney transplantation, and COVID-19, indicate the field's responsiveness to evolving societal and environmental changes. This study underscores the continued value of chronic disease nursing research and its ability to adapt to the evolving healthcare landscape.

Based on the findings of this study, it is recommended that healthcare professionals, researchers, and policymakers continue to invest in and support research in the field of chronic disease nursing, with a particular focus on emerging areas such as burnout, kidney transplantation, and COVID-19. This will enable a more comprehensive understanding of the evolving healthcare landscape and help address the challenges associated with chronic disease management. One limitation of this study is that it primarily relies on data from the WoS literature database, which may not encompass all relevant research. Additionally, the analysis is based on quantitative metrics and may benefit from qualitative insights and interdisciplinary perspectives to provide a more holistic understanding of the field. Future research could explore the qualitative aspects of chronic disease nursing, including the experiences of patients, healthcare providers, and caregivers. Lastly, interdisciplinary collaborations and mixed-methods approaches could provide a richer understanding of the field's dynamics and the impact of healthcare policies on chronic disease management.

Data availability. The data used to support the findings of this study are available from the corresponding author upon request.

Acknowledgment. The authors thank TopEdit (www.topeditsci.com) for its linguistic assistance during the preparation of this manuscript.

REFERENCES

[1] T. Bodenheimer and K. Grumbach, *Electronic technology: a spark to revitalize primary care?*, Journal of the American Medical Association, 290 (2003), pp. 259–264.

[2] J.-J. Boté-Vericad, M.-L. Ferrer-Mejía, M. Gorchs-Molist, and J. Begazo-Corahua, *Gender differences in peruvian nursing: a bibliometric analysis in scopus and web of science*, Telos: Revista de Estudios Interdisciplinarios en Ciencias Sociales, 24 (2022), pp. 302–328.

[3] V. Braun and V. Clarke, *Using thematic analysis in psychology*, Qualitative Research in Psychology, 3 (2006), pp. 77–101.

[4] R. Cant, C. Ryan, and S. Kardong-Edgren, *Virtual simulation studies in nursing education: A bibliometric analysis of the top 100 cited studies, 2021*, Nurse Education Today, 114 (2022), p. 105385.

[5] C. Chen, *Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature*, Journal of the American Society for information Science and Technology, 57 (2006), pp. 359–377.

[6] D. G. de Oliveira, A. da Cunha Reis, I. de Melo Franco, and A. L. Braga, *Exploring global research trends in burnout among nursing professionals: A bibliometric analysis*, in Healthcare (Basel), vol. 9, MDPI, 2021, p. 1680.

[7] A. Ghamgosar, M. Zarghani, and L. Nemati-Anaraki, *Bibliometric analysis on geriatric nursing research in web of science (1900–2020)*, BioMed Research International, 2021 (2021), p. 8758161.

[8] L. Guo, G. Lu, and J. Tian, *A bibliometric analysis of cirrhosis nursing research on web of science*, Gastroenterology Nursing, 43 (2020), pp. 232–240.

[9] S. Hahn and Y. M. Ryu, *Trends in research on clinical reasoning in nursing over the past 20 years: a bibliometric analysis*, Science Editing, 9 (2022), pp. 112–119.

[10] Q. Huang, Q. Ronghuang, R. Yinhuang, Y. Fanghuang, and H. Yansun, *Trends and hotspots of family nursing research based on web of science: A bibliometric analysis*, Japan Journal of Nursing Science, 18 (2021), p. e12401.

[11] K. R. Lorig and H. R. Holman, *Self-management education: history, definition, outcomes, and mechanisms*, Annals of Behavioral Medicine, 26 (2003), pp. 1–7.

[12] A. Molassiotis, C. Guo, H. Abu-Odah, C. West, and A. Y. Loke, *Evolution of disaster nursing research in the past 30 years (1990–2019): a bibliometric and mapping analysis*, International Journal of Disaster Risk Reduction, 58 (2021), p. 102230.

[13] W.-S. Su, G.-J. Hwang, and C.-Y. Chang, *Bibliometric analysis of core competencies associated nursing management publications*, Journal of Nursing Management, 30 (2022), pp. 2869–2880.

[14] Y. Teng, Y. Huang, and S. Yang, *Applying knowledge graph to analyze the historical landscape based on citespace*, Wireless Communications and Mobile Computing, 2022 (2022), p. 3867541.

[15] A. Tong, P. Sainsbury, and J. Craig, *Consolidated criteria for reporting qualitative research (coreq): a 32-item checklist for interviews and focus groups*, International Journal for Quality in Health Care, 19 (2007), pp. 349–357.

[16] M. Viswanathan, C. E. Golin, C. D. Jones, M. Ashok, S. J. Blalock, R. C. Wines, E. J. Coker-Schwimmer, D. L. Rosen, P. Sista, and K. N. Lohr, *Interventions to improve adherence to self-administered medications for chronic diseases in the united states: a systematic review*, Annals of Internal Medicine, 157 (2012), pp. 785–795.

[17] H. B. Vošner, H. Carter-Templeton, J. Završnik, and P. Kokol, *Nursing informatics: a historical bibliometric analysis*, CIN: Computers, Informatics, Nursing, 38 (2020), pp. 331–337.

[18] E. H. Wagner, B. T. Austin, C. Davis, M. Hindmarsh, J. Schaefer, and A. Bonomi, *Improving chronic illness care: translating evidence into action*, Health Affairs, 20 (2001), pp. 64–78.

[19] J. Wang, Y. Chen, X. Zhai, Y. Chu, X. Liu, and X. Ma, *Visualizing research trends and identifying hotspots of traditional chinese medicine (tcm) nursing technology for insomnia: A 18-years bibliometric analysis of web of science core collection*, Frontiers in Neurology, 13 (2022), p. 816031.

[20] A. WHO, *World health statistics 2016: monitoring health for the sdgs sustainable development goals*, World Health Organization, (2021).

[21] H. Yesilbas and F. Kantek, *Trends and hot topics in nurse empowerment research: A bibliometric analysis*, Japan Journal of Nursing Science, 19 (2022), p. e12458.

[22] Q. Zhang, S. Li, J. Liu, and J. Chen, *Global trends in nursing-related research on covid-19: a bibliometric analysis*, Frontiers in Public Health, 10 (2022), p. 933555.

[23] J. Zhao, Y. Lu, F. Zhou, R. Mao, and F. Fei, *Systematic bibliometric analysis of research hotspots and trends on the application of virtual reality in nursing*, Frontiers in Public Health, 10 (2022), p. 906715.

[24] Y. Zhu, M. C. Kim, and C. Chen, *An investigation of the intellectual structure of opinion mining research.*, Information Research: An International Electronic Journal, 22 (2017), p. 739.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Nov 4, 2023
Accepted: Jan 8, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1748–1756, DOI 10.12694/scpe.v25i3.2678

REMOTE INTELLIGENT MEDICAL MONITORING DATA TRANSMISSION NETWORK OPTIMIZATION BASED ON DEEP LEARNING

RUN WANG*

Abstract. A hospital operating status evaluation data analysis system was established based on the autoencoder's network. The Gibbs sampling method is used to obtain the approximate distribution of RBM. In addition, the Autoencoder neural network can also select feature dimensions that can better characterize the characteristics of financial operation data from a large amount of financial operation data. Deep learning methods are used to study the redundant information elimination method and the generation mechanism of multi-source heterogeneity in multi-source heterogeneous networks. The principle of intrinsic compression is used to reduce the dimensionality of the redundancy in the network and obtain the compression redundancy objective function. This article sets thresholds for information classification on the Internet. The approach was tested using financial data from a medical institution. Use smart encoders to extract 17 financial indicators from financial data that can be used for modeling. The evaluation results are used as the output vector of the model. Comparative experiments show that the AUC value and accuracy of the method proposed in this article can be improved by 0.84 and 83.33% compared with the AUC value of shallow logistic regression and BP neural network. This algorithm has apparent improvements.

Key words: Deep learning; DBN; RBM; Autoencoder; Network data; Redundant information; Optimization elimination; Medical data mining

1. Introduction. With the advent of the fourth industrial revolution, information on the Internet has also exploded. With the rapid development of big data, medical financial research will also face various challenges. As the scale of data continues to expand, higher requirements are placed on data processing technology. Rapid identification and processing of large amounts of data can efficiently extract useful information from these data. It can improve the efficiency of the financial system operation of medical institutions and provide a robust data basis for the operation and management of medical institutions. Computer processing capabilities have improved dramatically in recent years. The computing speed of classic machine learning methods can no longer meet the needs of practical applications. Due to their better computing power, deep neural networks are increasingly used in industrial fields. Due to the low degree of data structure, high feature dimension, and missing data of current medical financial data, it is challenging to directly apply deep neural networks to medical financial data. Literature [1] proposes a method for eliminating redundant information in network data transmission. This approach is based on grouping data in the network. A dynamic analysis of bimodality and packet characteristics in the network is performed. An algorithm based on sliding windows is proposed to realize the positioning of data grouping boundary points. In network data transmission, the data that is transmitted repeatedly is encoded. However, this method can easily cause the system's operating efficiency to decrease. Literature [2] provides a method to eliminate redundant information based on dynamic lookup tables during network data transmission. In network data transmission, the first-byte value of a data block with a high redundancy rate is selected as a mark. Update dynamic query tables promptly during network data transmission. Select data blocks based on tags in the lookup table. Blocks of redundant information are encoded in the network data that has been sent. Replace redundant information fragments in the original network data transmission with encoded data. This algorithm can effectively improve the average of the data, but it will cause data loss. This paper builds an intelligent analysis and identification system for medical imaging data. The system's data analysis capabilities were tested and analyzed using a hospital's relevant financial data set as an example.

*The Fifth Affiliated Hospital of Zhengzhou University, Zhengzhou 450052, China (Corresponding author, zzdxwfy123@126.com)

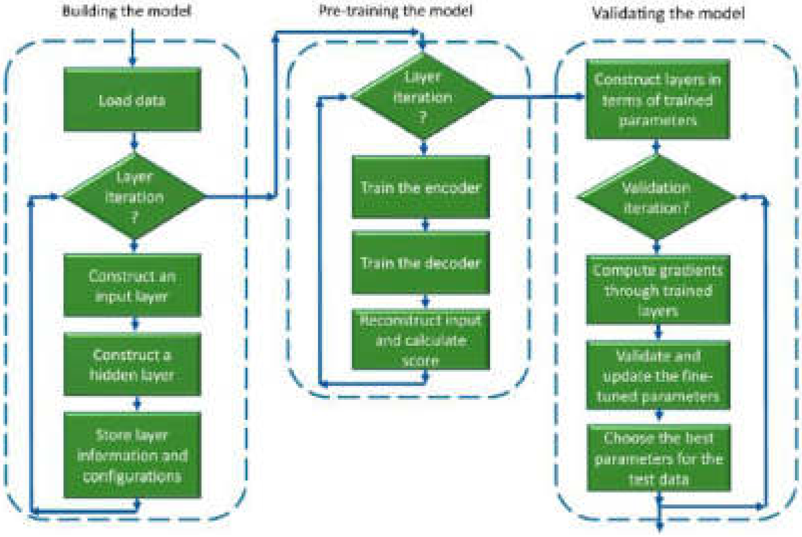

Building the model
Load data
Layer iteration ?
Construct an input layer
Construct a hidden layer
Store layer information and configurations
Pre-training the model
Layer iteration ?
Train the encoder
Train the decoder
Reconstruct input and calculate score
Validating the model
Construct layers in terms of trained parameters
Validation iteration?
Compute gradients through trained layers
Validate and update the fine-tuned parameters
Choose the best parameters for the test data

Fig. 2.1: Autoencoder method flow.

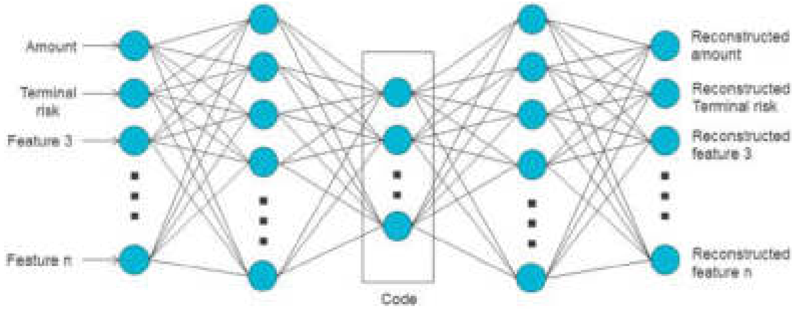


Fig. 2.2: Network architecture of autoencoder.

2. Autoencoder network feature extraction method. This paper uses an automatic encoding machine network for feature extraction. The network extraction process is given in Figure 2.1 (the picture is quoted from A deep learning method for lincRNA detection using auto-encoder algorithm).

Autoencoders can be thought of as unsupervised learning networks. It translates input data into passwords. Finally, the signal is decoded to obtain the final output decoding result [3]. Then, the encoding and decoding parameters are continuously adjusted according to the deviation between the input and the output, and finally, the required output is obtained. Its network structure is shown in Figure 2.2.

Determines the input quantity U with dimension m and the output quantity Y with dimension n. Determine the excitation functions h and f.

$$\begin{aligned} \rho = h(u) = s_h(\kappa u + \alpha) \\ v = f(u) = s_f(\bar{\kappa} u + \chi) \end{aligned} \tag{2.1}$$

where κ is the weight input to the hidden layer. α is the compensation matrix output to the hidden layer. $\bar{\kappa}$ is the weight of the hidden layer and output layer. $\chi$is the bias matrix of the hidden layer and output layer. The activation functions s_h and s_f used in this article are both Sigmoid functions [4]. Ideally, the output layer of Y should be a reproduction of the U data. So the relationship of κ is expressed as follows (2.2):

$$\bar{\kappa} = \kappa^T \tag{2.2}$$

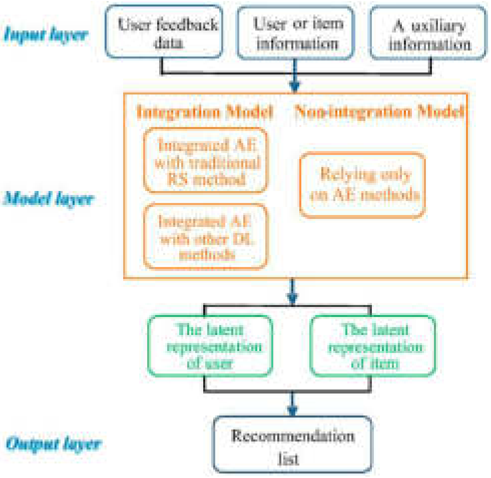

Input layer
User feedback data
User or item information
A uxiliary information
Integration Model
Non-integration Model
Integrated AE with traditional RS method
Integrated AE with other DL methods
Relying only on AE methods
Model layer
The latent representation of user
The latent representation of item
Output layer
Recommendation list

Fig. 2.3: System algorithm flow chart.

To reduce the deviation between input and output, the error distance $\Omega(u, v)$ must be determined. Determine$\Omega(u, v)$ as follows when using the Sigmund function.

$$\Omega(u, v) = -\sum_{i=1}^{n} [u_i \log v_i + (1 - u_i) \log(1 - v_i)] \tag{2.3}$$

The loss function can be determined in terms of $\Omega(u, v)$ in the automatic encoding process. If the training set has the following formula (2.4):

$$D = \{U^1, U^2, \cdots, U^N\} \tag{2.4}$$

Then, the loss function is expressed as (2.5).

$$l(\beta) = \sum \Omega(u, f(h(u))) \tag{2.5}$$

This paper uses the hidden layer of the self-organizing network obtained based on the gradient method. This method can be used as input to a deep learning network [5]. This article combines the DBN and Autoencoder self-encoding networks to construct the system algorithm flow shown in Figure 2.3 (picture cited in Mathematics 2023, 11(8), 1777). First, the required relevant medical data needs to be reprocessed. The Autoencoder method is used to realize automatic extraction of data. The extracted features are used for the training of deep neural networks. Finally, experiments were conducted on the established neural network, and the experimental results were verified.

3. Application of optimization removal principle in redundancy in network data transmission. A random sampling method based on wavelet transform is proposed [6]. It reflects the sample point distribution u_i and its uncertainty λ_i during the data transmission process in the network. Q^+, Q^- represents the average of the positive and negative values of the data sampling set D during network data transmission. $\beta^T = Q^+ - Q^-$ is the average vector of positive and negative values in the data sampling set during the network data transmission process, then the hyperplanes passing through the two categories of Q^+, Q^- can be expressed by formula (3.1)

$$\begin{cases} \beta^T(u_i - Q^+) = 0 \\ \beta^T(u_i - Q^-) = 0 \end{cases} \tag{3.1}$$

Use equation (3.2) to calculate the distance s_{i+}, s_{i-} between the sampling points in the positive and negative classes of data sampling in network data transmission and the hyperplane in the corresponding class.

$$\begin{cases} s_{i+} = \beta^T(u - Q^+)/||\beta|| \\ s_{i-} = \beta^T(u - Q^-)/||\beta|| \end{cases} \tag{3.2}$$

$S_+ = \max\{s_{i+}\}$ and $S_- = \max\{s_{i-}\}$ represent the farthest distance between the sample points in the positive class and the sample points in the corresponding class during the network data transmission process. β represents the standard vector element. Equation (3.3) can be used to determine the fuzzy coefficient

$$\lambda_i = \begin{cases} -1 + 2e^{-In2/(S_+ + \xi s_{i+})} \\ -1 + 2e^{-In2/(S_- + \xi s_{i-})} \end{cases} \tag{3.3}$$

ξ is the conversion factor. Because each feature of the data sample in the analyzed network data transmission contributes differently to the classification accuracy, it is necessary to use the feature validity h_t^iin the network data transmission to classify each feature of the sample [7] accurately. Quantify the correlation of rates. The data training sample set in the network data transmission under test is denoted as C. The total number of data samples in network data transmission is represented by $|D|$. Assume that there are m types of data sample sets in network data transmission $\sigma_i(i = 1, 2, \cdots, m)$, then $|\sigma_1| + \cdots + |\sigma_m| = |D|$. In any network data transmission, if the possibility that a sample is related to class σ_i is $P_i = |\sigma_i|/D$, then the information entropy of the data sample set D in network data transmission is:

$$Info(D) = -\sum_{i=1}^{m} P_i \log_2(P_i) \tag{3.4}$$

Assume that feature t_i in any network data transmission can divide the data training sample set D in network data transmission into $D = \{D_1, D_2, \cdots, D_q\}$. During network data transmission, D_i contains the number of samples $|D_i|$. Then, equation (3.5) can express the conditional entropy of t_i line classification in network data transmission.

$$Info_{t_i}(D) = \sum_{i=1}^{m} |D_i|/|D| Info(D) \tag{3.5}$$

Use the information of characteristic t_i in network data transmission to increase by $Gain(t_i)$ to represent the change in entropy:

$$Gain(t_i) = Info(D) - Info_{t_i}(D) \tag{3.6}$$

In the process of network data transmission, the feature vector $F = (Gain(t_1), Gain(t_2), \cdots, Gain(t_n))$ reflects the effectiveness of each feature. The feature validity in network data transmission can be defined by formula (3.7)

$$h_i' = (Gain(t_i)/\sum_{i=1}^{n} Gain(t_i))e^{Gain(t_i)} \tag{3.7}$$

$e^{Gain(t_i)}$is the conversion factor. On a specific network, when attribute t_i has a more considerable amount of information, it has a more significant impact on category $Gain(t_i)$. The feature validity matrix P in network data transmission describes the feature validity of n features of the sample.

$$P = (h_t^1 \cdots h_t^n) \tag{3.8}$$

h_i' represents the characteristic effect of the data characteristic when the i network data is transmitted. Among them, Z is the core function of FSVM. P is the $n\prime$ class characteristic performance matrix for network data transmission. The kernel function for the effectiveness of data characteristics in the network is

$$Z\prime(u_i, u_j) = Z(u_i^T P, u_j^T P) \tag{3.9}$$

$u_i^T P$ is the representation of the eigen normal vector matrix at the i transmission. $u_j^T P$ represents the eigen normal vector matrix at node j. This paper chooses the radial basic kernel function. where η is the characteristic concentration of the kernel function. Equation (3.10) can be used to express the characteristic effectiveness kernel function $Z\prime$ in network data transmission:

$$Z\prime(u_i, u_j) = \exp(-\eta||u_i^T P - u_j^T P||^2) = \exp\{-\eta(u_i - u_j)P \cdot h_t^i\} \tag{3.10}$$

Feature validity h_t^i is low during network data transmission, the $i$bit characteristics of data sampling have a lower impact on the performance kernel during network data transmission [8]. This shows that the intrinsically valid kernel function can avoid the influence of features with strong correlations or redundant features in some network data transmission processes.

4. Optimal method to remove redundant information in network data transmission.

4.1. Reconstruction based on phase space and feature extraction of network redundant information. i'^{th}represents the i' packet sent by the network. Its essence is to segment the packets transmitted by the block data chain in a static state. Data of various sizes can be obtained through this method. Map redundant data in the network into a high-dimensional space. t_0 represents the initial value processing of boundary characteristics in the network. t_f is the number of steps iteratively processing boundary features in the network in high dimensions. This article will reconstruct the phase space of the redundant information in this network [9]. Assume that the length of the information flow time series during the transmission of network redundant information is N. The limited network data set U is divided into σ categories. The reconstruction of network structure redundant information in phase space is expressed as follows:

$$\begin{aligned} &U = [u(t_0), u(t_0 + (Z-1)\Delta t)] = \\ &[u(t_0) \cdots u(t_0 + (1 + (\bar{m} - 1)l)\Delta t) \cdots u(t_0 + (N-1)\Delta t)] \end{aligned} \tag{4.1}$$

$u(t_0)$is the redundant data pattern of the timing-based network structure. l represents the space partitioning scale, which is used to reconstruct the redundant information of the network in phase space. $\bar{m}$represents the network redundant information reconstruction embedding dimension. This project intends to perform feature extraction on the constructed high-dimensional phase space. $U = (u_1, u_2, \cdots, u_{\breve{n}}) \subset R^s$ is a restricted set of vectors obtained when processing redundant information in the network [10]. Network information is obtained in U by shape vectors spanning network data. The information of the network is a finite element group containing $\breve{n}$ sample values. Using u_i as the sampling point, the network is sampled for redundant information. Use (4.2) to express the high-dimensional eigenvectors in the state space

$$u_i' = (u_{i1}, u_{i2}, \cdots, u_{is},)^T \tag{4.2}$$

The method based on entropy value is used to process the redundant information of multi-dimensional data. Redundant information encoding vectors are classified into category σ. Obtain the data's clustering center by characterizing and compressing the data.

$$B = \{b_{ij\prime} | i = 1, 2, \cdots, \sigma, j\prime = 1, 2, \cdots, s\} \tag{4.3}$$

$b_{ij\prime}$ is the processing of the disturbance vector $j\prime$ when processing the redundancy information. A multi-source data clustering method based on SVD is proposed. The obtained decomposition results are expressed as follows

$$A = \{\mu_{ik} | i = 1, 2, \cdots, \sigma, k = 1, 2, \cdots, n\} \tag{4.4}$$

μ_{ik} stands for decomposed elements. Determining the network redundancy indicator function based on sparse representation

$$l(A, B) = \sum_{k=1}^{n} \sum_{i=1}^{\sigma} \mu_{ik} (c_{ik})^2 \tag{4.5}$$

$\ddot{m}$ represents the weighted weight. μ_{ik} represents the weight of the split factor μ_{ik}. $(c_{ik})^2$ represents the Euclidean distance between network sampling points u_k and $b_{ij\prime}$ that contain redundant data. (4.1) is analyzed in the case that $\sum\limits_{i=1}^{\sigma} \mu_{ik} = 1$ is satisfied. The objective function is maximized to obtain the compression of the feature space of network redundant information.

$$\begin{cases} \mu_{ik} = 1/(\sum\limits_{j=1}^{\sigma} (c_{ik}/c_{jk})^{(2/m-1)}) \\ b_{ij\prime} = (\sum\limits_{k=1} (\mu_{ik}) u_k)/(\sum\limits_{k=1}^{n} (\mu_{ik})) \end{cases} \tag{4.6}$$

Initial values with perturbation vectors are given. A data extraction method based on a fuzzy index c_{jk} is proposed.

4.2. Eliminate redundant information during data transmission. s refers to the number of categories of redundant information in the network. $\{c_1, c_2, \cdots, c_s\}$is a set of network redundant information feature types [11]. The redundant information characteristic coefficient in the network is obtained through equation (4.8)

$$\begin{aligned} JH(u) = &\sum_{k=1}^{s} P(c_k) InP(c_k) + P(u) \sum_{k=1}^{s} P(c_k|u) InP(c_k|u) \\ &+ P(\bar{u}) \sum_{k=1}^{s} P(c_k|\bar{u}) InP(c_k|\bar{u}) \end{aligned} \tag{4.7}$$

$P(c_k)$ is the proportion of type c_k network data in all network data. $P(\bar{u})$ is the proportion of all network materials with characteristic u. $P(c_k|u)$ is the proportion of redundant information containing attribute u in class c_k. Set up the feature matrix. A represents the characteristics of redundant information in the network, which can be calculated using equation (4.9).

$$Y = (y_{\omega k})_{N\prime \times Q} \times p_\omega \tag{4.8}$$

$y_{\omega k}$ is the weight of attribute ω in the redundancy data set of network data. Q represents the number of samples in the redundancy data. $N\prime$ represents the data characteristics in the network redundant information sampling set [12]. Set the probability that a redundant information feature ω is included in a network data sample k to $f_{\omega k}$. The value of $y_{\omega k}$ can be solved by equation (4.10)

$$y_{\omega k} = \begin{cases} 1 & f_{\omega k} > 0 \\ 0 & f_{\omega k} = 0 \end{cases} \tag{4.9}$$

The weighting of the redundant information characteristics of the network can be obtained by Equation (4.10). The corresponding weights describe the impact of different network redundant information characteristics on the classification of redundant information. First, standardize the redundant information properties in the network

$$\delta = f_{\omega k} \times In(Q/c_k) \Bigg/ \sqrt{\sum_{k=1} [f_{\omega k} \times In(Q/p_\omega)]^2} \tag{4.10}$$

p_ω is the probability occupied by the network redundancy information characteristic B of the sample group. Therefore, accurate classification parameters for network redundant information can be obtained according to the processing result of equation (4.11).

5. Method implementation.

Table 5.1: Model input characteristics.

Feature name	Input variables	Feature name	Input variables
current ratio	x1	cash flow ratio	x10
cash ratio	x2	Assets and liabilities	x11
Assets and liabilities	x3	Equity Multiplier	x12
Tangible net worth debt ratio	x4	debt service coverage ratio	x13
Fixed asset turnover rate	x5	Interest coverage ratio	x14
Sub-asset turnover rate	x6	Inventory turnover	x15
Current asset turnover ratio	x7	revenue growth rate	x16
return on assets	x8	total assets growth rate	x17
return on equity	x9		

Table 5.2: Influence of the number of nodes in the first hidden layer and the second hidden layer.

h1 number of nodes	Accuracy (%)	h2 Number of nodes	Accuracy (%)
27	62.7	15	69.0
25	64.0	14	72.7
23	66.0	13	70.9
21	65.8	11	74.6
19	64.8	10	78.6
17	64.6	8	71.8
15	65.9	7	71.0
13	64.5	6	70.1

5.1. Data preprocessing. This project plans to integrate deep learning and autoencoding networks to build a new deep learning system. And use it to conduct intelligent analysis of medical financial information. This article takes a large hospital as an example to verify this method. Our goal is to evaluate the performance of this system for data analysis [13]. This data collection contains 154,688 financial data for a certain period. The hospital's operating status is evaluated through the analysis of financial data during each unit period. The evaluation is divided into two categories: good operation and poor operation. This article adopts an evaluation method based on weekly financial statement data of each hospital. The autoencoder neural network was used to classify 17 image features by analyzing the original data. Feature categories are listed in Table 5.1.

The 17 economic indicators listed in Table 5.1 can reflect the financial indicators of the hospital. The evaluation system covers operating costs, profitability and prospects for future development.

5.2. Simulation results. Corresponding parameter selection must be made before conducting simulation experiments. Among them, the number of input, input, and hidden layer nodes are the most critical parameters [14]. In this algorithm, the number of nodes in the input layer is related to the eigenvalues. The number of hidden layers and the number of nodes in each layer are the main factors affecting network performance. Too many layers and nodes in the network will significantly impact processing performance when the network's depth is insufficient. This method will have over-adaptation problems during learning, which will have a particular impact on the generalization ability of experimental data. This paper uses hierarchical experiments to calculate the number of nodes in each network layer. Table 5.2 and Table 5.3 illustrate the impact of different numbers of hidden layers and the number of hidden layer nodes on the modeling results.

In the case of 22 nodes, the model can achieve the best accuracy of 66.04%. When the number of nodes is 10, the system can obtain the best accuracy of 78.65%. When adding the third hidden layer, the optimal solution of the algorithm dropped from 83.33% to 81.98%. The network structure is relatively complex, and over-adaptation can occur during learning [15]. The network's performance combined with the autoencoder was evaluated through comparative experiments. Detailed results are listed in Table 5.4.

Table 5.4 shows the data results obtained by several shallow machine learning methods. Compared with the

Table 5.3: Influence of the number of nodes in the third and fourth hidden layers.

h3 number of nodes	Accuracy (%)	h4 number of nodes	Accuracy (%)
11	77.3	9	77.3
10	76.3	8	76.3
9	78.2	7	80.9
8	83.3	6	82.0
7	75.2	5	75.4
6	76.1	4	72.1
5	77.1	3	77.3
4	74.1	2	76.3

Table 5.4: Network parameters.

Network parameters	AUC	Accuracy (%)
LR	0.60	63.75
Random Forest	0.74	77.40
BP Network	0.74	64.90
The algorithm of this article	0.84	83.33

shallow method, the AUC value of the method proposed in this article is increased to 0.84, and the Accuracy value is increased to 84.38%. However, the random distribution model that currently performs best in shallow machine learning has an AUC value of only 0.74 and an Accuracy value of 77.40%. The algorithm in this article has improved by 0.10 in AUC and 5.94% in accuracy. Both indices have improved significantly.

6. Conclusion. This project proposes an automatic extraction method for financial big data based on autoencoding networks. Comparative experiments show that the medical financial data processing method based on deep neural networks and redundant data elimination has achieved significant results.

REFERENCES

[1] Rünzel, M. A., Hassler, E. E., Rogers, R. E., Formato, G., & Cazier, J. A. (2021). Designing a smart honey supply chain for sustainable development. IEEE Consumer Electronics Magazine, 10(4), 69-78.

[2] Li, J., & Zhang, Y. (2022). Construction of smart medical assurance system based on virtual reality and GANs image recognition. International Journal of System Assurance Engineering and Management, 13(5), 2517-2530.

[3] Imtiaz, A., Gousia, H., & Parmod, K. (2023). Machine Learning-based Peer-to-Peer Platform for Precision Agriculture in Crop Growth and Disease Monitoring. Journal of Computer Technology & Applications, 13(03), 26-39.

[4] Treleaven, P., Smietanka, M., & Pithadia, H. (2022). Federated learning: the pioneering distributed machine learning and privacy-preserving data technology. Computer, 55(4), 20-29.

[5] Santamaría, P., Tobarra, L., Pastor-Vargas, R., & Robles-Gómez, A. (2023). Smart Contracts for Managing the Chain-of-Custody of Digital Evidence: A Practical Case of Study. Smart Cities, 6(2), 709-727.

[6] Amalraj, J. R., & Lourdusamy, R. (2022). Security and privacy issues in federated healthcare—An overview. Open Computer Science, 12(1), 57-65.

[7] Bhardwaj, S., & Dave, M. (2022). Crypto-preserving investigation framework for deep learning based malware attack detection for network forensics. Wireless Personal Communications, 122(3), 2701-2722.

[8] Ghasemi, M., Anvari, D., Atapour, M., Stephen Wormith, J., Stockdale, K. C., & Spiteri, R. J. (2021). The application of machine learning to a general risk–need assessment instrument in the prediction of criminal recidivism. Criminal Justice and Behavior, 48(4), 518-538.

[9] Gómez, J. M., Fares, O. H., Mohan, M., & Lee, S. H. M. (2023). Blockchain in the Food Industry: Integrating Machine Learning in a Systematic Literature Review. Journal of International Technology and Information Management, 32(1), 32-58.

[10] Ibrahim, A. K., Hassan, M. M., & Ali, I. A. (2022). Smart Homes for Disabled People: A Review Study. Science Journal of University of Zakho, 10(4), 213-221.

[11] Keefe, R. F., Zimbelman, E. G., & Picchi, G. (2022). Use of individual tree and product level data to improve operational forestry. Current Forestry Reports, 8(2), 148-165.

[12] Tan, T. F., Li, Y., Lim, J. S., Gunasekeran, D. V., Teo, Z. L., Ng, W. Y., & Ting, D. S. (2022). Metaverse and virtual health care in ophthalmology: Opportunities and challenges. The Asia-Pacific Journal of Ophthalmology, 11(3), 237-246.

[13] Pothula, S. R. (2023). Review and analysis of FinTech approaches for smart agriculture in one place. Journal of Agriculture, Science and Technology, 22(1), 60-69.

[14] Thomason, J. (2021). Big tech, big data and the new world of digital health. Global Health Journal, 5(4), 165-168.

[15] Yathiraju, N. (2022). Investigating the use of an Artificial Intelligence Model in an ERP Cloud-Based System. International Journal of Electrical, Electronics and Computers, 7(2), 1-26.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Oct 10, 2023
Accepted: Oct 23, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1757–1764, DOI 10.12694/scpe.v25i3.2680

DESIGN OF SMART HOME SYSTEM BASED ON WIRELESS SENSOR NETWORK LINK STATUS AWARENESS ALGORITHM

RONG XU*

Abstract. When wireless sensor networks are used in smart homes, the connection state will be unstable due to signal masking attenuation. This will cause low packet rate, high time delay and high cost in the network. In this paper, a network routing algorithm for wireless sensing based on connection conditions is designed. Secondly, the expected number of sends is proposed to evaluate the stability of links. Based on this, the following network signal delivery situation is forecasted in real time and quickly. According to the estimated expected number of transmissions, the path is dynamically corrected to effectively avoid attenuation in the channel and achieve optimal system performance. Experimental results show that the method proposed in this paper can improve the efficiency of message sending and reduce the routing cost under the condition of masking effect.

Key words: Wireless sensor network; Routing; Smart family; Shielding attenuation; Link potential sense

1. Introduction. Due to the particularity of the building structure, it has a solid electromagnetic interference ability and then has a more significant impact on the performance of the smart home wireless sensor network (WSNSH) system. The communication between perceptron's is affected by ground, wall and human movement. Because of the instability of the link connection, the network communication often fails. Some critical alarm information will likely be sent out too late, bringing significant security risks to users.

ZigBee network has been widely used in the practical application of smart home wireless sensor networks. Its default routing algorithm, LEACH, generally adopts flood routing to ensure that efficient routes can be found in the case of attenuation. However, this method also has two problems: First, the flood routing cost of the LEACH method is high; The existing methods do not fully use the wireless transmission characteristics in the environment and cannot accurately estimate the wireless transmission performance. It cannot be dynamically adjusted according to the actual situation. Therefore, its fundamental transmission characteristics in wireless channels, such as wireless SNSH, are not ideal. Second, although link quality indication (LQI) is used to evaluate the link state in the ZigBee node neighbor table and can help routing decisions, LQI needs to determine its value by repeatedly sending and receiving beacon frames. This increases the routing burden. In smart home networking systems, the energy-saving technologies of data communication between data nodes of wireless sensor networks and the whole sensor network mainly revolve around low energy consumption media access control technology, compressed sensing technology, low duty ratio working technology, low energy routing technology and so on. This paper [1] proposes a path selection method to solve the swarm clustering problem.

When the algorithm is used to select the cluster head of WSN in the smart home system network system, a series of factors such as the residual energy of data nodes, the location information of data nodes and the node density of data nodes in WSN should be considered comprehensively. Although this algorithm is reasonable for cluster selection in WSN, it also has a significant defect. Its convergence is poor. This paper studies a WSN path selection method for smart homes. Based on LEACH, the scalable Transmission Count (ETX-SH) is proposed to replace LQI to describe the transmission status of WSNSH under channel conditions [2]. The routing cost is reduced by using the directed routing method. At the same time, a path planning method based on optimal state is proposed.

2. Smart home wireless sensor network. The construction of WSNSH is usually similar to Figure 2.1 (image cited in Wireless Personal Communications, 2018, 101:1019-1055.). The network generally consists

*School of Information and Architectural Engineering, Anhui Open University, Hefei, 230022, China; Corresponding author's e-mail: ahouxurong@126.com

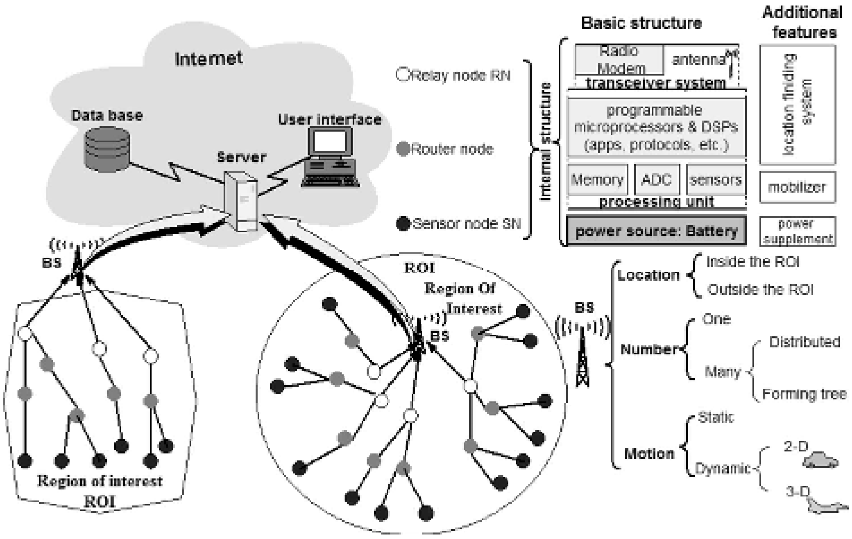


Fig. 2.1: Topology of smart home wireless sensor network.

of terminal, routing, and coordinator nodes. In a network, nodes are usually stationary [3]. In these sensors, the end nodes often collect data from sensors such as temperature, humidity, combustible gases, and infrared monitoring. The routing node is used to transmit data to the partner point. The coordinator node completes the network and data summary of the whole system.

3. LEACH algorithm. Classical LEACH is a hierarchical network structure control scheme based on the "wheel." Establishing a cluster head in each cycle and the information transfer between nodes in the cluster is the essential work of WSN. In the classical LEACH method, when the cluster head is selected, each node in WSN will generate a corresponding random sequence [4]. This random number is on a scale of 0 to 1. Compare this random number with the threshold $S(m)$ determined by formula (1) and determine whether this data node can serve as the cluster head.

$$S_j(m) = \begin{cases} \dfrac{q}{1 - q[d^* \bmod (\frac{1}{q})]}, m \in R \\ 0, m \notin R \end{cases} \tag{3.1}$$

The number of clusters in WSN that currently have several votes in the cluster is d. A cluster head algorithm based on cluster theory is proposed. The ratio of cluster heads to the number of data nodes in the wireless sensor network is represented by q. R is a cluster of nodes without cluster heads.

3.1. Improvement of LEACH algorithm.

3.1.1. Classification method of "hot area" and "non-hot area.". In the organizational structure of smart homes, the central node is generally the core to complete the control of the entire WSN. There must be a central node. Nodes need to have two main functions to realize the control of each node through the control terminal: one is to be able to carry out information transfer, and the other is to be able to carry out a network connection. Secondly, there must be adequate data collection capabilities [5]. The central node broadcasts a message to the object monitoring area. The monitored objects are partitioned according to the relationship between them and the central node and are divided into "hot areas." The subregions far from the central node are collectively called "non-hot areas" based on the information received. Each node divides its location into a "hot zone" and a "non-hot zone." The upper and lower bounds of the subzone j are:

$$W_j = c_{\min} + j \times \frac{c_{\max} - c_{\min}}{v}, j = 1, 2, \cdots v \tag{3.2}$$

$$LB_j = c_{\min} + (j-1) \times \frac{c_{\max} - c_{\min}}{v}, j = 1, 2, \cdots v \tag{3.3}$$

$c_{\max}$ represents the maximum distance from the central node. $c_{\min}$ represents the shortest distance from the data child node to the central node. v is the number of "hot zones" and "non-hot zones" in the target area.

3.1.2. Dynamic regulation principle of cluster radius. The number of nodes in each cluster in the "hot zone" and "non-hot zone" depends on two essential factors: monitoring range

$$D$$

and node density

$$\varphi$$

. The following formula can calculate the number of nodes in a cluster:

$$v = \pi D^2 \varphi \tag{3.4}$$

In WSN, each cluster selection process will generate the corresponding node energy consumption. The energy consumption of the cluster head is calculated according to the following formula:

$$Y_{total} = Y_{rec} + Y_t{=}\pi D^2 \varphi \times t + s) \times Y_e + \sigma \times (Y_e + \delta_{amp} c^2) \tag{3.5}$$

σ represents the amount of data owned by each member node in the cluster. s represents the amount of data passed by other cluster heads. According to formula (3.6), the relationship between each data node's competition radius and the group head's position is determined.

$$D = (1 - \alpha \times \frac{c_{\max} - c}{c_{\max} - c_{\min})} \times D_0 \tag{3.6}$$

c is the distance between the head of the cluster and the central node. α represents the influence of the distance between cluster heads and central nodes in the "hot zone" and "non-hot zone" on their competition radius [6]. The competition radius of "hot zone" and "non-hot zone" is shown in formulas (3.7) and (3.8).

$$D_{hot} = [\delta_1 \times (1 - \alpha \times \frac{c(S_j, BS) - c_{\min}}{c_{\max} - c_{\min}} + \delta_2 \times \frac{Y_{res}}{Y_{init}}] \tag{3.7}$$

$$\begin{aligned} D_{unhot} &= [1 - \alpha \times \frac{c_{\max} - c(S_j, BS)}{c_{\max} - c_{\min}}] \times D_0 \\ &+ [\lambda_1 \times sgn(Y_{res} - Y_{ave} + \lambda_2) \times \frac{x}{v}] \times \Delta D \end{aligned} \tag{3.8}$$

Y_{res} indicates the residual power of the node. Y_{ave} is the average value of the remaining node energy of the two adjacent data nodes. $c(S_j, BS)$ represents the distance between the node S_j and the central node. ΔD represents the competitive radius adjustment value of "hot zone" and "non-hot zone." x and v indicate the number of nodes in an area and an entire area, respectively.

3.2. Cluster Header Selection. Each node generates a random number when selecting the cluster head under initial conditions [7]. This random number is between 0 and 1. Compare the random number to the new threshold $S(m)$ calculated from the formula (3.9). If the random number is lower than the new threshold, it will be played in the whole network, and the new threshold$S(m)$ is determined to be the new cluster head.

$$S(m) = \begin{cases} \frac{q}{1 - q^*(d^* \bmod \frac{1}{q})} \cdot \left(\mu_1 \frac{Y_{cur}(m)}{Y_{init}} + \mu_2 (1 - \frac{c_{ctos}}{c_{\max}}\right) m \in R \\ 0, m \notin R \end{cases} \tag{3.9}$$

An improved LEACH method is proposed, which introduces parameters $Y_{cur}(m), Y_{init}$ and c_{ctos}. It allows for better consideration of cluster selection. $Y_{cur}(m)$ represents the energy left at the current data node. Then $Y_{init}$is the initial amount of energy for that data point. c_{ctos}represents the distance between the current node and the Sink node [8]. Two methods μ_1 and μ_2 are used to reduce the weight of each parameter in the network. Where $\mu_1 + \mu_2 = 1, \mu_1 \geq 0, \mu_2 \geq 0$.

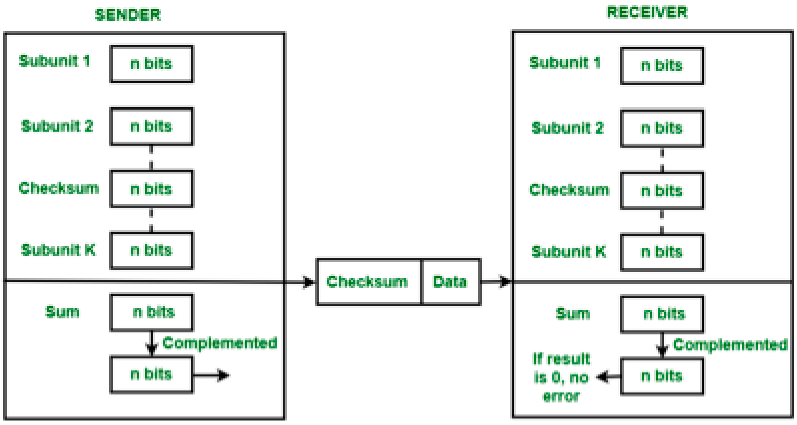


Fig. 3.1: Schematic diagram of node fault correction.

3.3. Data communication mode. The data communication mode is mainly optimized according to the ”hot area” and ”non-hot area” divisions. The process is as follows.

1) If the cluster head node S_j is in the ”hot zone,” then this node does not need to exchange data with other cluster nodes. The data information between the front-end data node and the central node is exchanged to realize the energy saving of the node.

2) If the location of the group head node S_j is in the ”hot zone” that has not been determined in advance, then a group head group near the group head node must be reconstructed with it. The network transmission mechanism is proposed. This mechanism uses the cost function of network transmission information to find the minor network transmission node S_j. Repeat the above steps until you find the data to transfer to the central node. Formula (3.10) represents an expression used to calculate the cost function for data communication.

$$\psi = \omega \times \frac{c^2(S_i, S_j) + c^2(S_j, BS)}{c^2(S_j, BS)} - \upsilon \times \frac{Y_{res}}{Y_{init}} + \psi \times \frac{N_{member}}{N} \tag{3.10}$$

3.4. Network organization self-recovery mechanism. Even if a small number of wireless communication nodes fail in a smart home, it will not affect the essential characteristics of the entire home network [9]. According to the characteristics of the smart home, the local repair is carried out in four states: node failure, adding, deleting and moving.

3.4.1. Node fails. Network failure occurs when the communication node encounters a short circuit or power consumption in the power system. Communication lines need to be repaired. The patching process is shown in Figure 3.1.

If data node 01 is faulty, links 00-01, 01-05, and 01-08 are also faulty. Data node 01 that fails without data transfer will not be immediately detected. Communication was cut off until 01’s data node was asked to transmit data. In this case, the link must be repaired [10]. The detailed repair process for node data transmission is as follows. 1) If central node A wishes to send data to node 09, the information should be sent at 00-01-05-07-09 according to the original path. When node 01 cannot be found when data is transmitted through node 01, node 00 sends a command to the control terminal. This means that node 01 has failed and needs to be maintained. At this moment, you need to check whether node 05 is the neighbor of node 00 according to the route information recorded during the networking.

3.4.2. Adding Nodes. According to the existing routing information, a new node is added to the existing network [11]. Each node has its route information. When a new data node sends a request to its neighbor to join the neighbor cluster, the smart home network usually responds to the response of another data node. At the same time, the intermediate node can be found according to the path of other data nodes.

Table 4.1: Simulation scenario parameters.

Parameter	Value
Nodal area	(0,0) ∼(20,10) m
Total inductor node	10, 20, 30, 40
Fading link | Fading link	8
Arc of oriented conduit	2m
Link attenuation threshold	1.651×10^5
Smoothness coefficient α	0.7

3.4.3. Node shift. The methods of solving the two types of nodes of motion data and motion alone are also different. When a large number of data nodes are migrated, the central node is used to reissue new networking instructions to realize the reconstruction of the whole networking process. When the movement of a single data node is low, insert a new data node again [12]. In this case, the central node does not need another significant network restart.

4. Algorithm simulation. MATLAB software simulates the method's effectiveness in the WSNSH environment. This paper compares and analyzes the improved LEACH algorithm, LEACH algorithm, and multipath routing algorithm AOMDV. In this way, the characteristics of different methods can be shown from different angles.

4.1. Setting of simulation scenarios and parameters. The simulated smart home network area is 20*10 square meters [13]. A wireless SNSH scheme based on WSNSH is proposed in this paper. Real applications were simulated using between 10 and 40 different nodes. The channel attenuation threshold is determined based on the nominal transmit-receive energy ratio of TI's CC2530 chip. The simulation parameters are listed in Table 4.1.

4.2. Simulation performance index. The paper evaluated three performance indicators to compare the effects of the above three methods in WSNSH during the simulation process:

(1) Message transfer rate: the ratio between the number of messages received by the target node and the number of messages sent from the source node. This value can be used to describe the probability of success of packet sending.

(2) Average: The average number of path nodes that must pass when transmitting information groups from the source node to the target node [14]. This value describes the time it takes for a packet to be sent. It's proportional to the time delay.

(3) Routing cost: the number of instruction groups in routing processing. This value can be used to describe the characteristics of network congestion. It's proportional to the time delay.

4.3. Experimental Results.

4.3.1. Packet Sending Rate. The more nodes there are, the more links are available. Compared with the LEACH method, the improved LEACH algorithm has a higher submission rate when the number of nodes is the same. Since the LEACH method cannot control the link state effectively in the transmission process, the possibility of packet loss is very high, so this paper proposes an improved LEACH method based on ETX-SH. The expected number of links between nodes in the network is calculated, and the optimal path is selected to minimize the ETX-SH in the network [15]. Determine the most stable path on the link for sending packets. In a sense, this is also a way to improve the delivery success rate. When the number of nodes is small, the recurrence rate of AOMDV is not high. However, its delivery rate increases rapidly with the number of knot points, consistent with the improved LEACH method. This is because small networks do not take full advantage of multiple paths. The AOMDV method will generate more backup chains as the network scale increases. When the primary link fails to be sent, it can be transferred to the secondary link and sent again in time [16]. In wireless SNSH, multipath transmission technology reduces channel influence on system performance. Figure 4.1 shows the simulation comparison of packet delivery rates of the three algorithms.

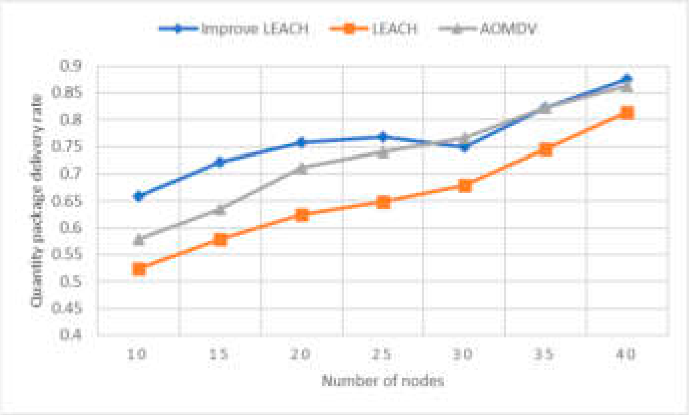


Fig. 4.1: Comparison of packet transmission rate simulations.

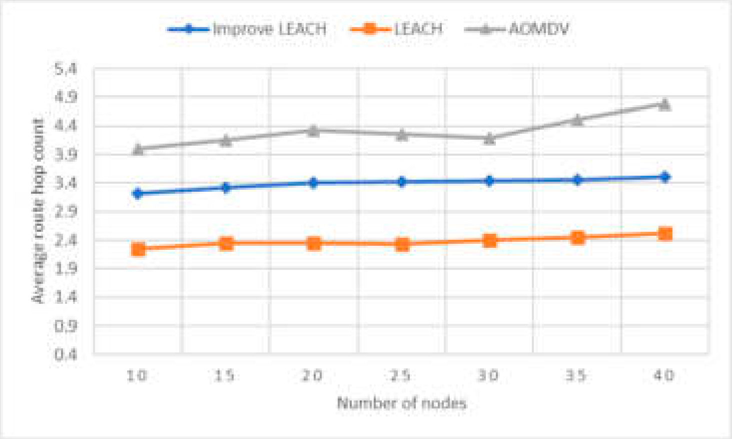


Fig. 4.2: Comparison of the average number of paths jumps under different algorithms.

4.3.2. Average number of hops. Figure 4.2 shows the results of the average number of jumps simulation. When the number of nodes is large, the size of the network will increase, and the number of nodes required for each transmission will also increase. Therefore, the number of paths hops the three methods require to send packets increases. AOMDV has the most significant number of hops when the number of knots is equal [17]. This is because the poor condition of the primary link in the case of channel attenuation leads to multiple transmissions when the AOMDV switches to other links. This will lead to more path jumps. At the same time, improving the LEACH algorithm requires more hops. This is because the improved LEACH algorithm selects lower links to ensure the success rate of data exchange [18]. This attenuation connection can be ignored when the shortest connection is in poor condition. However, the result is that the number of path hops increases relative to the LEACH method.

4.3.3. Route Cost. Figure 4.3 shows the results of the routing load simulation. Network complexity increases as the number of nodes increases, and more command packets must be sent and received during routing [19]. The routing cost of the three methods increases with the number of nodes. The minimum routing cost is obtained by improving the LEACH algorithm when the junction number is constant. There are two reasons for this conclusion. One is to improve the LEACH algorithm to use ETX-SH instead of LQI, thus reducing the cost of repeatedly sending and receiving signals when obtaining LQI. The second is to improve

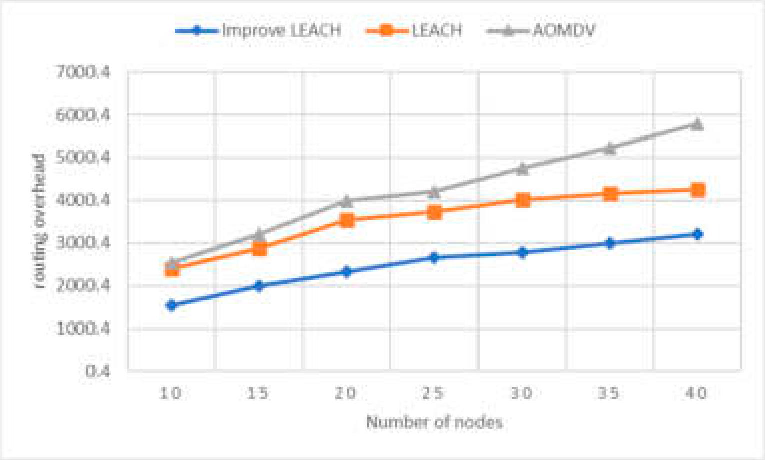


Fig. 4.3: Simulation comparison of path load.

the LEACH algorithm to use the directed route, thus limiting the direction and scope of finding the route. Compared with LEACH and AOMDV global flood forecasting methods, this method can significantly reduce the time needed to find the path. At the same time, AOMDV has a higher routing cost than LEACH. This is mainly due to the large number of instructions AOMDV consumes to construct and maintain routing tables when multipathing is performed.

The improved LEACH algorithm has apparent advantages over LEACH and AOMDV regarding transmission rate and routing cost. This algorithm can improve the success rate of transmission, reduce the additional cost of repeated transmission after transmission failure, and reduce the broadcast storm caused by high transmission costs. It plays a vital role in improving the performance of the network. A new LEACH method is proposed on delay, requiring only an average of 1-2 hops. Because the link is in good condition, the time spent on adding one or two hops is negligible. At the same time, the delay caused by the reduced routing cost is offset by the delay caused by the extra jump. In general, improving the LEACH algorithm and AOMDV can better solve the high-speed and stable data transmission in the case of occlusion.

5. Conclusion. This paper presents an improved LEACH method, which can effectively guarantee the WSNSH system's performance in occlusion. The key is to calculate the ETX-SH value of the link to improve the network performance. Through the analysis of ETX-SH data, the ETX-SH data of the next time point can be obtained to help the user choose the path. Avoid attenuated links as much as possible during the sending process. Select stable links with reasonable expectations during the sending process to ensure the sending rate. Directional routing is used to limit the search scope of the route and further improve the accuracy of the route at a lower cost. The simulation results show that the WSNSH network with channel attenuation has higher transmission efficiency and lower routing cost than the AOMDV network with multipath. This algorithm improves the performance of the WSNSH system well.

Acknowledgments. Natural science research project of Higher Education Institutions of Anhui Province, China (Grant No. 2022 AH052683).

REFERENCES

[1] Cui, Y., Zhang, L., Hou, Y., & Tian, G. (2021). Design of intelligent home pension service platform based on machine learning and wireless sensor network. Journal of Intelligent & Fuzzy Systems, 40(2), 2529-2540.

[2] Pirzada, P., Wilde, A., Doherty, G. H., & Harris-Birtill, D. (2022). Ethics and acceptance of smart homes for older adults. Informatics for Health and Social Care, 47(1), 10-37.

[3] Torad, M. A., Bouallegue, B., & Ahmed, A. M. (2022). A voice controlled smart home automation system using artificial

intelligent and internet of things. TELKOMNIKA (Telecommunication Computing Electronics and Control), 20(4), 808-816.

[4] Abdulrahman, L. M., Zeebaree, S. R., Kak, S. F., Sadeeq, M. A., AL-Zebari, A., Salim, B. W., & Sharif, K. H. (2021). A state of art for smart gateways issues and modification. Asian Journal of Research in Computer Science, 7(4), 1-13.

[5] Hamdan, Y. B. (2021). Smart home environment future challenges and issues-a survey. Journal of Electronics, 3(01), 239-246.

[6] Touqeer, H., Zaman, S., Amin, R., Hussain, M., Al-Turjman, F., & Bilal, M. (2021). Smart home security: challenges, issues and solutions at different IoT layers. The Journal of Supercomputing, 77(12), 14053-14089.

[7] Zou, S., Cao, Q., Wang, C., Huang, Z., & Xu, G. (2021). A robust two-factor user authentication scheme-based ECC for smart home in IoT. IEEE Systems Journal, 16(3), 4938-4949.

[8] Verma, R. (2022). Smart city healthcare cyber physical system: characteristics, technologies and challenges. Wireless personal communications, 122(2), 1413-1433.

[9] Mohammad, Z. N., Farha, F., Abuassba, A. O., Yang, S., & Zhou, F. (2021). Access control and authorization in smart homes: A survey. Tsinghua Science and Technology, 26(6), 906-917.

[10] Luo, H., Wang, C., Luo, H., Zhang, F., Lin, F., & Xu, G. (2021). G2F: a secure user authentication for rapid smart home IoT management. IEEE Internet of Things Journal, 8(13), 10884-10895.

[11] Tanveer, M., Abbas, G., Abbas, Z. H., Bilal, M., Mukherjee, A., & Kwak, K. S. (2021). LAKE-6SH: Lightweight user authenticated key exchange for 6LoWPAN-based smart homes. IEEE Internet of Things Journal, 9(4), 2578-2591.

[12] Ahmed, A. A., Belrzaeg, M., Nassar, Y., El-Khozondar, H. J., Khaleel, M., & Alsharif, A. (2023). A comprehensive review towards smart homes and cities considering sustainability developments, concepts, and future trends. World J. Adv. Res. Rev, 19(1), 1482-1489.

[13] Duric, I., Barac, D., Bogdanovic, Z., Labus, A., & Radenkovic, B. (2023). Model of an intelligent smart home system based on ambient intelligence and user profiling. Journal of Ambient Intelligence and Humanized Computing, 14(5), 5137-5149.

[14] Khalid, N., Mirzavand, R., Saghlatoon, H., Honari, M. M., Iyer, A. K., & Mousavi, P. (2021). A Batteryless RFID sensor architecture with distance ambiguity resolution for smart home IoT applications. IEEE Internet of Things Journal, 9(4), 2960-2972.

[15] Jurado-Lasso, F. F., Clarke, K., Cadavid, A. N., & Nirmalathas, A. (2021). Energy-aware routing for software-defined multihop wireless sensor networks. IEEE Sensors Journal, 21(8), 10174-10182.

[16] Bajaj, K., Sharma, B., & Singh, R. (2022). Implementation analysis of IoT-based offloading frameworks on cloud/edge computing for sensor generated big data. Complex & Intelligent Systems, 8(5), 3641-3658.

[17] Liu, F., Cui, Y., Masouros, C., Xu, J., Han, T. X., Eldar, Y. C., & Buzzi, S. (2022). Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond. IEEE journal on selected areas in communications, 40(6), 1728-1767.

[18] Cui, Y., Liu, F., Jing, X., & Mu, J. (2021). Integrating sensing and communications for ubiquitous IoT: Applications, trends, and challenges. IEEE Network, 35(5), 158-167.

[19] Tan, S., Ren, Y., Yang, J., & Chen, Y. (2022). Commodity WiFi Sensing in Ten Years: Status, Challenges, and Opportunities. IEEE Internet of Things Journal, 9(18), 17832-17843.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Oct 9, 2023
Accepted: Oct 19, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1765–1772, DOI 10.12694/scpe.v25i3.2682

E-COMMERCE DATA MINING ANALYSIS BASED ON USER PREFERENCES AND ASSICIATION RULES

ZHIYING FAN[*]

Abstract. Improving the sales of e-commerce platforms is the primary goal of this paper. This paper studies the data of e-commerce product recommendations from the perspective of user preference and association rules. The characteristics of positive and reverse association rules in data mining are analyzed. Then, a multi-dimension association rule calculation method is proposed. Create a data attribute unit set. By analyzing each attribute's weighted coefficient and similarity, the attribute confidence degree is obtained, and the data is preprocessed. An example is given to verify the effectiveness of the proposed method. The recommendation engine based on user preferences and association rules significantly improves the accuracy, recall rate and prediction coverage of e-commerce recommendation systems.

Key words: User preference; Association rules; Electronic commerce; Data mining; Recommendation system

1. Introduction. Nowadays, most e-commerce websites have adopted various ways to provide information services. A sound referral system can improve the chances of individual items being viewed. It can improve the time consumers spend in online stores, help consumers find products they are genuinely interested in, and improve their purchasing experience. This will increase site visits and product sales. Generally speaking, product recommendations on an e-commerce website can be based on the product's characteristics, for example, brand, category, applicable age group, etc. However, this requires a professional to review the product. Suggestions can also be made based on the user's browsing habits. Some scholars have proposed an automatic classification method based on association rules [1]. It can find the correlation law between commodities based on analyzing various transaction records in the transaction database, and then assist merchants in making corresponding business decisions—for example, purchase, sales and inventory management, shelf placement, etc. Compared with similar methods, this method has higher computational speed and accuracy. Meaningful associations can be found by analyzing the correlation of massive user behaviors and product metadata in the database to realize personalized customer recommendations based on the association rule algorithm. E-commerce extensive data processing recommendation systems can provide convenience to customers and bring more revenue to the company [2]. Therefore, this paper analyzes the big data of e-commerce systems based on user preference and association rule algorithm. The conclusion obtained in this paper has significant theoretical research value and practical guiding significance.

2. Research on data mining methods and user preferences based on multidimensional association rules.

2.1. Application of multidimensional association rule method in e-com merce data preprocessing. The preprocessing of e-commerce data is the premise of data mining for the same structure data [3]. Feature extraction is the first step to preprocessing e-commerce data. Aiming at the sparse problem of e-commerce data, an e-commerce data association method based on multidimensional association rules is proposed. This allows you to find features and properties in your e-business data. The set of characteristic elements of the e-business data can then be expressed as:

$$R = \{[r_1, C(r_1)], [r_2, C(r_2)], \cdots, [r_i, C(r_i)]\} \tag{2.1}$$

[*] Shanxi Institute of Mechanical and Electrical Technology, Changzhi, Shanxi, 046011, China (Corresponding author, F13191259977@163.com)

r_i stands for the characteristic unit of electronic commerce data. $C(r_i)$ represents the number of characteristic units in electronic commerce data. When the value of the multidimensional correlation function is significant, the parameter value is also considerable. In this case, the EPC model parameter estimation obtained is relatively accurate. In this way, the parameter estimation of e-commerce big data is transformed into a target-optimal problem with limitations [4]. The generated target functionality is described as follows:

$$h(x) = \max_x y(x) \tag{2.2}$$

The multidimensional association function $h(x)$ is a fitness function of e-commerce data set parameters. A matching index based on multidimensional association rules is proposed to measure the matching degree of e-commerce data features and attributes [5]. These formulas are:

$$W(a_i, r_j) = \sum_{i=1}^{n} C_i(r_j) \frac{U}{\sum\limits_{j=1}^{n} C_i(r_j)} \tag{2.3}$$

$C_i(r_j)$ represents the statistic of the e-commerce data attribute. U represents the statistical number of all units with characteristics in the e-commerce data [6]. If the $V = \{v_1, v_2, \cdots, v_n\}$set is used to define the relevant rules of EPC data, the correlation between e-commerce data can be expressed as:

$$L\{v_1, v_2, \cdots, v_n\} = \frac{1}{h_i} \sum_{i=1}^{n} \max v_i \tag{2.4}$$

$H = \{h_1, h_2, \cdots, h_n\}$ represents the weight vector for e-commerce data. n represents the range of sub-business data. Assuming that there are weights in the e-commerce data in the attribute set, the similarity between e-commerce data c_i and attribute a_i is expressed as:

$$Sam(h_i, a_i) = \frac{h_i}{||h_i||} \tag{2.5}$$

The attribute set X^a is labeled by the similarity of the data and its attributes [7]. The reliability analysis of the characteristics of e-commerce data is carried out. The calculation formula is:

$$Cor_{sam} = \sum_{sam_{\in} \in SamX^a} \frac{Sam_i}{|X^a|} \tag{2.6}$$

$SamX^a$ in formula (2.6) is the similarity group of e-commerce data and its attributes. Sam_i stands for characteristics similar to electronic commerce data. $|X^a|$ represents the number of attributes present in the attribute set X^a of the e-commerce data. This paper transforms the parameter estimation problem of e-commerce big data into a multi-objective optimal problem with constraints [8]. Then, the method of solving multidimensional association rules is proposed. E-commerce data is preprocessed by combining similarity and confidence.

2.2. Construction of e-commerce data model. Electronic commerce data includes a lot of network information. Website Q_i also contains a large amount of content and structure of electronic transaction data [9]. The structure of web pages is analyzed by constructing an e-commerce data model. The e-business data schema is represented as follows:

$$Q_i = (Z_i, Y_i, T_i) \tag{2.7}$$

Z_i represents electronic business data in the organized form of Web pages. Y_i represents the electronic transaction data target in the Web page, which is detected by the entity. T_i stands for electronic business data included on Web pages. E-commerce data priority value is calculated according to association rules and user preference algorithms.

$$W\left(\psi_i^j\right)=\frac{\frac{1}{\sigma_{ij}^k}\left(\sum_{k=1}^{m}\varphi_{ij}^k\right)^2}{\Omega\left(C_i^j\right)} \tag{2.8}$$

σ_{ij}^k and φ_{ij}^k in formula (2.8) represent the threshold for determining the e-commerce data block j in the adjacent window k. $\left(\psi_i^j\right)$stands for the j e-commerce data block in the web page window i. $\Omega\left(C_i^j\right)$ represents the amount of ψ_i^j contained in the page. m represents the number of electronic commerce data in adjacent Windows [10]. The data are arranged in order of importance. If the maximum sorting time of e-commerce data is set to $t_{\max}$, the sorting result of e-commerce data can be expressed as:

$$C^h=\frac{1}{t_{\max}}\sum_{t=1}^{n}Q_i \tag{2.9}$$

When classifying e-commerce information, the whole minimization principle is used to classify the data and determine its suitability [11]. The model of the e-commerce data object is established. The following formula is used to calculate fitness:

$$Fithess=\sum_{i=1}^{N}\left(x_i^2-x_i\right)^2 \tag{2.10}$$

x_i represents the expected output. N represents the sample size of e-commerce data. The $\tilde{x}_i$ stands for the actual result. The e-commerce data thus constructed can be expressed as:

$$\phi=\frac{1}{A_i}\sum_{i=1}^{m}\lambda_i\frac{L_i}{\vartheta_i\cdot\kappa_i}+Cor_{sam} \tag{2.11}$$

L_i represents the E-Business Data item at bit i in E-Business Data object Q_i. ϑ_i stands for the name of the e-commerce data item i. κ_i represents the value of the e-commerce data item i. Cor_{sam} stands for the degree of trust of item i of e-commerce data. λ_i represents the weighting of e-commerce data item i. A complete target model of electronic commerce business is formed by calculating the above links.

2.3. E-commerce data mining algorithm flow. According to the target pattern of e-commerce data, the data segmentation scheme with the highest priority is found [12]. The objective function of e-commerce data optimization is defined.

$$\min\varepsilon=\frac{1}{h_i^{(a)}}\sqrt{\frac{q_i\Omega(Z(\theta)-Z(r))}{\phi_n(v)}} \tag{2.12}$$

$h_i^{(a)}$ of formula (2.12) represents the characteristics of e-commerce data. q_i represents the eigenvector of e-commerce data. $Z(\theta)$ stands for the amount of information contained in the e-commerce data. $\phi_n(v)$ represents the difference component in the e-commerce data of the two characteristics. Ω stands for the optimal classification threshold of e-commerce data. $Z(r)$ represents the number of e-commerce data characterized by r.

2.4. User preference degree model. A preference modeling method based on user behavior data is proposed. Because the display area of the page in the recommendation scenario is limited, the items ranked higher in the recommendation list will be more likely to attract users' attention. If the user's classification model can be used to classify products, it will have a practical guiding effect [13]. This method is of great significance to improve the system's accuracy and customer loyalty. A description of the user priority algorithm is shown in Figure 2.1.

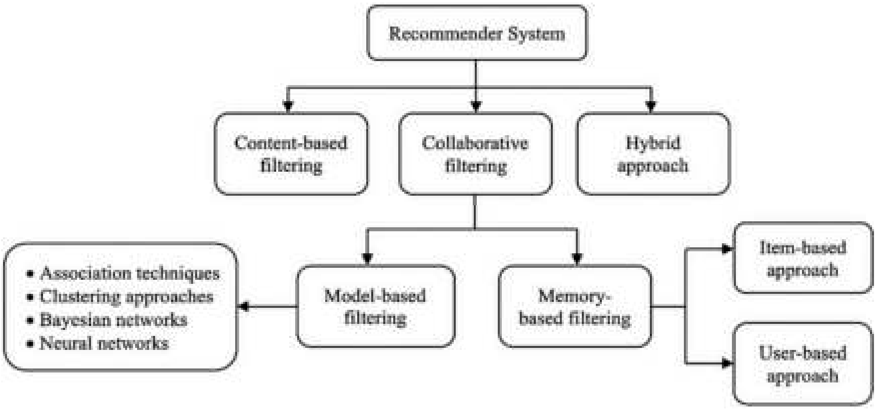


Fig. 2.1: User preference algorithm description.

3. Design of e-commerce data mining system.

3.1. System Structure. Establish an intelligent e-commerce site recommendation system. First, we must go through the data collection and extraction process. The log center of the mall is used to extract the information of users and items related to the algorithm. In addition, the matching process between the converted data and the loading module is carried out to improve the algorithm's effectiveness and the service's confidentiality. Before the establishment of the model, the corresponding countermeasures are put forward for the violation of ”brushing the list.” The aim is to improve the resistance of the rule association algorithm [14]. Minimize the direction of the specific combination of the modeled objects and the product, and store the modeled objects in the modeled database of the recommendation system. The intelligent recommendation system reads the algorithm input model required by the algorithm from the corresponding modeling database for algorithm calculation. Finally, the recommended results are transmitted to the relevant commercial system of the mall in the form of a protocol. Figure 3.1 shows the overall architecture design of the e-commerce recommendation system (the picture is quoted in Egyptian Informatics Journal, Volume 23, Issue 1, March 2022, Pages 33-45).

3.2. Functional architecture design. It is necessary to design and implement many functions in network information service reasonably to successfully implement good network information service. The differences in modular particles, coupling and cohesion among modules will directly affect the development efficiency and operation performance of the whole system [15]. This paper completes the essential system management, realizes the control of the running process of the system, and quickly configures and updates. The basic framework for intelligent recommendations in the marketplace is shown in Figure 3.2 (Frontiers in big Data, 2023, 6:1157899). The electronic commerce intelligent recommendation system consists of nine main functional modules.

Data Collection Module. The task of this module is to collect item metadata, user metadata, user usage habits and other data required by each storage and record center of the mall. According to the way and length of information storage in each storage center, the data collection module must extract relevant data from the journal center by SFTP and then save it to the distributed file system for use in the recommendation system. Given the characteristics of numerous product types and complex product levels in commercial networks, a model based on commercial relationships is proposed. Some data of each business department, such as user transaction data, are stored separately [16]. The data acquisition module is required to set up multiple data collection points. An effective data transmission method is proposed to ensure the system's data transmission speed and reliability.

Data preprocessing module. This module is mainly divided into data ETL and data cleaning modules. The two are combined to complete the preprocessing of business data. Finally, standard and efficient transaction

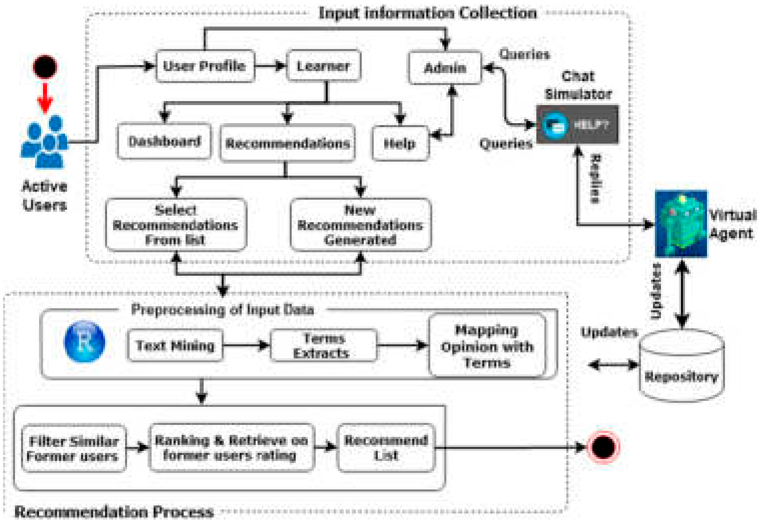

Input information Collection
User Profile
Learner
Admin
Queries
Chat Simulator
HELP?
Active Users
Dashboard
Recommendations
Help
Queries
Replies
Select Recommendations From list
New Recommendations Generated
Virtual Agent
Updates
Preprocessing of Input Data
R
Text Mining
Terms Extracts
Mapping Opinion with Terms
Updates
Repository
Filter Similar Former users
Ranking & Retrieve on former users rating
Recommend List
Recommendation Process

Fig. 3.1: Overall architecture design of e-commerce recommendation system.

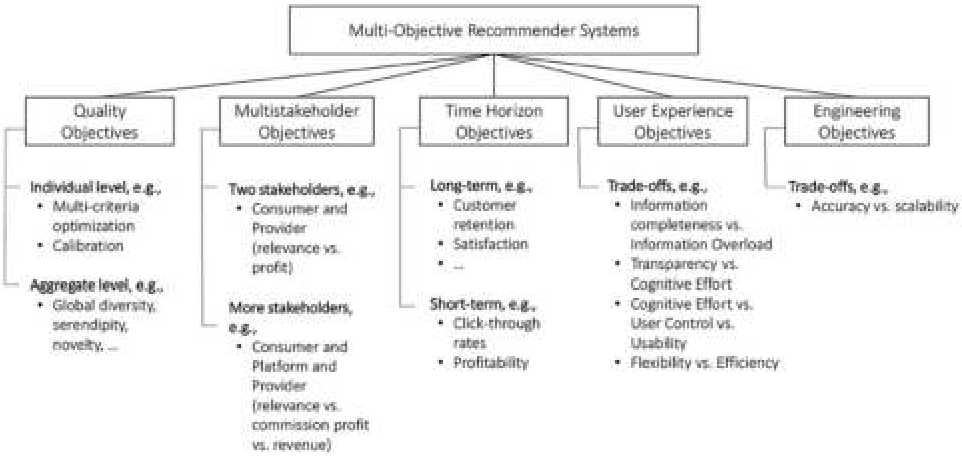

Multi-Objective Recommender Systems
Quality Objectives
Multistakeholder Objectives
Time Horizon Objectives
User Experience Objectives
Engineering Objectives
Individual level, e.g.,
• Multi-criteria optimization
• Calibration
Aggregate level, e.g.,
• Global diversity, serendipity, novelty, ...
Two stakeholders, e.g.,
• Consumer and Provider (relevance vs. profit)
More stakeholders, e.g.,
• Consumer and Platform and Provider (relevance vs. commission profit vs. revenue)
Long-term, e.g.,
• Customer retention
• Satisfaction
• ...
Short-term, e.g.,
• Click-through rates
• Profitability
Trade-offs, e.g.,
• Information completeness vs. Information Overload
• Transparency vs. Cognitive Effort
• Cognitive Effort vs. User Control vs. Usability
• Flexibility vs. Efficiency
Trade-offs, e.g.,
• Accuracy vs. scalability

Fig. 3.2: Functional architecture of e-commerce recommendation system.

data is generated. Among them, data ETL is mainly based on the data accumulated by the data acquisition module, the format requirements of the algorithm input for the data, and the processing of large-scale data expansion data extraction, transformation and loading. The data cleaning module mainly deletes or performs other operations on invalid data, missing field data and other illegal data in the process of data ETL to avoid the impact of illegal data on the recommendation quality of the algorithm.

Data mode and parsing module. This module aims to realize the analysis and modeling of ETL and cleaned data. Attribute analysis and canonical modeling of such data are carried out according to input data format and data type requirements. In the rule association algorithm, it is necessary to normalize each transaction in the transaction database [17]. How to define consumer behavior is an important question. This is because there are many scenarios where users purchase fewer items at a time. If every purchase is treated as a transaction, the links between items become small and complex, quantitative and efficient. In terms of user/commodity metadata information, it can be seen that the algorithm will analyze the user's preference degree based on the commodity's level, so the commodity's level attribute and user preference should be part of the algorithm

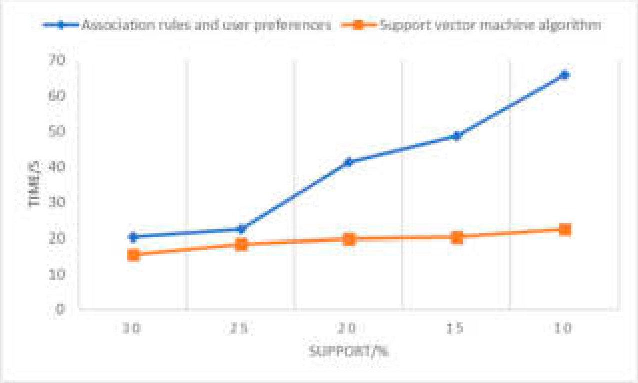


Fig. 3.3: Comparison of support vector machine algorithm and association rules algorithm in execution time.

modeling. In addition, for the recommendation system, the input model of the algorithm can be correlated with and sorted out from multiple data tables and generate specific shopping or browsing records. Finally, according to the required characteristics or other requirements, the corresponding data content and the corresponding association format are generated, and the information is saved to the unassociated database MongoDB cluster to achieve access to the recommendation algorithm.

Recommended Algorithm module. An e-commerce information recommendation method is proposed based on association rules and user preferences. The association rules algorithm presented in Chapter 2 is used in this module [18]. The method is divided into two stages: offline operation and online operation. In practical application, the offline operation mainly aims at the problem of high complexity and extended time in the solving process. The online calculation is based on the initial recommendation value and combined with the corresponding optimization algorithm to get the final recommendation value.

Distributed Recommendation Engine Module. This paper presents the implementation method of Hadoop based on the Jar package. It is done according to a particular order and needs. Select the corresponding recommendation list for the specific recommendation environment and target. The corresponding suggestions are provided to users by using the HTTP protocol.

Optimized Modules are recommended. The task of this module is to optimize the recommendation results generated by various recommendation algorithms to generate the final user recommendation results. The system mainly realizes the following three optimization aspects: initial recommendation filtering, ranking, and interpretation.

Recommended Data Interaction Interface Modules. This module aims to realize the adaptation of commodity recommendation and trading interface [19]. The data collection problem of the mall record center involves data type, data date, data channel and so on. The second is an interactive recommendation. The commercial system of the mall must respond to the recommendation system according to different recommendation scenarios, people and products. The recommendation system recommends the corresponding list to the user based on the data.

Recommended Evaluation Module. Three experimental methods evaluate the recommendation system. It includes an AIB test that evaluates the index offline, surveys users, and online trials [20]. In the practical implementation, the offline evaluation of the method is emphasized. Among them, there are mainly precision, recall rate and other indicators.

Basic Management Function Modules. A recommendation algorithm based on a data warehouse is proposed, and the algorithm is analyzed in detail. It includes two parts of: data storage and processing. The system management module includes interface management, log management, data maintenance, parameter configuration and rule intervention.

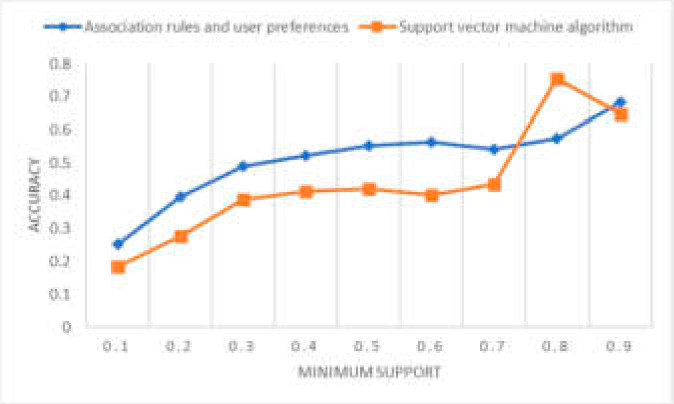


Fig. 4.1: Comparison of product recommendation accuracy between the SVM and association rules algorithms..

4. Application of user preference and association rule algorithm in product recommendation. The method uses a rotating database. When dealing with K-entry candidate sets, the search times of the original database can be reduced effectively, and the efficiency of data mining can be improved. Figure 4.1 compares the association rule algorithm and the support vector machine algorithm regarding execution time. As you can see in Figure 4.1, the relationship between user priority and the implementation time of the relevant rules is almost horizontal [21]. The support vector machine (SVM) 's running speed shows a significantly monotonically increasing trend. Compared with the traditional support vector machine method, the user preference and association rules method can obtain the initial candidate set only through a single scan of the initial transaction database. Subsequent processing, such as generating candidate sets, calculating support numbers, etc., no longer requires re-accessing the original database. This can significantly reduce the execution time and increase the operation's efficiency. The proposed method has obvious advantages over the SVM method in computation time.

If we want to discover all the e-commerce data, we need to do a lot of operations. Usually, the data in the database is divided into daily necessities, food, clothing, sporting goods, etc. There are many subdivisions under each category. This paper proposes a method based on association rules. This is also consistent with the general rules of what consumers buy. For example, when a customer searches for a digital camera, the site will display the digital camera the customer is searching for. Want to buy an SD card too? When consumers want red women's clothing, the site will suggest red leather bags and socks. Because it is only for product recommendation, only two frequent sets must be mined when data mining. In general, the number of recommendations that each project can provide is limited. When many items match the suggestions, you can select the most popular items from them [22]. The correctness of the method is tested by statistical analysis of the purchase records of different types of products. In this way, the precision comparison curve of the product can be obtained (Figure 4.2). From Figure 4.2, we can see that the accuracy of product recommendation of user preference and association rule algorithm is higher than that of support vector machine algorithm in terms of corresponding support degree.

5. Conclusion. The product recommendation function on the e-commerce platform is becoming increasingly prominent. A data mining method based on association rules and user preferences is presented. The algorithm can recommend products that users may be interested in more accurately based on sales history data. The results show that the method utilizes a rotating database and bit operation. The computational efficiency of this method is much higher than that of the conventional SVM method. Product recommendations are more accurate after introducing rule mining. But this approach has its limits. For example, if there is too much traffic on the site, the demand for storage will increase. This is the next step that needs to be addressed.

REFERENCES

[1] Zhang, H. N., & Dwivedi, A. D. (2022). Precise marketing data mining method of E-commerce platform based on association rules. Mobile Networks and Applications, 27(6), 2400-2408.

[2] Zhang, Y. (2021). Sales forecasting of promotion activities based on the cross-industry standard process for data mining of E-commerce promotional information and support vector regression. Journal of Computers, 32(1), 212-225.

[3] Loukili, M., Messaoudi, F., & El Ghazi, M. (2023). Machine learning based recommender system for e-commerce. IAES International Journal of Artificial Intelligence, 12(4), 1803-1811.

[4] Xie, C., Xiao, X., & Hassan, D. K. (2020). Data mining and application of social e-commerce users based on big data of internet of things. Journal of Intelligent & Fuzzy Systems, 39(4), 5171-5181.

[5] Massaro, A., Mustich, A., & Galiano, A. (2020). Decision support system for multistore online sales based on priority rules and data mining. Computer Science and Information Technology, 8(1), 1-12.

[6] Sjarif, N. N. A., Azmi, N. F. M., Yuhaniz, S. S., & Wong, D. H. T. (2021). A review of market basket analysis on business intelligence and data mining. International Journal of Business Intelligence and Data Mining, 18(3), 383-394.

[7] Bakar, W. A. W. A., Zuhairi, M. A., Man, M. U. S. T. A. F. A., Jusoh, J. A., & Triana, Y. S. (2022). a Critical Review of Deep Learning Algorithm in Association Rule Mining. J. Theor. Appl. Inf. Technol, 100(5), 1487-1494.

[8] Xu, B., Huang, D., & Mi, B. (2020). Research on E-commerce transaction payment system basedf on C4. 5 decision tree data mining algorithm. Computer Systems Science and Engineering, 35(2), 113-121.

[9] Ünvan, Y. A. (2021). Market basket analysis with association rules. Communications in Statistics-Theory and Methods, 50(7), 1615-1628.

[10] 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.

[11] 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.

[12] Wu, Z., Li, C., Cao, J., & Ge, Y. (2020). On scalability of association-rule-based recommendation: A unified distributed-computing framework. ACM Transactions on the Web (TWEB), 14(3), 1-21.

[13] Murthy, T. S., Roy, M. S., & Varma, M. K. (2020). Improving the performance of association rules hiding using hybrid optimization algorithm. Journal of Applied Security Research, 15(3), 423-437.

[14] Putra, A. A. C., Haryanto, H., & Dolphina, E. (2021). Implementasi Metode Association Rule Mining Dengan Algoritma Apriori Untuk Rekomendasi Promo Barang. CSRID (Computer Science Research and Its Development Journal), 10(2), 93-103.

[15] Zhao, Z., Jian, Z., Gaba, G. S., Alroobaea, R., Masud, M., & Rubaiee, S. (2021). An improved association rule mining algorithm for large data. Journal of Intelligent Systems, 30(1), 750-762.

[16] Han, Q. Y. (2020). The study of personalized recommendation algorithm in e-commerce system. International Journal of Education and Economics, 3(2), 1-6.

[17] Zhang, Y. (2021). The application of e-commerce recommendation system in smart cities based on big data and cloud computing. Computer Science and Information Systems, 18(4), 1359-1378.

[18] Perumal, S. P., Sannasi, G., & Arputharaj, K. (2020). REFERS: refined and effective fuzzy e-commerce recommendation system. International Journal of Business Intelligence and Data Mining, 17(1), 117-137.

[19] Tingting, W. (2020). Research on user access pattern mining based on web log. Asia-pacific Journal of Convergent Research Interchange (APJCRI), 6(8), 135-148.

[20] 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.

[21] Urbancokova, V., Kompan, M., Trebulova, Z., & Bielikova, M. (2020). Behavior-based customer demography prediction in E-commerce. Journal of Electronic Commerce Research, 21(2), 96-112.

[22] Rani, L. N., Defit, S., & Muhammad, L. J. (2021). Determination of student subjects in higher education using hybrid data mining method with the k-means algorithm and fp growth. International Journal of Artificial Intelligence Research, 5(1), 91-101.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Oct 9, 2023
Accepted: Oct 23, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1773–1780, DOI 10.12694/scpe.v25i3.2683

UNCREWED BOAT PATH PLANNING ALGORITHM BASED ON EVOLUTIONARY POTENTIAL FIELD MODEL IN DENSE OBSTACLE ENVIRONMENT

WEI ZHENG*AND XIN HUANG †

Abstract. In the trajectory planning of crewless ships, the artificial potential field method is commonly used. The results obtained using the classic potential field model for path design are not optimal and cannot fully meet the trajectory design requirements of uncrewed ships. This paper uses the evolutionary potential field model for trajectory planning. The evaluation formula of the potential path is combined with the differential evolution algorithm to evaluate and optimize the potential. A quadratic optimization smoothing algorithm is designed to limit the maximum turning angle of the uncrewed ship. Simulation experiments show that this method is effective and reliable.

Key words: Uncrewed boat; Path planning; Potential modeling; Differential evolution algorithm; Track optimization; Maximum turning angle limit

1. Introduction. The intelligent system of underwater uncrewed ships (UV) consists of motion control, sensors and communication. Among them, the path planning subsystem under the motion control system is the core for the autonomous navigation of uncrewed boats. Finding a new and effective trajectory optimization method is essential in this field.

A commonly used method in trajectory optimization is the artificial potential field method. This method has the characteristics of a simple model, fast calculation speed and smooth path. This method is currently the most widely used underwater uncrewed ship trajectory analysis method. However, the potential model itself has limitations. In practice, further improvement is often needed. Literature [1] uses various escape methods based on the existing potential method to study the problem that uncrewed ships are prone to falling into local minima during movement. In literature [2], the path obtained by integrating the potential field method with the grid model is safe and short but not smooth enough. Literature [3] proposed a tangent potential field method that can solve the local flutter problem of the path. However, the above methods are all based on the classic potential field theory. Because the optimality of the trajectory and rationality are not considered, the trajectory is only optimal from a certain angle. Since the maneuverability of uncrewed ships is not fully utilized, it cannot sufficiently meet the actual needs of uncrewed ships. There has been a lot of theoretical and practical work on the trajectory planning problem of uncrewed ships. Literature [4] proposes a new multi-target aircraft path planning method for the multi-target aircraft path problem. A path hazard evaluation method based on wind, waves and navigability is proposed, but the impact of other vessels and obstacles on path hazards is not considered.

Existing research results [5] have solved the problem of ship path re-planning in complex ocean environments by establishing a conflict hazard model between dynamic obstacles and ships based on the modified rate obstacle method. However, this algorithm only targets a single ship and ignores the influence of stationary obstacles on the course. Literature [6] uses a route planning method that combines A* with the dynamic window method to optimize road length and turning point problems. However, when making local map selection, the dynamic obstacles in the entire environment must first be judged before they can be adjusted, so real-time performance is lacking. Literature [7] presents a fast path optimization algorithm based on Fuzzy greedy. This algorithm can effectively control the growth direction and distribution density of various parts of trees in the

*1. School of Computer Science and Technology, Nanyang Normal University, Nanyang, Henan 473061, China; 2. Henan Engineering Research Center of Service and Guarantee for Intelligent Emergency, Nanyang, Henan 473061, China (Corresponding author, 15538757501@163.com)

†School of Computer Science and Technology, Nanyang Normal University, Nanyang, Henan 473061, China

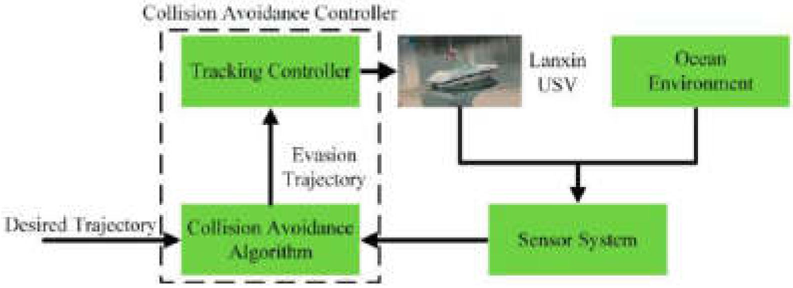


Fig. 2.1: Path collision avoidance coefficient.

configuration space, thereby avoiding various forms of stationary obstacles. However, this method requires high computing power and cannot avoid moving obstacles. Literature [8] uses an improved compressed sensing (CS) algorithm for trajectory planning. It adds a competitive filtering function to optimize the compressed sensing algorithm. Under stationary conditions, this algorithm can solve the local extreme value problem in the motion trajectory. However, due to the poor convergence of the algorithm, collisions in the motion trajectory cannot be effectively avoided. Through research on current uncrewed ship collision avoidance technology, few people can simultaneously conduct real-time collision avoidance against dynamic and static obstacles and static obstacles and discuss how to avoid collision when the two types of obstacles are approaching.

Differential evolution (DE) is a heuristic global optimization method that exploits population differences. This method is obtained by solving the Chebyshev polynomials in literature [9]. Compared with the classic evolution method, the DE method has the advantages of a simple model, fewer control parameters, and good robustness. In recent years, it has been used in areas such as optimization. This paper proposes an artificial potential energy field model based on differential evolution. The optimal properties of DE are added to the classic potential model. The evaluation formula of the geopotential channel is given, and the geopotential channel is modified using an evolutionary strategy. Then, a preliminary optimization of the geopotential trajectory was carried out. At the same time, the maximum turning angle of the uncrewed ship is included in the path planning as a limiting factor of its maneuverability. The smoothing method is used to realize the secondary optimization of the local track. Finally, simulation experiments were conducted under various circumstances to test the method's effectiveness.

2. Evolutionary potential field model for uncrewed ship path optimization.

2.1. Evaluation formula of a geopotential field path. Path optimization is to select the best path based on the path cost. The existing geopotential model lacks a mechanism for path evaluation, but this does not mean it is the best route. To this end, this paper establishes a channel evaluation formula based on the potential field method to compare and analyze various types of channels [10]. When planning the trajectory of an uncrewed ship, the collision avoidance, length, and smoothing coefficients should be comprehensively considered. The collision avoidance factor in the ship's trajectory is essential in ensuring the trajectory's safety. The motion performance of an uncrewed ship is affected by its length and smoothness. The collision avoidance factor is expressed by the sum of the distances between all access points in the area where the repulsive force acts and the corresponding obstacles (Figure 2.1 is cited in Appl. Sci. 2021, 11, 9741). It is the i obstacle in the environment. C_{inf} is the scope of action of the barrier. $O_j, \cdots, O_{j+m}$ refers to the route that falls within this obstacle exclusion circle.

The global avoidance factor of a path is

$$g_{\alpha} = \sum_{i}^{M} \sum_{j}^{j+m} c(W_i, O_j) \tag{2.1}$$

M is for all obstacles. m is the number of points on all routes that the obstacle passes under the repulsive effect. $c(W_i, O_j)$ represents the shortest distance between obstacle W_i and the j point G on route i. When

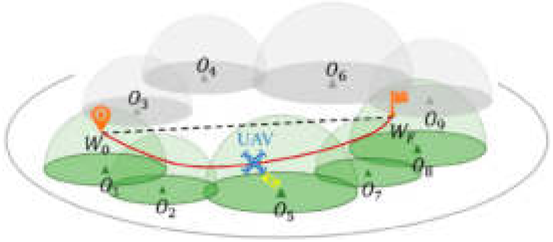


Fig. 2.2: Path Smoothing Factor.

H is larger, the distance between the entire route and the obstacle is more significant. The greater the ship's avoidance, the safer the route. The smoothness of the pathway is expressed by the sum of the spacing between all adjacent pathway points on the pathway (Figure 2.2 is cited in Electronics 2023, 12(11), 2358). O_i, O_{i+1}, O_{i+2} is three adjacent points. c is the length on the straight line from $O_i, O_{i+2}$to. From its structure, it can be seen that when the value of c is more significant, and the angle between the three points O_i, O_{i+1}, O_{i+2} is larger, the route becomes smoother in this area. The overall smoothness of the trajectory is

$$g_s = \sum_{i=1}^{N-2} c(O_i, O_{i+2}) \tag{2.2}$$

NAll access points. $c(O_i, O_{i+2})$is the length of the straight-line distance between $(O_i, O_{i+2})$and the path. As g_s increases, the change in curvature of the entire trajectory gradually decreases. There will be fewer unnecessary turns, making the operating system more efficient. The sum of the distances between adjacent path points approximates the path length. The length factor of this path is

$$g_1 = Ns \tag{2.3}$$

N is the total number of waypoints. s represents the traveling distance of the uncrewed boat. The results show that a smaller value of g_s results in a smaller total distance of the trajectory. Resources and time can be saved while completing this task while improving efficiency. A geopotential pathway efficiency evaluation formula was established based on the pathway as mentioned above efficiency factors.

$$g = \gamma g_\alpha + \delta g_s - \zeta g_1 \tag{2.4}$$

γ, δ, ζ represents the corresponding coefficient weight. It achieved a $\gamma + \delta + \zeta = 1$ grade. These three factors can be set according to their required weights in practical applications.

2.2. Trajectory smoothing based on maximum turning angle. Conventional robots can calculate the potential field force in the potential field through the potential field equation to determine the forward direction of each step of the robot and thus determine the entire path. The trajectory planning problem of uncrewed ships is very different from that of ordinary robots. Given the exceptional environment in which uncrewed ships travel in the ocean and the limitations of their maneuverability, it is often challenging to meet the requirements of engineering applications by only using potential-based methods for route design. This limitation must be considered when determining an uncrewed ship's actual trajectory. The maximum rudder angle is a significant parameter to measure the navigation performance of uncrewed ships [11]. The maximum swing distance is defined on the planned route. Assume that the current track point O_i and track point $\overrightarrow{O_iO_{i+1}}$ are given. The maximum steering angle of the uncrewed ship is κ . The step length is s. Then, the point O_{i+2} of the following straight line can only be limited to the 2κ arcuate AB with the angle DD (Figure 2.3 is cited in Scientific Reports, 2022, 12(1): 13997).

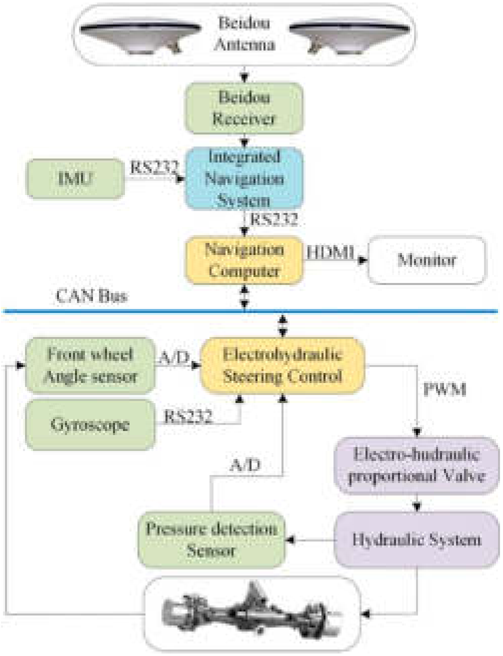


Fig. 2.3: The maximum turning angle of the uncrewed ship in the adjacent stairs.

It can be seen from Figure 2.3 that the maximum turning angle between the two steps is expressed as follows

$$\kappa = 2acr\tan\left(\frac{s}{D}\right) \tag{2.5}$$

Use the potential field method to set the path step size s. The minimum rotation radius D can be obtained through the rotation test of the uncrewed ship. The trajectory generated based on the evolved geopotential model has dramatically improved regarding collision avoidance coefficient, smoothness and trajectory length. However, the potential field method has limitations and cannot simultaneously meet the maximum rudder angle limit. This may result in excessive trajectory angles and unnecessary arcs. It cannot ensure the planned trajectory is reasonable and feasible [12]. This article gives a quadratic optimization method based on the maximum turning angle - the smoothing method. The three routes from the starting point to the end take three adjacent routes O_i, O_{i+1} and O_{i+2} respectively. The angle φ formed by these three points was judged to determine whether it could meet the requirements of the maximum control angle. If the angle κ is larger than the maximum turning angle, the point O_{i+1} passed halfway will be removed. At the same time, the route is updated until all points on the route comply with the maximum turning angle limit (Figure 2.4 is cited in Electronics 2023, 12(11), 2358).

2.3. Improved artificial potential field method. Khatib proposed the artificial potential field method in 1986. The artificial potential field method was used to plan the robot's trajectory.

Its core idea is to regard people's sensory space as virtual power. Obstacles or dangerous areas repel robots. The target point exerts a gravitational force on the robot, and the closer it is to the obstacle, the greater the repulsive force is, and the closer it is to the target point, the greater the gravitational force is. Their combined force then moves the robot towards the target. Establish gravitational and repulsive fields in artificial potential fields. Simplify robots and obstacles into dots to facilitate analysis [13]. The robotic arm and obstacles are converted into dots to facilitate calculation. This paper proposes a modified artificial potential field method to overcome the local minimum problem that is prone to occur in the classic artificial potential function method.

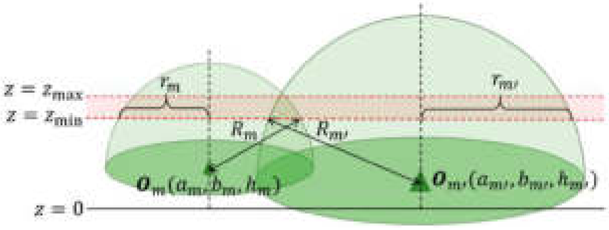

$z = z_{\max}$
$z = z_{\min}$
$z = 0$
r_m
$r_{m'}$
R_m
$R_{m'}$
$\boldsymbol{O}_m(a_m, b_m, h_m)$
$\boldsymbol{O}_{m'}(a_{m'}, b_{m'}, h_{m'})$

Fig. 2.4: Track smoothing method under maximum turning angle.

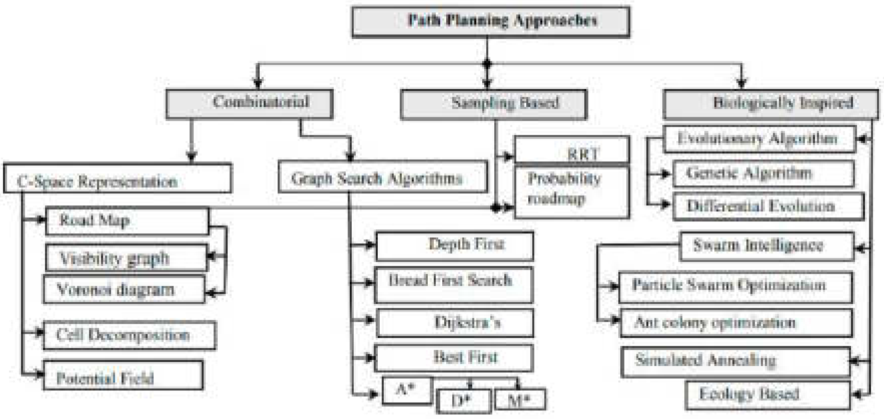


Fig. 2.5: Evolutionary potential field method for uncrewed ship trajectory planning.

The commonly used repulsive potential field is expressed as follows:

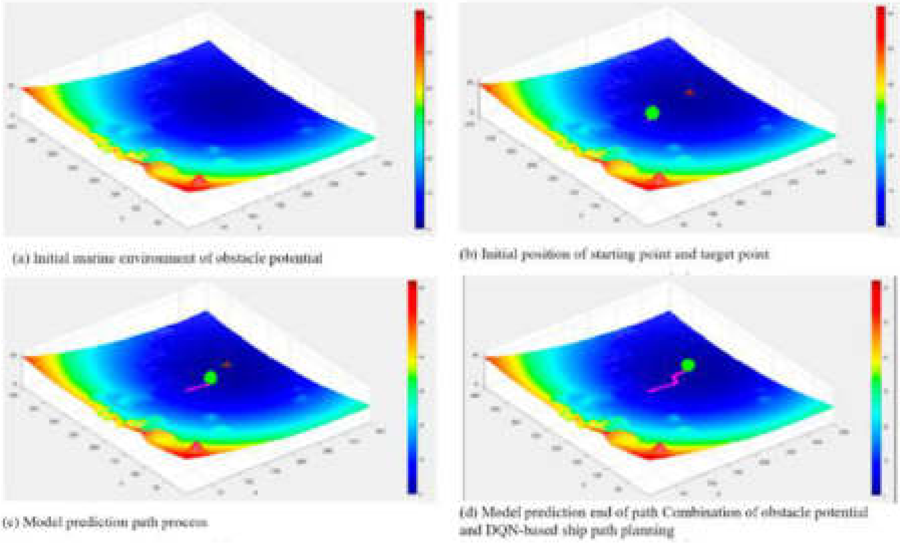
$$A_{rep}(w) \begin{cases} \frac{1}{2}\lambda_{rep}\left(\frac{1}{\omega w, w_{obs})} - \frac{1}{\omega_0}\right)^2 \omega^n(w, w_{goal}) & \omega(w, w_{obs}) \leq \omega_0 \\ 0 & \omega(w, w_{obs}) > \omega_0 \end{cases} \tag{2.6}$$

$\omega w, w_{obs})$ represents the distance between the uncrewed ship and the obstacle. ω_0 represents the area of influence of the repulsive potential field. λ_{rep} is a proportional increase in force. n is the distance coefficient by which the distance to the target can be adjusted. n is 2. The repulsive force formula can be obtained by applying a negative gradient to the repulsive potential field in equation (2.7).

$$G_{rep}(w) = \begin{cases} \lambda_{rep}\left(\frac{1}{\omega w, w_{obs})} - \frac{1}{\omega_0}\right) \frac{\omega^n(w, w_{goal})}{\omega^2(w, w_{obs})} & \omega(w, w_{obs}) \leq \omega_0 \\ 0 & \omega(w, w_{obs}) > \omega_0 \end{cases} \tag{2.7}$$

This paper uses the heading selected when avoiding obstacles at the fastest speed to replace the artificial gravity potential, preliminarily changing rates, and replacing rate obstacles in artificial potential fields to ensure that the uncrewed ship leaves the threat area.

2.4. Algorithm flow chart. An uncrewed ship trajectory planning method is proposed. Based on the known environmental factors, the route evaluation formula of the potential field method is established to evaluate the route [14] comprehensively. The parameters in the potential field method were evolutionarily optimized using the DE algorithm, and the optimal trajectory under the potential field method was obtained. Considering the potential field model's limitations, secondary smoothing is performed on the local path points of the potential field path. This achieves the maximum turning angle requirement for uncrewed ships. The specific process of this algorithm is shown in Figure 2.5 (picture quoted from/Proceedings of the 11th National Technical Seminar on Uncrewed System Technology 2019: NUSYS'19. Springer Singapore, 2021: 99-111).



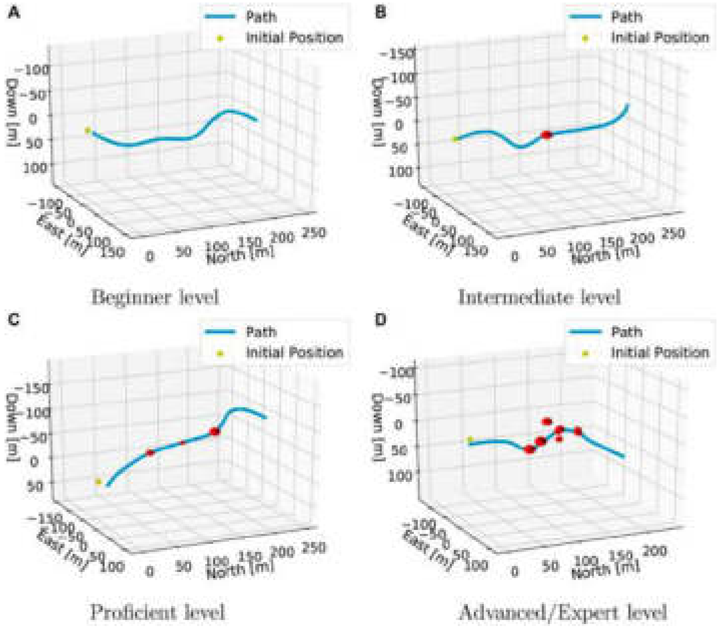
(a) Initial marine environment of obstacle potential

(b) Initial position of starting point and target point

(c) Model prediction path process

(d) Model prediction end of path Combination of obstacle potential and DQN-based ship path planning

Fig. 2.6: Simulation situation of traditional artificial potential field method.



Beginner level

Intermediate level

Proficient level

Advanced/Expert level

Fig. 2.7: Improved simulation results.

3. Simulation experiments and analysis. This paper tests the effectiveness of this method. According to the obstacle conditions encountered by the uncrewed ship in natural waters, regular obstacles are replaced with specific radius obstacles. And the setting is more realistic. The simulation results are shown in Figure 2.6 (picture quoted from Journal of Marine Science and Engineering, 2021, 9(2): 210.). As shown in Figure 2.6(a), the uncrewed ship is prone to collision when it is close to obstacles and far from the target point. This is because, in the classic ”artificial potential” method, the strength of the gravitational potential increases exponentially with the increase in distance, causing collisions [15]. It can be seen from Figure 2.7 (a) that no

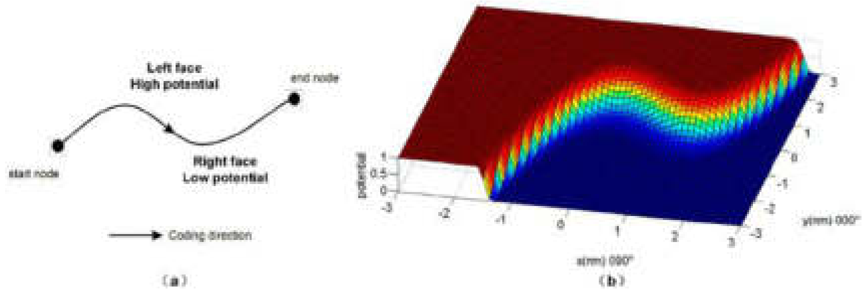


Fig. 3.1: Simulation situation before and after local minimum point situation improvement.

matter how far away the drone is from the target, as long as the gravitational potential field strength exceeds a certain distance, it will always be a specific value and safety accidents can be avoided.

From Figure 3.1, we can see that under the action of 0 net force, an uncrewed ship will stop at the minimum value of a certain point. The uncrewed ship was optimized using differential equations. Although there is still some oscillation, it finally leaves the target point.

4. Conclusion. The artificial potential field method is a local trajectory optimization method that is very suitable for uncrewed ships. However, this method has some flaws that make it problematic in practical applications. This paper proposes a gravitational potential field function and threshold and modifies them with the differential equation method. Finally, the obtained results were tested using the Matlab simulation program. The method proposed in this article can effectively improve the above problems. It can reach the target point safely, even in complex situations.

Acknowledgments. Research on Emergency Technology for Water Source Protection in the Mid-route of South-to-North Water Transfer Project Based on Machine Vision (No.: KJGG011).

REFERENCES

[1] Cao, Y., Cheng, X., & Mu, J. (2022). Concentrated coverage path planning algorithm of UAV formation for aerial photography. IEEE Sensors Journal, 22(11), 11098-11111.

[2] Vahid, S., & Dideban, A. (2022). Optimal path planning for uncrewed surface vehicle using new modified local search ant colony optimization. Journal of Marine Science and Technology, 27(4), 1207-1219.

[3] He, Z., Dong, L., Sun, C., & Wang, J. (2021). Asynchronous multithreading reinforcement-learning-based path planning and tracking for uncrewed underwater vehicle. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 52(5), 2757-2769.

[4] Zhu, M., Xiao, C., Gu, S., Du, Z., & Wen, Y. (2023). A circle grid-based approach for obstacle avoidance motion planning of uncrewed surface vehicles. Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment, 237(1), 132-152.

[5] Liu, J., Yang, J., Guo, Z., Cao, H., & Ren, Y. (2021). Simulation of uncrewed ship real-time trajectory planning model based on Q-learning. International Journal of Simulation and Process Modelling, 16(4), 290-299.

[6] Gu, Y., Goez, J. C., Guajardo, M., & Wallace, S. W. (2021). Autonomous vessels: state of the art and potential opportunities in logistics. International Transactions in Operational Research, 28(4), 1706-1739.

[7] Chowdhury, R., & Subramani, D. (2022). Optimal Path Planning of Autonomous Marine Vehicles in Stochastic Dynamic Ocean Flows Using a GPU-Accelerated Algorithm. IEEE Journal of Oceanic Engineering, 47(4), 864-879.

[8] Han, X., & Zhang, X. (2022). Multi-scale theta* algorithm for the path planning of uncrewed surface vehicle. Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment, 236(2), 427-435.

[9] Li, M., Mou, J., He, Y., Zhang, X., Xie, Q., & Chen, P. (2022). Dynamic trajectory planning for uncrewed ship under multi-object environment. Journal of Marine Science and Technology, 27(1), 173-185.

[10] Liu, G., An, Z., Lao, S., & Li, W. (2022). Firepower distribution method of anti-ship missile based on coupled path planning. Journal of Systems Engineering and Electronics, 33(4), 1010-1024.

[11] Zhang, X., & Shu, W. (2021). An obstacle avoidance route planning method for uncrewed surface vessel based on multi-objective evolutionary algorithm. Int. Core J. Eng, 7(3), 382-387.

[12] Wang, R., Miao, K., Li, Q., Sun, J., & Deng, H. (2022). The path planning of collision avoidance for an uncrewed ship navigating in waterways based on an artificial neural network. Nonlinear Engineering, 11(1), 680-692.

[13] Zhuang, X., Zhuang, S., Su, D., Du, S., & Liu, Y. (2023). TPS-Genetic Algorithm for Real-Time Sailing Route Planning based on Potential Field Theory. European Journal of Engineering and Technology Research, 8(3), 86-99.

[14] Yu, H., Murray, A. T., Fang, Z., Liu, J., Peng, G., Solgi, M., & Zhang, W. (2021). Ship path optimization that accounts for geographical traffic characteristics to increase maritime port safety. IEEE Transactions on Intelligent Transportation Systems, 23(6), 5765-5776.

[15] Radmanesh, M., Sharma, B., Kumar, M., & French, D. (2021). PDE solution to UAV/UGV trajectory planning problem by spatio-temporal estimation during wildfires. Chinese Journal of Aeronautics, 34(5), 601-616.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Oct 10, 2023
Accepted: Oct 30, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1781–1789, DOI 10.12694/scpe.v25i3.2695

RESEARCH ON INVENTORY CONTROL METHOD BASED ON DEMAND RESPONSE OF POWER BIG DATA

HUIXUAN SHI [1,2]*, ZHENGPING GAO [3]†, LI FANG [4]‡, JIQING ZHAI [5]§, AND HONGZHI SUN [5]¶

Abstract. The supply chain functions as a complex web, interconnecting various stakeholders such as suppliers, manufacturers, wholesalers, retailers, and ultimately, end consumers. Central to effective supply chain management is the meticulous handling of inventory, a critical factor influencing both cost efficiency and service excellence throughout the entire network. The management of inventory within this context extends beyond the confines of individual enterprises, bearing significance across the entirety of the supply chain. Consequently, achieving optimal performance necessitates a cohesive, holistic approach to management, aligning with the overarching objectives of the system.

Through dynamic data analysis of multiple types of power materials, a dynamic inventory control model for power materials is constructed to achieve optimal adjustment of inventory management. Ultimately, a multi-granularity inventory control method based on big data analysis of power warehousing is constructed, which effectively improves inventory management efficiency and reduces logistics management costs for enterprises.

Through big data analysis of power material warehousing, the characteristics of power material demand are excavated, and a classification method for power material demand is constructed to achieve an overall inventory control strategy for power materials.

The implementation results show that the controlled inventory can better meet the changing demand, thereby improving inventory management efficiency.

The multi-granularity inventory control method based on big data mining of warehousing combines inventory and multi-objective optimization theories, proves the applicability and feasibility of the proposed method, effectively improves inventory management efficiency, reduces logistics management costs for enterprises, and provides practical guidance and decision-making reference for improving the intensive management level of power production and maintenance materials.

Key words: Power big data; Demand side response; Inventory control; Intelligent system; Warehouse data mining

1. Introduction. The rapid development of China's power industry has driven the continuous expansion of the operation scale of power groups and the increase of the demand and variety of power materials. Effective inventory management can not only satisfy the normal production and operation needs of enterprises, but also prevent delays in production progress or power supply due to shortages, and lower enterprise costs and risks. Currently, power groups mainly use three inventory management modes: physical reserve, agreement reserve and dynamic turnover. Among them, physical reserve is the most direct and effective method, but long-term and large-scale reserve will consume a lot of funds, impede the flow of materials, and may cause material waste. Agreement reserve and dynamic turnover can enhance the efficiency of material use and reduce inventory costs. However, with the rapid expansion of power grid scale and the growing number of engineering projects, traditional distribution networks need to urgently transform and upgrade to smart distribution networks. Power grid business data exhibits explosive growth, and big data ”volume, variety and velocity” features become more evident. The traditional static power material supply chain will be gradually replaced by a highly flexible and data-driven supply chain. This poses higher demands for the precision of inventory control.

Inventory management of power groups is vital for enterprise operation and development. First, a reasonable inventory level can satisfy the normal production and operation needs of enterprises, and prevent delays in production progress or power supply caused by shortages. Second, efficient inventory management can lower enterprise costs and risks. By rationally determining the order quantity and order time, over-purchasing and

*1. State Grid Electric Power Research Institute, Nanjing, China; 2. Wuhan Nari Limited Liability Company of State Grid Electric Power Research Institute, Wuhan, China (Corresponding author, shi_huixuan@sina.com)

†State Grid Jiangsu Electric Power Co., Ltd., Nanjing, China

‡State Grid Fujian Electric Power Co., Ltd., Fuzhou, China

§State Grid Shandong Electric Power Co., Ltd., Jinan, China

¶State Grid Shandong Electric Power Co., Ltd., Jinan, China

material accumulation can be avoided, and enterprise procurement costs and storage costs can be reduced. Moreover, good inventory management can also enhance the service quality and customer satisfaction of enterprises. When enterprises respond to customer needs promptly and provide high-quality power materials, customer satisfaction will improve, thereby laying a foundation for the long-term development of enterprises.

Inventory control of power groups is essential for enterprise sustainable development. First, a reasonable inventory level can enhance enterprise operation efficiency. By optimizing inventory structure and increasing material turnover rate, enterprises can better cope with market demand fluctuations and improve operation efficiency. Second, effective inventory control can reduce enterprise environmental impact. Excessive physical reserve will lead to material accumulation and waste, increasing environmental burden. By rationally using agreement reserve and dynamic turnover, enterprises can lower inventory level and mitigate environmental impact. Furthermore, good inventory control can also strengthen enterprise competitiveness. In the context of intensifying market competition, by fine-tuning inventory management and precise inventory control strategy, enterprises can better satisfy customer needs and boost market competitiveness.

However, there are some drawbacks in the current inventory management. First, inventory control strategy is not refined enough, often only considering the demand and cost of materials, while neglecting other influencing factors, such as the quality, procurement cycle, transportation time of materials, etc. Second, there is a lack of tracking and monitoring of the usage of materials, resulting in the inability to adjust the inventory management strategy timely, and difficulty in achieving fine-grained inventory control. In addition, power groups also face challenges such as low level of informatization and insufficient data sharing, which hamper the improvement of inventory management level. Therefore, with the diversification and uncertainty of demand increasing, the traditional inventory management method is hard to cope with new challenges. Hence, exploring the multi-granularity inventory control method for demand-side response of power system has significant theoretical and practical value.

This paper analyzes the classification features of power materials, and models their demand characteristics. Based on historical consumption data, it designs a general inventory control strategy for power materials. For different categories of power materials, it constructs a dynamic inventory control strategy, and innovatively uses multi-objective optimization theory to implement a multi-granularity inventory control method. This research can help power enterprises develop reasonable inventory control strategies and optimize processes, thus enhancing their operation efficiency, lowering their costs and achieving sustainable development.

2. Related Work. The inventory control problem is a well-established topic in the field of optimization theory. According to literature [1], both the deterministic demand model and the random demand model for a given demand distribution have reached a relatively mature level. In terms of related demand, the theory of material requirement planning (MRP) and enterprise resource planning (ERP) have been developed based on these models, and have effectively solved these types of problems.However, with the progress of time, market demand has undergone significant changes, and the demand for many products exhibits non-stationary distribution characteristics. In the context of inventory control in a two-level supply chain, literature [2] compared the models of retailers holding inventory separately and a central inventory, and found that by establishing multiple retailers, inventory managers can reduce the total cost of inventory control and increase company income.In the aspect of cooperation and competition in the supply chain, literature [3] conducted a detailed analysis and used the game theory method to derive the optimal strategy. Literature [4] established a system that includes multiple retailers, multiple suppliers, and a warehouse model. Literature [5] examined the advantages and disadvantages of information sharing under the multi-retailer model. Finally, literature [6] provided a comprehensive summary of current supplier inventory management models.

In the domain of multi-level supply chain inventory control, literature [7] delved into the optimal purchase quantity problem under the supplier-managed inventory mode in the multi-level supply chain. Literature [8] investigated the issue of economical batch ordering of consumable inventory earlier. The literature [9] enhanced the dynamic inventory model originally proposed by Arrow-Harris-Marshak, factoring in the evolution of customer demand distribution over time and also examined the inventory control problem when demand information is unknown. In this context, the demand distribution remains stable with unknown parameters, but with some prior distribution knowledge. Literature [10] assumed that the prediction error follows a normal distribution and, from the perspective of satisfying service level constraints, proposed a non-stationary demand inventory control strategy based on heuristic algorithm.

Jun [11] has developed a mathematical model for the emergency material supply network, successfully reducing the dimensionality of both the cost function and time function. The composite of these two functions was then weighted, enabling a comprehensive analysis of the time and cost components of emergency material supply. The study also implemented an optimal material supply chain plan for emergency materials, recommended an emergency supply reserve model, and verified the feasibility of the mathematical model through specific case studies.

The demand sensitivity of inventory control is primarily manifested in the timely acquisition and dynamic optimization of demand. For instance, in a smart home setting, HEMS receives real-time electricity prices through a smart meter. With this price information, the EMC can utilize optimization algorithms to perform the scheduling of home appliances, thereby fulfilling one or more objectives from the consumer's perspective [12].

Based on the aforementioned research on adaptive inventory control, both domestic and international, it is evident that adaptive inventory control can effectively handle non-stationary demand and enhance system efficiency. Utilizing adaptive inventory control allows for the adaptive tracking of inventory control targets, resulting in an optimal model. However, most of the aforementioned documents are continuous demand-based and do not account for the adaptive inventory control strategy for discrete random demand. This paper aims to address the inventory control problem in non-stationary random demand supply chains and proposes an improved control strategy that can ensure the inventory model meets the given service level while reducing the bullwhip effect.

3. Demand-side response power material inventory control.

3.1. Demand-side classification method of electric power materials. To explore inventory control methods for different types of power materials in response to demand, it is essential to first grasp the properties of power materials and the fundamental workings of inventory management, specifically the classification of power supplies. Nevertheless, a noteworthy challenge lies in the classification of power materials due to their diverse types, varieties, and applications. Hence, the classification principles may vary. From different perspectives, power materials can be classified as follows:

(1) According to the nature of the project. Based on the nature of the project, power materials can be classified into three categories: overhaul materials, daily maintenance materials, and emergency repair materials. Among them, overhaul materials mainly refer to the replacement and planned overhaul of power supplies and equipment, daily maintenance materials mainly refer to the planned maintenance of non-overcurrent power materials, and emergency repair materials mainly refer to emergency repairs caused by an increase in power load, such as weather-related repairs, external forces, and general emergency repairs.

(2) By material category. Power materials can be divided into different categories based on their material type. These include raw materials (such as metal and building materials), overhead line equipment, substation equipment (including primary, secondary, and tertiary components), cables and accessories, tools, and office supplies (including non-production equipment), among others.

(3) According to the use of the team. Based on the team's usage, power materials can be classified into three categories: materials for the electric test shift, materials for the relay protection shift, and materials for the installation and succession shift.

(4) According to the supply method. Power materials can be classified based on their supply method. These include supply materials provided by the warehouse, direct transfer materials sent directly to the site by the supplier, and adjustment materials transferred from the warehouse.

(5) According to purchasing frequency. Based on the frequency of procurement, power materials can be divided into three categories: weekly procurement of materials, monthly procurement of materials, and quarterly procurement of materials.

(6) According to grid standards. Based on grid standards, power materials can be classified into large categories of materials (including primary and secondary equipment, installation materials, metal materials, and tools), medium-level materials (such as transformers, fuse boxes, and cables), and small materials (such as line angle iron cross arm, ordinary bolts, and cement products, commonly used fittings, etc.).

3.2. Extraction and analysis of demand-side features of power materials. The extraction and analysis of the demand characteristics of power materials serve as the cornerstone for both the classification of power materials and the design of strategies for inventory control of power materials. This holds significant theoretical value and practical relevance for managing power materials inventory. Electric power supplies typically exhibit a wide array of types, diverse specifications, varying degrees of standardization, distinct levels of planned demand for materials, diverse rules, varying degrees of volatility, and varying rates of update speed. Therefore, taking into account the specifics of warehouse power materials, we establish the demand characteristic model Q for devising strategies related to power materials inventory as follows:

$$\mathrm{Q}{=}\{\mathrm{M,U,P,V,R,S,L,G}\} \tag{3.1}$$

where M represents importance, U represents urgency, P represents periodicity, V represents universality, R represents regionality, S represents substitution, L represents liquidity, and G represents regularity.

(1) Key characteristics are employed to delineate the value attributes of power materials and the pivotal metrics for gauging service levels. Utilizing the ABC analysis principle, power materials are predominantly categorized based on their significance. The underlying principle of ABC analysis posits that ”the vital few outweigh the trivial many,” wherein goods are prioritized according to cumulative turnover as the yardstick for classification. Consequently, power materials are segmented into three distinct categories: A, B, and C. Category A comprises materials characterized by either substantial quantities or high demand, earmarked for intensified management and control, with a designated importance level of 3. Category B encompasses materials with moderate quantities or demand, managed and controlled through conventional means, with an assigned importance level of 2. Category C encompasses materials with limited quantities or demand, meriting straightforward management and control methods, with an importance level of 1. Thus, the importance value range, denoted as M, is {3, 2, 1}, signifying three tiers of significance: very important, generally important, and unimportant, respectively. In accordance with actual scenarios, guidelines are established as follows: materials of high value or substantial demand (approximately 80%) merit an importance level of 3, designating them as very important; materials of moderate value or demand (approximately 15%) with a moderate quantity of items (ranging between 20% and 50%) are assigned an importance level of 2, categorizing them as generally important materials; materials of lower value (approximately 5%) with a larger quantity of items (exceeding 50%) are deemed unimportant, warranting an importance level of 1.

(2) The concept of urgency is pivotal in delineating the intensity of customer demand for the prompt delivery of required materials. This characteristic is intricately linked to the prevalence of emergency repairs, serving as a barometer for measuring the immediacy of response required. In the context of business operations, each outbound record for electrical materials is accompanied by a unique order number, with the leading digit indicative of the specific purpose of the outbound materia – be it for overhaul, emergency repair, infrastructure enhancements, among others. Consequently, the urgency index (U) is derived through a nuanced assessment of these factors, encapsulating the imperative need for swift action in fulfilling customer requirements. The urgency index (U) can be quantified using the following formula:

$$\mathrm{U} = \mathrm{q}/\mathrm{X} \times 100\% \tag{3.2}$$

where q represents the quantity of materials dispatched for emergency repairs from the warehouse, while X denotes the total volume of materials dispatched overall. An urgency index (U) is established based on the ratio of emergency repair dispatches to total dispatches. When the urgency index surpasses 48%, it signifies a relatively high level of urgency, indicating a significant proportion of materials allocated for emergency repairs. Conversely, when the urgency index falls below 15%, it suggests a relatively low urgency level, indicative of a lesser portion of materials allocated for emergency situations.

(3) Periodicity encapsulates the temporal spacing between significant junctures within the power supply procurement process, extending from the preceding purchase to the current acquisition. This encompasses various temporal metrics, notably including the purchase lead time and the purchase cycle duration. These metrics serve as vital benchmarks in understanding the rhythm and cadence of procurement activities, offering insights into the timing and frequency of resource replenishment.

(4) The concept of universality pertains to the alignment between the existing condition of power materials stored in the warehouse and the technical specifications mandated by the State Grid Corporation. It essentially gauges the extent to which inventory adheres to standardized norms and requirements. This metric serves as a barometer of consistency and compliance within the inventory management framework, highlighting the efficacy of standardization practices in ensuring operational efficiency and regulatory compliance.

(5) Regional characteristics refer to the classification of power materials based on their specific applicability within certain geographic areas. This classification is determined by assessing the attribute characteristics and technical specifications of the materials. It involves discerning whether a material is specialized for a particular locale or if it serves a broader function across the entire Fujian Province or Fuzhou area.

(6) The substitution characteristics of electric power materials elucidate their capacity to fulfill similar functions across diverse applications, thereby determining their interchangeability. This evaluation goes beyond a simple comparison of item counts and delves into the nuanced aspects of functionality, technical compatibility, and operational efficacy. It entails a comprehensive analysis of whether materials can effectively substitute for one another under varying order requirements, reflecting the dynamic nature of their utility in different contexts.

(7) The liquidity characteristics of power materials encompass the velocity and frequency of their circulation within a defined timeframe. This includes metrics such as the monthly quantity of materials leaving the warehouse, the average monthly rate of material outflows, and the typical volume of each outgoing batch. Liquidity (L) can be quantified as follows:

$$L = \varepsilon_1 N + \varepsilon_2 T + \varepsilon_3 E \tag{3.3}$$

In the given equation, N denotes the total count of monthly shipments, T signifies the mean monthly shipment frequency, and E represents the average volume of each individual shipment. The variables ε_1, ε_2, and ε_3 denote the corresponding weight coefficients, the optimization of which will be conducted as part of the experimental process.

(8) The regularity characteristic assesses the presence of patterns or regularities in historical demand for power materials. This evaluation typically involves statistical analysis, such as computing the coefficient of variation for consumption data. The regularity (G) can be quantified using the following formula:

$$\mathrm{G} = \sigma/\mathrm{x} \times 100\% \tag{3.4}$$

where σ is the standard deviation and x is the average.

In terms of material utilization, maintenance materials exhibit low regularity. A regularity index (G) exceeding 80% indicates a high degree of consistency, while values between 50% and 80% suggest moderate regularity, and G below 50% indicates weak regularity. Materials demonstrating strong regularity facilitate more accurate prediction of future demand, enabling precise inventory planning

3.3. Inventory control strategy for demand side response. The power materials supply chain encompasses suppliers, multi-tiered warehouses, and project sites, each with distinct demand characteristics necessitating varied distribution modes and inventory management strategies. Special regulations govern aspects such as urgency and safety. Typically, materials are sourced from suppliers and routed through specialized warehouses before reaching project sites. General planned materials are subject to a two-tier inventory control strategy, with suppliers provisioning regional distribution centers for centralized inventory management. These centers sort and distribute materials to front-end warehouses, employing circular distribution or cross-docking methods for secondary distribution. Framework agreement materials and emergency procurement items are expedited directly from suppliers to project sites to minimize construction delays.

Prior to formulating a comprehensive inventory control strategy for electric power materials, it is essential to categorize the inventory based on the degree of item overlap. Item overlap refers to the extent to which various specifications of each material, serving different purposes, coincide. This categorization yields two distinct classifications: planned inventory and order inventory. Planned inventory comprises materials with a high overlap ratio, necessitating a predetermined quantity to be stored in warehouses throughout the year. Conversely, materials with low overlap ratios are categorized as order inventory, requiring procurement in accordance with specific orders placed.

Based on the distinct demand patterns of electric power materials, the division between planned inventory and order inventory can be further refined. This paper proposes an inventory control strategy rooted in demand-side responsiveness, deploying diverse control algorithms tailored to different material demand profiles. For planned inventory materials with irregular demand, particularly emergency repair items, a dynamic inventory control approach is advocated. This entails dynamic adjustment of inventory levels at specific intervals based on the demand characteristics of different power materials. Control strategies are implemented by flexibly modifying inventory upper and lower limits. In contrast, for planned inventory materials with consistent demand patterns, such as overhaul materials, the MRP (Material Requirements Planning) system is recommended for inventory control. Under this replenishment strategy, materials of greater importance are subject to more frequent inspections, typically on a weekly basis, while less critical items undergo less frequent assessments, perhaps on a monthly basis.Similarly, order inventory can be stratified based on demand regularity. For materials with sporadic demand, adopting the Vendor Managed Inventory (VMI) replenishment model[13] is advocated. In this approach, a collaborative relationship is established with suppliers, allowing them to manage inventory based on a master plan formulated by the power material company. This facilitates timely and accurate demand information transmission, ensuring precise inventory control. Conversely, for order inventory materials characterized by consistent demand, the Just in Time (JIT) distribution model is proposed. JIT distribution, orchestrated from a coordination center, emphasizes timely delivery of the appropriate products in the exact quantities specified by the customer. Leveraging small-batch and multi-frequency delivery methods, JIT distribution aims to minimize inventory and waste while accommodating the diverse and personalized needs of customers. To enhance supply chain stability, inventory control strategies may involve forging alliances with nearby suppliers or agents and entering into contractual agreements, such as framework agreements, to ensure timely power supply distribution.

(1) Real-time inventory management approach. The real-time inventory management paradigm operates on a swift and continuous inventory inspection cycle, ensuring constant oversight of stock levels. Through meticulous parameterization, lower and upper inventory thresholds are meticulously defined. Upon reaching the lower threshold, an automatic replenishment signal is swiftly dispatched to restore inventory levels to the upper limit. This approach is particularly well-suited for the management of materials characterized by their high value, liquidity, and urgency, ensuring optimal inventory levels to meet dynamic demand fluctuations.

(2) Material requirement planning inventory management. The inventory management approach within the material requirement planning system typically adopts an extended inspection cycle, commonly on a monthly basis. Replenishment occurs upon reaching the inspection point, provided that the safety stock threshold is not breached, resulting in the addition of inventory up to the upper limit. This strategy is generally applicable to materials of moderate value and liquidity, encompassing supplies with varying levels of urgency.

(3) Vendor managed inventory stock refill strategy. Vendor Managed Inventory (VMI) involves suppliers managing the inventory of users with the users' consent. This approach relies on close collaboration between the parties to ensure efficient material delivery. The supplier assumes responsibility for determining inventory levels and devising strategies to maintain them. VMI is particularly effective in the relationship between suppliers and distributors at the first tier of the supply chain. Typically, the inventory control strategy in VMI operates on a monthly inspection cycle. Replenishment signals are promptly triggered when inventory levels fall below the designated threshold. The replenishment quantity can be calculated using predetermined models. VMI is generally suitable for managing high-value, low-liquidity, and emergency materials of superior quality.

(4) Just in Time strategy.. The fundamental principle of Just in Time (JIT) is to deliver the precise materials to the designated location exactly when they are needed, while considering excess inventory as wasteful. Typically, the JIT control strategy aligns its inspection cycle with the occurrence of demand, often employing a monthly inspection frequency. Replenishment is triggered upon reaching the inspection point, provided that the safety stock threshold is not breached. The replenishment quantity can be calculated using predetermined models. This approach is generally suited for materials of moderate value, with relatively low turnover and urgency.

$$\mathrm{E} = \mathrm{e} \times \mathrm{MAD} \tag{3.5}$$

$$\mathrm{Q} = \mathrm{q} \times \mathrm{n} \tag{3.6}$$

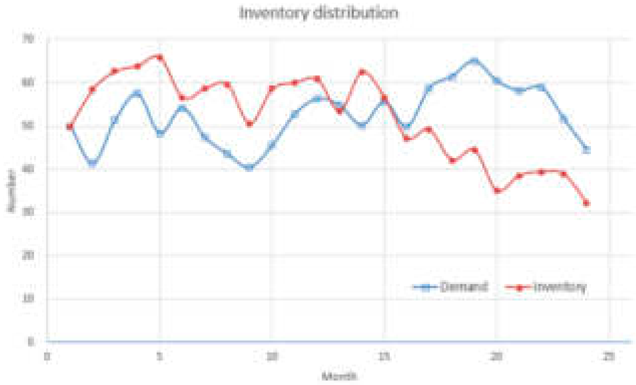


Fig. 4.1: Inventory and demand curve.

Through a meticulous examination of demand characteristics and cluster analysis, an adaptive inventory control strategy is devised for each category of power materials. However, it's essential to recognize that the demand landscape and clustering outcomes of power materials are dynamic and subject to change. Consequently, the company undertakes annual reassessments to refine its inventory control strategy in response to evolving analysis results and real-world conditions. This underscores the remarkable adaptability and responsiveness of the inventory control strategy grounded in demand characteristics analysis, facilitating a more nuanced and effective approach to inventory management that better aligns with the company's operational needs and objectives.

4. Experiment. In the simulation example, a two-level supply chain consisting of a warehouse and multiple suppliers is considered. The MATLAB simulation test is mainly performed when the demand distribution of power projects is unknown, and then the experimental model when the demand distribution is known is compared with the former. A certain type of commonly used materials is used as the object to experiment to evaluate the pros and cons of the model established in this chapter.

4.1. Inventory simulation when demand distribution is known. When the demand is known, the demand information is predicted by exponential smoothing once. Assuming that the demand per unit time is D~N(50,102), the actual initial demand is the random number generated in the interval [45,55], the predicted initial demand is 50. The model runs for 24 cycles (months). And the simulation runs 20 times to get the average data. The model time is taken as the unit time.

Figure 4.1 shows the actual data of demand and inventory when the demand distribution obeys the normal distribution, where the box represents the demand number, and the dot represents the inventory number. We can see that there are 13 cases where the inventory is greater than the demand, and the average demand overflow is 21%; there are 10 times when the inventory is less than the demand, and the average under-demand rate is 24%. Excessive inventory will lead to occupation of inventory space, thereby resulting in waste; insufficient inventory will lead to unsatisfied demand. Therefore, it is not good if the inventory is too large or too small.

4.2. Inventory control simulation. Through the dynamic inventory control strategy, MRP system inventory control strategy, VMI inventory replenishment control strategy, and JIT control strategy proposed in this paper, multi-granular inventory control is carried out. The results are shown in Figure 2 as follow.

Figure 4.2 reflects that the controlled inventory can better meet the changes in demand, thereby improving the efficiency of inventory management.

5. Conclusion. In light of the expanding landscape and escalating intricacies of power projects, the imperative for sophisticated inventory control mechanisms has been underscored. Within this evolving context, the paradigm of multi-granularity inventory control, particularly from a demand-centric perspective, has emerged as a pivotal domain of interest, heralding its agility in real-time assessment and adaptability. This discourse

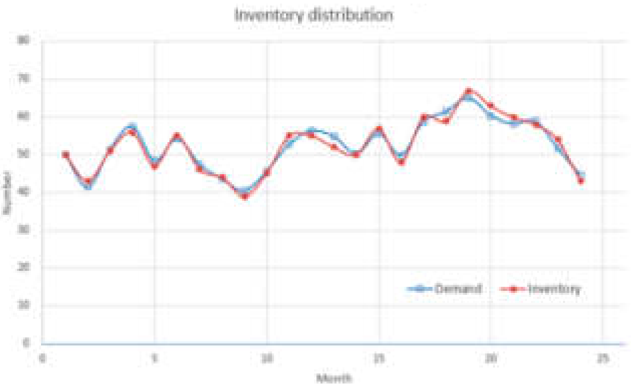


Fig. 4.2: Curve of multi-granularity inventory control results.

unveils a pioneering methodology in multi-granularity inventory control, synthesizing tenets from inventory management and multi-objective optimization theory. Beyond mere conceptualization, empirical validation of this framework not only attests to its practical viability but also illuminates actionable insights and strategic imperatives aimed at optimizing the management of power production and maintenance materials.

As we navigate the next phase of this scholarly endeavor, the spotlight shifts decisively toward the conception and realization of a dynamic inventory management architecture bespoke to the exigencies of power materials. This ambitious undertaking encompasses a comprehensive reassessment of dynamic safety inventory thresholds, meticulously calibrated against a nuanced backdrop of supply cycles, lead times, historical safety stock data, and procurement trajectories. In parallel, the augmentation of this framework entails the dynamic modulation of parameters within the demand management framework, ultimately coalescing into the construction of a dynamic, multi-tiered inventory management infrastructure. This adaptive infrastructure, characterized by its responsiveness to evolving demand dynamics, not only serves to mitigate inventory bottlenecks but also acts as a bulwark for ensuring the resilience and continuity of the power grid, all while effectuating optimal capital deployment across inventory holdings.

Acknowledgements. This paper is supported by State Grid Corporation Headquarters Technology Project (Research on Key Technologies of Modern Intelligent Warehousing System and Logistics Equipment, No:1400-202118268A-0-0-00).

REFERENCES

[1] Raza S A. Supply chain coordination under a revenue-sharing contract with corporate social responsibility and partial demand information[J]. International Journal of Production Economics, 2018, 205: 1-14.

[2] Mishra U, Wu J Z, Sarkar B. Optimum sustainable inventory management with backorder and deterioration under controllable carbon emissions[J]. Journal of Cleaner Production, 2021, 279: 123699.

[3] Nazari L, Seifbarghy M, Setak M. Modeling and analyzing pricing and inventory problem in a closed-loop supply chain with return policy and multiple manufacturers and multiple sales channels using game theory[J]. Scientia Iranica, 2018, 25(5): 2759-2774.

[4] Li Z, Hai J. Inventory management for one warehouse multi-retailer systems with carbon emission costs[J]. Computers & Industrial Engineering, 2019, 130: 565-574.

[5] Colicchia C, Creazza A, Noè C, et al. Information sharing in supply chains: a review of risks and opportunities using the systematic literature network analysis (SLNA)[J]. Supply chain management: an international journal, 2018.

[6] Song J S, van Houtum G J, Van Mieghem J A. Capacity and inventory management: Review, trends, and projections[J]. Manufacturing & Service Operations Management, 2020, 22(1): 36-46.

[7] Gharaei A, Pasandideh S H R, Akhavan Niaki S T. An optimal integrated lot sizing policy of inventory in a bi-objective multi-level supply chain with stochastic constraints and imperfect products[J]. Journal of Industrial and Production Engineering, 2018, 35(1): 6-20.

[8] Sebatjane M, Adetunji O. Three-echelon supply chain inventory model for growing items[J]. Journal of Modelling in Management, 2019.

[9] Kaijun L, Wang Yuxia W. Research on inventory control policies for nonstationary demand based on TOC[J]. International Journal of Computational Intelligence Systems, 2010, 3(sup01): 114-128. [10] Bookbinder. J. H, Tan..

[10] J. Y, Strategies for the Probabilistic Lot-sizing Problem with Service-level Constraints [J], Management Science, 1988, 34(9): 1096-1108.

[11] J. Y, Strategies for the Probabilistic Lot-sizing Problem with Service-level Constraints [J], Management Science, 1988, 34(9): 1096-1108.

[12] Ahmad S, Ahmad A, Naeem M, et al. A compendium of performance metrics, pricing schemes, optimization objectives, and solution methodologies of demand side management for the smart grid[J]. Energies, 2018, 11(10): 2801.

[13] Kaasgari M A, Imani D M, Mahmoodjanloo M. Optimizing a vendor managed inventory (VMI) supply chain for perishable products by considering discount: Two calibrated meta-heuristic algorithms[J]. Computers & Industrial Engineering, 2017, 103: 227-241.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Oct 15, 2023
Accepted: Oct 25, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1790–1798, DOI 10.12694/scpe.v25i3.2705

APPLICATION OF SPORTS VIDEO IMAGE ANALYSIS BASED ON FUZZY SUPPORT VECTOR MACHINE

LICHENG GAO [1*] AND YAWEN ZHAO [2†]

Abstract. Sports video image has always been a hot topic in sports video processing. The theoretical and experimental analysis of digital image noise reduction technology is a challenging topic. In this paper, a sports video denoising algorithm is designed by combining the excellent characteristics of curvilinear transformation theory and fuzzy support vector machine. Firstly, the image with noise is curvilinear, and the conversion coefficient is obtained. Then, according to the distribution characteristics of the system noise, the system parameters are divided into space, and the system learning features are constructed. The fuzzy classification of high-frequency curves is realized using the adaptive threshold denoising method. Then, the noise reduction coefficient is reconstructed by the curve-wave method to obtain the processed image. The simulation results show that this method can overcome the pseudo-Gibbs effect effectively and suppress the noise well. This algorithm has a good application prospect in sports video image processing.

Key words: Sports video; Image denoising; Curve-wave transformation; Fuzzy support vector machine; Adaptive threshold

1. Introduction.

Image noise reduction is always a hot topic in image processing. Firstly, the image is denoised to provide more accurate information for later image processing (edge detection, object recognition). Second, the development of image noise reduction technology provides a new way for image restoration, image segmentation and other image processing and analysis. In recent years, noise reduction of digital images has become a hot issue in image processing.

The existing denoising methods can be divided into two sides: filter denoising, conditional random field denoising, anisotropic denoising, non-local average denoising and statistical model denoising. The bidirectional filter can eliminate noise and keep boundary information well, but it cannot suppress Speckle noise well and can easily cause excessive smoothness [1]. The feature selection in the conditional random field modelling method has excellent flexibility, and it is unnecessary to give accurate prior data. However, the algorithm faces two difficulties: first, the solution of the energy function of the conditional constraint factor must meet certain conditions, and the global minimization of the conditional constraint factor is an NP-hard problem under most conditions. The second is finding the appropriate energy function to obtain the ideal global minimum [2]. The advantage of the anisotropic diffusion image denoising method is that the image can be denoised without affecting the target features, but this algorithm has some problems, such as over-processing and over-dependence on the target features. Non-local mean methods take advantage of the properties of repetitive structures in the image to remove noise, but their objective quality and visual effect are generally worse than other denoising methods. In recent years, scholars at home and abroad have studied various image noise reduction algorithms based on statistical models [3]. This kind of algorithm mainly uses the inter-scale and intra-scale correlation to reduce noise, but the experiment proves that this kind of algorithm cannot get good noise reduction results.

Sports video is a kind of multimedia data with voice and image as the main content. This data contains some complex semantics. Classification of semantic information in videos is the first step for users to obtain information. Most of the existing video classification methods are for the lower level of image classification, such as shot type, shot action and so on. The use of high-level semantic features can make the classification broader. Some scholars have studied the multi-modal information extraction method. Some scholars classify it according to the rules of basketball [4]. The most significant disadvantage of the lack of semantic guidance for the underlying features is that the moving images cannot be expressed efficiently, and the helpful information

[*] College of P.E, Jiaozuo Normal College, Jiaozuo, Henan 454000, China

[†] College of Arts, Jiaozuo Normal College, Jiaozuo, Henan 454000, China (Corresponding author, mini@jzsz.edu.cn)

contained in the moving images cannot be accurately evaluated. For example, the overall feature has a lot of listener noise. Neither from the motion point of view nor from the static point of view can ensure the correct recognition result. There are sports venues, athletes, referees, a large number of spectators and other subjects. Parsing domain rules can obtain the corresponding relationship between low-level attributes and high-level attributes of actions. Therefore, a set of meaningful feature libraries is obtained. This method can effectively classify video objects by an auxiliary classifier.

In this paper, a sports video denoising algorithm is designed by combining the excellent characteristics of curvilinear transformation theory and fuzzy support vector machine [5]. The region features of the target in motion video are obtained by analysing the region boundary and attention mode in motion video. The motion video is classified using a support vector machine (SVM) meta-classifier.

2. Fuzzy support vector machine algorithm.

2.1. Standard Support Vector Machine Algorithm (SVM). The basic idea of SVM is to train the data into nonlinear high-dimensional data. The best categorical hyperplane is constructed in high-dimensional nucleon space, consistent with the VC dimension. The method combines test risk with confidence intervals [6]. The risk classification function with the maximum limit is obtained according to the principle of reducing structural risks to the maximum extent and weighing them. Let the separable sample set be $(u_i, f(u)), i = 1, \cdots, n, u \in R^n, f(u) \in \{+1, -1\}$. The goal of SVM is to create a single-class hyperplane. Separate samples from two different classes to achieve maximum classification spacing. It can get the quadratic optimal problem:

$$\begin{aligned} &\min\left(\frac{1}{2}||\delta||^2 + \lambda\sum_{i=1}^{n} e_i\right) \\ &s.t. \quad f(u)_i[\delta u_i + \sigma] - 1 + e_i \geq 0 \\ &e_i \geq 0 \end{aligned} \tag{2.1}$$

The Lagrange multiplier $\varphi_i (i = 1, 2, \cdots, n)$ is introduced to obtain the duality of the formula:

$$\begin{aligned} &\max\sum_{i=1}^{n}\varphi_i - \frac{1}{2}\sum_{i,j=1}^{n}\varphi_i\varphi_j f(u)_i f(u)_j \mu(u_i, u_j) \\ &s.t. \quad 0 \leq \varphi_i \leq \beta \\ &\sum_{i=1}^{n}\varphi_i f(u)_i = 0 \end{aligned} \tag{2.2}$$

e_i is for the relaxation variable, β represents the penalty factor. $\mu(u_i, u_j)$ is the kernel function. Therefore, the decision function can be obtained as

$$g(u) = sign\left[\sum\nolimits_{i=1}^{n}\varphi_i f(u)_i \mu(u_i, u) + \sigma\right] \tag{2.3}$$

2.2. Multi-classification support vector machine. A multi-class classification method based on SVM is proposed. It can be broadly divided into two categories:

1. One-to-one: SVM classifier is used to learn between two types of samples to obtain samples of $t(t-1)/2$ category. The number of classifiers for each category is $t - 1$. Each classifier decides based on its classification criteria when forecasting a new sample. And vote for the corresponding category. The category with the highest number of votes is the category of the unknown sample.
2. One to many: Support vector machine classification distinguishes each classification from others in order. A total of t classifiers are generated. When forecasting uncertain samples, divide the samples into a class with the maximum determination function value.

If the number of samples is too large, the method's learning efficiency and classification efficiency will be reduced, and both have unidentifiable areas [7]. Then, the improvement method, including ECOC, DAG, etc., is implemented. Choosing the appropriate code book in the ECOC method is a complex problem. There is a shortcoming of the DAG algorithm. Its learning efficiency is not high when dealing with many classes. The selection of the root node significantly influences the classification effect.

2.3. Fuzzy SVM.

2.3.1. Fuzzy SVM Overview. SVM uses the optimal hyperplane to divide data into two opposite categories. In practice, however, each sample cannot be classified into a specific category. Sample and classification have a certain fuzziness. Therefore, many people apply Fuzzy theory to SVM and put forward Fuzzy SVM. This method is an improvement on the classical SVM method [8]. The main idea of this method is to introduce fuzzy theory into the SVM model. The value of the penalty weight varies depending on the size of the sample. In this way, different samples play different roles in the construction process. The sample containing noise or abnormal data is given a low weight to reduce the impact of noise and abnormal data on it.

When using a fuzzy support vector machine for classification, it differs from the traditional support vector machine in the representation of training samples in the following ways: In addition to the characteristics and class recognition of samples, it adds a membership degree to each part of the training. Suppose A training sample set is represented by

$$\begin{aligned}&(u_i, f(u)_i, \eta(u_i)),\\ &i = 1, \cdots, n, u \in R^n,.\\ &f(u) \in \{+1, -1\}\end{aligned}$$

where $\eta(u_i)$ stands for the degree of subordination and $0 < \eta(u_i) \leq 1$. Because the slave attribute $\eta(u_i)$ represents class confidence. e_i is the category error term in the objective function of SVM. So $\eta(u_i)e_i$ is the error with the weight. The optimal classification surface obtained is the optimal solution of the following objective functions:

$$\begin{aligned}&\min \left[\frac{1}{2}||\delta||^2 + \beta \sum_{i=1}^{n} \eta(u_i)e_i\right]\\ &s.t \quad f(u)_i[\delta u_i + \sigma] - 1 + e_i \geq 0\\ &e_i \geq 0\end{aligned} \tag{2.4}$$

The Lagrange multiplier $\varphi_i(i = 1, 2, \cdots, n)$is introduced to obtain the duality of the function:

$$\begin{aligned}&\max \sum_{i=1}^{n} \varphi_i - \frac{1}{2} \sum_{i,j=1}^{n} \varphi_i \varphi_j f(u)_i f(u)_j \mu(u_i, u_j)\\ &s.t \quad 0 \leq \varphi_i \leq \beta\eta(u_i)\\ &\sum_{i=1}^{n} \varphi_i f(u)_i = 0\end{aligned} \tag{2.5}$$

Thus, the decision function is

$$g(u) = sign\left[\sum_{i=1}^{n} \varphi_i f(u)_i \mu(u_i, u) + \sigma\right] \tag{2.6}$$

The comparison of formula (2) and (5) shows that SVM and Fuzzy SVM are different in terms of restrictions. In the SVM model, β is a customizable penalty factor. The algorithm can punish the wrong samples. The more β there is, the heavier the penalty factor. It has an extensive limit system of right and wrong samples and short intervals of class surfaces. When the β value decreases, the SVM will ignore more samples [9]. Thus, a class surface with larger boundary spacing is obtained. Set β more significant in Fuzzy SVM. If all dependencies $\eta(u_i)$ are set to 1, the algorithm reduces the error probability to a normal SVM. In this paper, a fuzzy SVM classification method based on degree $\eta(u_i)$ is proposed to make the classification results more accurate. The lower affiliation has less effect on learning outcomes. Fuzzy SVM has better anti-noise performance than ordinary SVM.

2.3.2. Fuzzy weight calculation. The most crucial thing in using fuzzy technology is determining its attribution function. Different membership degree functions will have different effects on the processing result of the algorithm and the difficulty of the algorithm implementation. Establishing membership functions that can objectively and accurately reflect various uncertainties in the sample is necessary. There is no uniform regulation on establishing membership functions [10]. In practical applications, different problems are often solved from different angles. Many scholars have done some research on this issue. However, most existing FSVM algorithms use the distance from the sample point to the classification center as an evaluation index. More and more people accept this algorithm because of its slight complexity and high robustness. However, in the current fuzzy support vector machine method, the membership degree is mainly measured based on the distance between the sample and the class center. This paper will use the fuzzy membership degree measurement method to determine the fuzzy membership degree $\eta(u_i)$. Set u_0 as the center of the classification. s is the radius of the class representing the system. s is determined by the following formula:

$$s = \max_i ||u_i - u_o|| \tag{2.7}$$

So, the degree of membership of each sample is

$$\eta(u_i) = 1 - \frac{||u_i - u_o||}{s} + \xi \tag{2.8}$$

To prevent the case of $\eta(u_i) = 0$, ξ is pre-set to a tiny constant, $(\xi > 0)$.

3. Fuzzy weight calculation. A set of adaptive image noise reduction methods based on Fuzzy SVM is designed, and good noise reduction results are obtained.

3.1. Fuzzy SVM is used for image denoising in the curvilinear wave domain. The working procedure of the classification method based on curve-wave transformation and Fuzzy SVM is as follows:

Step 1. Perform curve-wave processing on noisy images. FTW processing of the original noise image can extract a single low-frequency and several high-frequency bands. The results show that the noise's main component is in the bending region's higher frequency band.

Step 2. Generate feature vectors and train them. The noise distribution and space law characteristics are combined, and the feature vector is constructed [11]. The specific methods are as follows:

(1) The high-frequency subband coefficients of curved waves are initialized with binary tables.

$$\beta_\mu(i,j) = \begin{cases} 1, & |\beta_\mu(i,j)| > \xi \\ 0, & |\beta_\mu(i,j)| \le \xi \end{cases} \tag{3.1}$$

$\beta_\mu(i,j)$ represents the high-frequency subbands in the frequency range of the curve. $\xi = \varphi v + \sigma$ is the threshold function [12]. It is used when building binary tables to select sub-band widths with higher frequencies. v represents the variance of the noise. Where $v = Median(|\beta_t|)/0.6745, \beta_t$ is the maximum subbandwidth of the frequency in the bending frequency range. In addition, several test pictures were used to detect. This paper selected $\varphi = 0.475, \sigma = -3.75$.

(2) Construct a continuous road map and feature vector according to the spatial law. Once the binary table is formed, the function of some high-frequency subbands can be determined according to its spatial distribution law. The coefficient belongs to a subpart of a spatial feature. This paper analyzes the frequency spectrum characteristics of high-frequency signals by the continuous channel method. This is the identification of two high-frequency subbands. If there is a continuous channel between the two, it means that the parameters of the two labeled high-frequency subbands are spatially correlated. After processing the existing data, determine whether each parameter has spatial characteristics and whether each parameter is the noise point [13]. When the continuous path value exceeds a particular threshold value S, it is considered a spatial feature. Otherwise, it is regarded as a noise point.

(3) Select the local High subband coefficient with 0 continuous channels in the most extended continuous channel as the eigenvector.

(4) The obtained feature vectors and labels are used for SVM learning. The training mode of FSVM is obtained by an input feature vector and training sample.

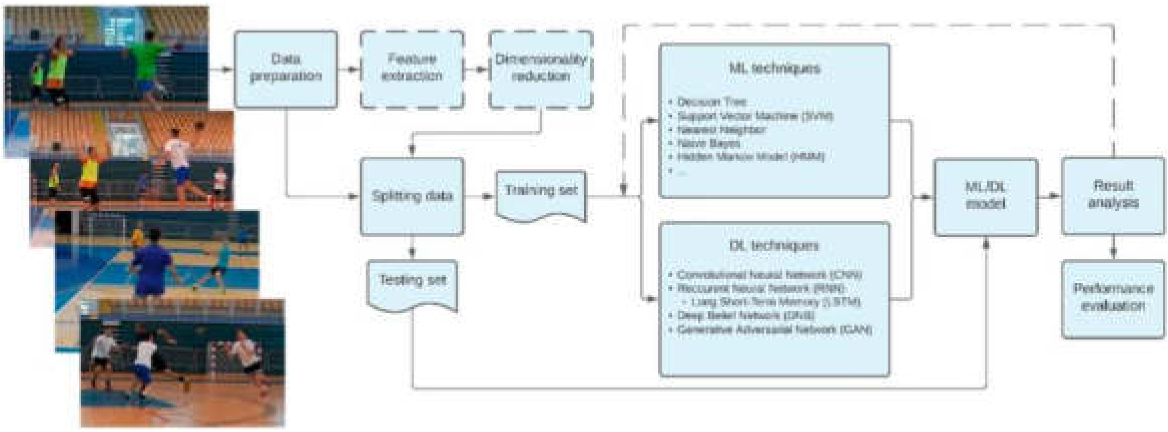

Data preparation
Feature extraction
Dimensionality reduction
Splitting data
Training set
Testing set
ML techniques
• Decision Tree
• Support Vector Machine (SVM)
• Nearest Neighbor
• Naive Bayes
• Hidden Markov Model (HMM)
• ...
DL techniques
• Convolutional Neural Network (CNN)
• Recurrent Neural Network (RNN)
◦ Long Short-Term Memory (LSTM)
• Deep Belief Network (DNB)
• Generative Adversarial Network (GAN)
ML/DL model
Result analysis
Performance evaluation

Fig. 3.1: Sports video image processing flow based on fuzzy support vector machine.

Step 3. Remove the high-frequency subband coefficient that contains noise.

Step 4. The fuzzy SVM learning model can divide the high-frequency subband parameters into two types: noisy and noiseless. The output marked 0 of the fuzzy support vector machine training models is set as the noise factor [14]. The output represented by one is noiseless. A noise reduction method based on an adaptive threshold is proposed according to the characteristics of curve wave conversion and the distribution characteristics of noise. Here is the detailed calculation method:

$$S_\mu = \begin{cases} \frac{\hat{v}_\varepsilon^2(\mu)}{\hat{v}_g(\mu)} \frac{1}{\theta_\mu} & \hat{v}_g(\mu) \neq 0 \\ \max(|\beta_\mu(i,j)|) & \hat{v}_g(\mu) = 0 \end{cases} \tag{3.2}$$

S_μ gives the adaptive threshold used for denoising the high-frequency subband coefficients in the K scale and D direction [15]. The local contrast of the image corresponding to the high-frequency subband in the Kscale D direction is expressed by$\theta_\mu.\theta_\mu = \frac{v_\mu}{\mu_\mu}, v_\mu$ represents the variance of the curve's waveform. μ_μ represents the average value of the curve waveform.

$$\hat{v}_\varepsilon(\mu) = Median(|\beta_\mu|)/0.6745 \tag{3.3}$$

The standard deviation of the original image signal is estimated, as shown below

$$\hat{v}_g(\mu) = \sqrt{\max(\hat{v}_\beta^2(\mu) - \hat{v}_\varepsilon^2(\mu), 0)} \tag{3.4}$$

Step 5. Reconstructing the high-frequency subband coefficient using the curve method. The reconstructed noise reduction image can be obtained by calculating the reverse bending wave in the existing high-order frequency band. Fig. 3.1 shows a sports video image's fuzzy support vector machine processing flow.

3.2. The number of categorical characteristics of intervention domain knowledge. The information on sports categories can be normalized at a higher semantic level by extracting the long shot fragments in the pre-processing process [16]. There are pronounced differences among various sports types in the characteristics of the field, the characteristics of the sports object, and the ratio of similar characteristics between the sports field and the sports object. You can see the differences in each feature in Table 3.1 and Figure 3.2.

This paper chooses FAR, MR, MG, MB, VR, VG, VB, NA and AAR as feature vectors. FAR refers to the ratio of the area on the field to the picture. MR, MG, and MB represent the average colour of the field RGB colour space. VR, VG, and VB are the colour variances of the site. The NA value is N (Aa). It refers to the number of contestants. AAR is expressed as S (Aa)/S (F), which is the ratio of the average area occupied by the players to the area of the field and the ratio of the area of the sports field to the screen (FAR) feature has

Table 3.1: The average value of the tuple feature vector.

Category	**Eigenvector mean**								
	FAR	MR	MG	MB	VR	VG	NB	NA	AAR
Soccer	0.921	83.313	103.719	103.656	14.293	69.869	14.454	12.339	0.003
Volleyball	0.347	199.813	100.990	91.594	55.423	15.432	15.845	8.658	0.047
Tennis	0.887	81.615	111.885	169.990	26.367	19.095	45.594	3.243	0.015
Table tennis	0.068	86.844	84.073	175.646	11.701	15.855	65.927	2.408	0.541

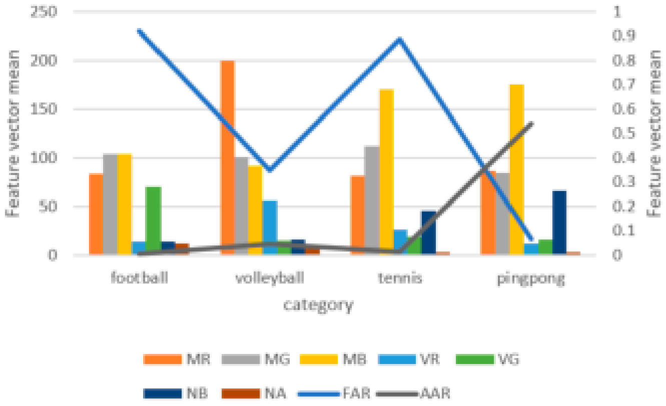


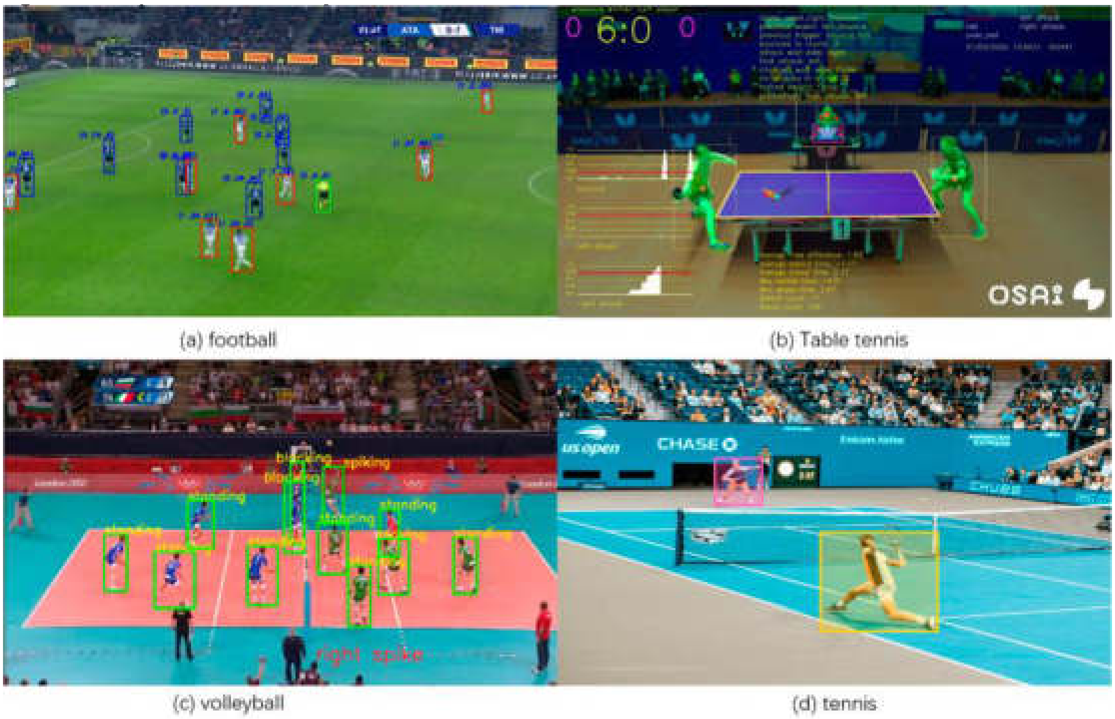
Fig. 3.2: 9 Mean value of tuple feature vector.

a good recognition for games with significant differences in the proportion of the field such as football, hockey and badminton. The venue's colour can make a big difference in outdoor and indoor competitions. Individual sports, such as tennis and rugby, are better distinguished by the characteristics of the crowd and the ratio of the players' pitches ({NA, AAR}) than team sports. Using the SVM method to analyse a 9-tuple vector, most motion videos can be classified and distinguished accurately.

4. Analysis and comparison of experimental results. In this study, football, volleyball, tennis, table tennis, etc., are taken as the research objects, and video collection cards are used to obtain data from TV sets. A picture of a football contains 352x288 pixels. Other video images are 640x480 pixels. The hardware used in the test was a P4-1.4G CPU, a PC with 256 MB of storage space, and a Windows 2016 operating system.

4.1. Feature Extraction. Fig. 4.1 shows the effect of segmenting the arena and players using the method described in Section 1 of this paper. The fields in the video are marked with gray rectangles, while the competitors are marked with a white rectangle. The experimental results show that this algorithm is the most effective for football target detection. The location and size of the players and the course can be marked. In the detection of table tennis, the Gaussian distribution weight is selected to make the central position of the Dalian Tong district more prominent [17]. This is the correct position of the ping-pong ball. The process of a volleyball match is the same as that of a table tennis match. But the method couldn't tell how many people there were, so their portraits had a lot of overlap. The amount of detection area is estimated in the experiment, but the result is unsatisfactory. In the tennis video, the colour of the court and the courtside is similar, resulting in a large screen selection area.

4.2. Classification and classification test of moving images by support vector machine. SVM inputs are based on features in base nine extracted from video clips of football, volleyball, tennis, etc. The relevant parameters of the support vector machine are obtained by cross-checking. After cyclic testing, the final SVM model is obtained. The classification results of four types of video clips are shown in Table 4.1 and Figure 4.2, respectively.



(a) football

(b) Table tennis

(c) volleyball

(d) tennis

Fig. 4.1: Semantic object detection and segmentation of sports video.

Table 4.1: The average value of the tuple feature vector.

Data set	Training episodes	Number of test sets	Cross-validation accuracy / %	Test accuracy / %
Soccer	620	290	98.9	97.4
volleyball	1080	610	98.1	96.8
tennis	419	255	95.4	94.1
Table tennis	910	646	97.4	95.9

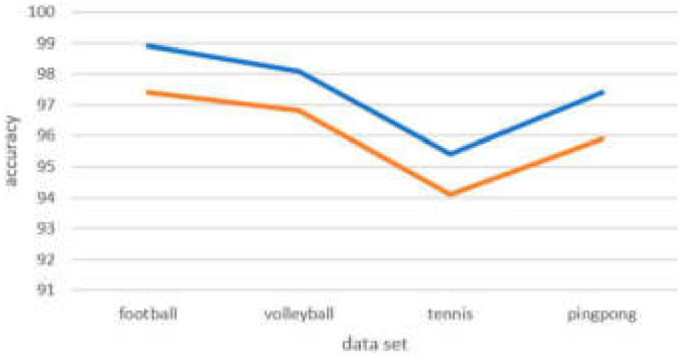


Fig. 4.2: Results of the grading test.

Table 4.2: Comparison of classification results.

Method	**Average accuracy / %**
Semantic feature vector + Fuzzy support vector machine	96.07
Acoustic properties + element hybrid model	88.90
Mobile feature + HMM+ Serial feature strategy	93.98
Action characteristics + speech characteristics + fuzzy matrix + neural network	95.21
Action characteristics + speech characteristics + HMM+ synthesis probability multiplication	95.95

Table 4.2 compares the more commonly used classification methods and classification methods. The experimental results show that the algorithm has a high average accuracy. In addition, the method in this paper can also increase the number of meta-support vector machines to improve the classification of video so that it has better adaptability.

5. Conclusion. A motion video automatic recognition model is established. The existing domain knowledge matches the features and semantics in the image to get the semantic features in the image. Fuzzy SVM is used to reduce the noise of curved waves. A method based on bending waveform is proposed. The spatial characteristics of each parameter are analysed according to the distribution characteristics of noise. The constructed feature vector is input into Fuzzy SVM. The fuzzy classification of high-frequency curves is realized using the adaptive threshold denoising method. Then, inverse curve wave processing obtains the reconstructed image after noise reduction. Experimental results show that the proposed method is superior to the asemantic supervised classification method. This method has good anti-noise performance. Adding a meta-classifier can allow the system to be expanded to classify more kinds of motion video.

REFERENCES

[1] Sontayasara, T., Jariyapongpaiboon, S., Promjun, A., Seelpipat, N., Saengtabtim, K., Tang, J., & Leelawat, N. Twitter sentiment analysis of Bangkok tourism during COVID-19 pandemic using support vector machine algorithm. Journal of Disaster Research,2021; 16(1): 24-30.

[2] Junaid, M., Sohail, A., Turjman, F. A., & Ali, R. Agile Support Vector Machine for Energy-efficient Resource Allocation in IoT-oriented Cloud using PSO. ACM Transactions on Internet Technology (TOIT),2021; 22(1): 1-35.

[3] Li, B., & Xu, X. Application of artificial intelligence in basketball sport. Journal of Education, Health and Sport, 2021;11(7): 54-67.

[4] Ren, J., Zhang, B., Zhu, X., & Li, S. Damaged cable identification in cable-stayed bridge from bridge deck strain measurements using support vector machine. Advances in Structural Engineering,2022; 25(4): 754-771.

[5] Wu, S., Chen, X., Shi, C., Fu, J., Yan, Y., & Wang, S. Ship detention prediction via feature selection scheme and support vector machine (SVM). Maritime Policy & Management,2022; 49(1): 140-153.

[6] Qiu, S., Hao, Z., Wang, Z., Liu, L., Liu, J., Zhao, H., & Fortino, G. Sensor combination selection strategy for kayak cycle phase segmentation based on body sensor networks. IEEE Internet of Things Journal,2021; 9(6): 4190-4201.

[7] Zheng, Y. G., Zhang, H. S., & Wang, Y. Q. Stripe detection and recognition of oceanic internal waves from synthetic aperture radar based on support vector machine and feature fusion. International Journal of Remote Sensing, 2021;42(17): 6706-6724.

[8] Khanam, F., Hossain, A. A., & Ahmad, M. Electroencephalogram-based cognitive load level classification using wavelet decomposition and support vector machine. Brain-Computer Interfaces,2023; 10(1): 1-15.

[9] Ghazali, N. F., Sanat, N., & As' ari, M. A. Esports Analytics on PlayerUnknown's Battlegrounds Player Placement Prediction using Machine Learning. International Journal of Human and Technology Interaction (IJHaTI),2021; 5(1): 17-28.

[10] Liu, L., Chen, X., & Wong, K. C. Early cancer detection from genome-wide cell-free DNA fragmentation via shuffled frog leaping algorithm and support vector machine. Bioinformatics,2021; 37(19): 3099-3105.

[11] Lee, K., Wu, X., Lee, Y., Lin, D. T., Bhattacharyya, S. S., & Chen, R. Neural decoding on imbalanced calcium imaging data with a network of support vector machines. Advanced Robotics,2021; 35(7): 459-470.

[12] Hu, J., & Zhang, H. Support vector machine method for developing ground motion models for earthquakes in western part of China. Journal of Earthquake Engineering,2022; 26(11): 5679-5694.

[13] Afrifa, S. Cyberbullying detection on twitter using natural language processing and machine learning techniques. International Journal of Innovative Technology and Interdisciplinary Sciences, 2022;5(4): 1069-1080.

[14] Radhika, A., & Syed Masood, M. Premier League Table Prediction Using Machine Learning Algorithms. Webology, 2022;19(1): 6379-6395.

[15] Abdelbaky, A., & Aly, S. Human action recognition using three orthogonal planes with unsupervised deep convolutional neural network. Multimedia Tools and Applications, 2021;80(13): 20019-20043.

[16] RangaNarayana, K., & Rao, G. V. Action recognition in low resolution videos using FO-SVM. Indian Journal of Computer Science and Engineering (IJCSE),2021;12(4): 1149-1162.

[17] Barbon Junior, S., Pinto, A., Barroso, J. V., Caetano, F. G., Moura, F. A., Cunha, S. A., & Torres, R. D. S. Sport action mining: Dribbling recognition in soccer. Multimedia Tools and Applications,2022; 81(3): 4341-4364.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Oct 17, 2023
Accepted: Nov 11, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1799–1806, DOI 10.12694/scpe.v25i3.2711

A MULTI-SENTENCE MUSIC HUMMING RETRIEVAL ALGORITHM BASED ON RELATIVE FEATURES AND DEEP LEARNING

YELIN ZHANG*

Abstract. This project will study a fast retrieval method for music humming speech recognition based on sentence features and deep learning. The method proposed in this paper can realize the fast extraction of songs. According to the characteristics of the natural pause mode of the song, the song database and the song fragments provided by the user are divided into different sentences. The deep learning algorithm of BDTW is used to calculate the similarity of the song's pitch, and users can set matching conditions according to their preferences. It can identify the most significant differences between music fragments and the order of queries in the database. Then, a retrieval method of a music database based on DIS is proposed. It can shorten the acquisition time. Experiments show that the algorithm can recognize humming songs quickly and efficiently.

Key words: Related features; Deep learning; The songs hum; BDTW algorithm; Search algorithm

1. Introduction. Digital music has gradually developed and formed a popular model in recent years. Now, China's online music industry is growing. Total mobile music sales in China reached 5.987 billion yuan in the first three months of this year. The number of mobile music users has reached 836 million. Major Internet sites currently store millions of electronic music. This makes it more complicated for users to search, retrieve, and find relevant songs. Now, the central system in the industry is in the form of a manual query. Such searches are often based on metadata such as artist name, song title, etc. The few semantically expressive features are musical styles. Usually, only the words mentioned above or other written information can be used when searching for music. In addition, the system only provides the most basic music recommendations and personalized service. It does not depend on the content of its musical signal. The system also provides only one user feature document generation method based on user information. In such a system, the user is represented as a vector, a measure of the click-through rate or number of plays. With the help of a vector library, we can realize the joint recommendation of similar users and similar songs. The same can be done by describing documents according to semantically labelled users, finding similar users or finding a music document. Realizing the automatic, accurate and fast location of music objects is an urgent problem in music notation.

Reference [1] describes a method that indirectly uses user-listening behaviour to generate semantic description documents. They took advantage of users' listening habits and metadata extracted from users' private music profiles. Music services like last.fm are available, as are reviews, biographies, journals, and music-related RSS links on the World Wide Web. Literature [2] uses joint filtering to implement music recommendations. It has an excellent statistical effect on popular songs. But there is a big downside to this. For example, the required user click rate, social tags and other metadata are missing for non-hit songs. Some signs using sound-based advice can ameliorate this problem.

At present, object-oriented analysis and fusion technology are rare. But this approach has its drawbacks. Each of them uses tone and rhythm to convey their meaning. This sound information is low-level and cannot be directly translated into high-level semantic information. Some recent experiments prove that the semantic gap can be bridged by corresponding work in semantics. The "semantic gap" is mainly reflected in the lack of correlation between the low-level feature information in speech and the semantic information of the human brain. Much research has been done on the current music content search methods. Literature [3] gives a new kind of music ontology aiming at the problem of the "semantic gap" problem. A new semantic ontology of music is constructed by using the characteristics of the lower and higher levels of music. Literature [4] provides a semantic model based on spatial context. The technology of situational cognition defines the semantic meaning.

*Zhengzhou Technology and Business University, Zhengzhou 450000, China (Corresponding author, 20100001@ztbu.edu.cn)

It constructs a text-oriented music retrieval system that includes emotion and non-emotion. Semantic features are used to analyse the similarity of search results. Reference [5] uses Dirichlet mixed modes to propose a token-based semantic polynomial. Some background information is added when annotating songs to improve the accuracy of annotations.

When labelling music, literature [6] introduces a method to synthesize the characteristics of various sound sources. Two phonological features and two social features are used in this method. Literature [7] proposes an algorithm based on social labelling and compares it with vector space models, singular value matrix factorization, non-negative matrix factorization, and probabilistic latent semantic analysis. Simulation results show that this method has a better effect than other methods. Literature [8] combines fuzzy music scene features and sound features into the ACT algorithm. The accuracy of content-based music retrieval is improved by using the fuzzy music scene characteristic with expressive semantics. Literature [9] uses an algorithm that uses the compressed string as the time characteristic of music to calculate the similarity of music. The semantic description-based query proposed in the literature [10] is a more natural query method. A good music information search model is developed. However, the biggest obstacle of this algorithm is the lack of clearly labelled, public and open, heterogeneous labelled song data sets. A CAL500 database was constructed in reference [11]. The study fuses users' listening habits with song labels to obtain semantic associations between lyrics and music. Multiple guide classifiers are used to label the pattern. By training the CAL500 library, a song search method based on the CAL500 library is obtained. This mode cannot only tag a new track but also perform text queries and corresponding return values for multiple tracks.

Humming query is a significant component of song retrieval. This search method retrieves the nearest k songs from a single lyric. The research of song extraction technology is relatively backward. A pioneering exploration of humming queries was made in literature [12]. The researchers used a rough string alignment algorithm to complete the humming search. Literature [13] adopts Dynamic Time Organization (DTW) technology to perform complete sequence matching between humming songs and other tracks in the database. Literature [14] uses the N-gram backward index structure in the music data mined by the topic, thus improving retrieval efficiency. Literature [15] uses subsequence as a matching method to solve the problem of missing specific notes or syllables in humming fragments. In addition, the rapid development of string-matching technology in recent years has also played an excellent guiding effect for music search. Literature [16] proposes a multi-dimensional music sequence matching technique. His research uses a combination of sentence features and deep learning to identify songs. The method proposed in this paper can realize the fast extraction of songs. According to the characteristics of the natural pause mode of the song, the song database and the humming fragments provided by the user are divided into different song sentences. The deep learning algorithm of BDTW is used to calculate the similarity of the intonation of humming songs, and users can set matching conditions according to their preferences. It limits the maximum difference between music fragments and the order of queries in the database.

2. Humming music retrieval system framework based on sample semantics. In the humming music retrieval system based on sample semantics, the focus of work is no longer to extract the features of intensity, pitch, beat and melody from the audio signal. First, the recognition accuracy cannot be significantly improved because of the "bottleneck" in recognizing sound level. Secondly, there is no clear research on the relationship between the meaning of humming songs and the sound level. Meaning is often determined by more than just one or a few sound features [17]. The traditional way of acquiring speech features based on speech information is faced with the problem of the "semantic gap." The construction of a music search system based on the example semantics is described in Figure 2.1 (image cited in Bioengineering 2023, 10(6), 685). This project intends to use deep learning methods to find the relationship between raw signals and semantics. The music in the established music library is classified and the semantic analogy is made. This search method can get the user's search intention more naturally and accurately.

3. Music retrieval algorithm based on humming. A sentence-based contour feature library of humming music is established. Calculating the front and back notes can determine the piece's profile. Its composition is "sentence length, first note interval, last note interval, the difference between adjacent notes interval."

According to the above characteristics, the tunes of multiple sentences are searched, and the corresponding characteristics are required to match the tunes of the songs. It requires the production of a candidate set of

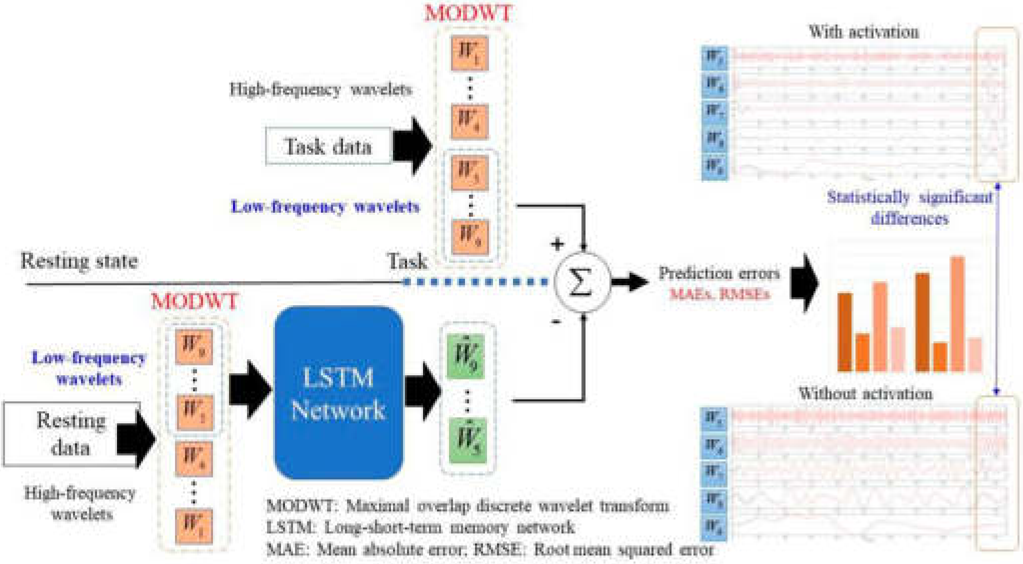


Fig. 2.1: Humming music retrieval model based on sample semantics.

music [18]. By summing the sentence length of adjacent sentences, the corresponding candidate music can be obtained, and the eigenvalue of the candidate music can be extracted. The sentence length feature is the number of music features retained in the music library in sentence units. This is the distance between two consecutive sounds in a sentence. It extracts the number of features in the sentence the user wants to hum. There are two ways to generate eigenvalues:

a) Direct generation of the characteristics of each tonal profile. When the contour features of two sentences are connected, a feature value needs to be added between the two sentences. This value is the interval difference between the latter sentence's first note and the previous sentence's last note. However, when the user connects two sentences, it is easy to make the connection between them inaccurate. Increasing the value of this feature results in lower detection efficiency.

b) Ignore the eigenvalues between different sentences. In this case, it just matches the attributes in the sentence.

3.1. Alternate feature extraction algorithm in multi-sentence search. By preprocessing, feature extraction and melody contour extraction of the song fragment that the user wants to search, the feature string is: $F = \{f_1, f_2, \cdots, f_l\}$. Its sequence is l. Any track in a music library $S = \{L_1, L_2, \cdots, L_d\}$ can be represented by $L_i(1 \leq i \leq d)$. The set of sentences is$L_i = \{R_1, R_2, \cdots, R_{lr}\}$.$R_i(1 \leq i \leq lr$ is for any line in a song. lr is the number of sentences in the entire song. Rearrange the letters that make up R_i into $R_i = \{\varphi_1, \varphi_2, \cdots, \varphi_n\}$. Where n is its length. A multi-sentence search algorithm is used to determine the candidate set of music fragments:

Definition 1. A collection of candidate music pieces. Include the following in the database

$$L_i = \{R_1, R_2, \cdots, R_{lr}\} \tag{3.1}$$

The sentence length is ordered as $D = \{d_1, d_2, \cdots, d_{lr}\}$. The length of the track F to be retrieved is ld. Its fragment G is identified as

$$\begin{aligned} &G = \{R_i + R_{i+1} + \cdots + R_j|| \\ &d_i + d_{i+1} + \cdots + d_j - ld| \\ &\leq \varepsilon_G, (1 \leq i \leq j \leq k)\} \end{aligned} \tag{3.2}$$

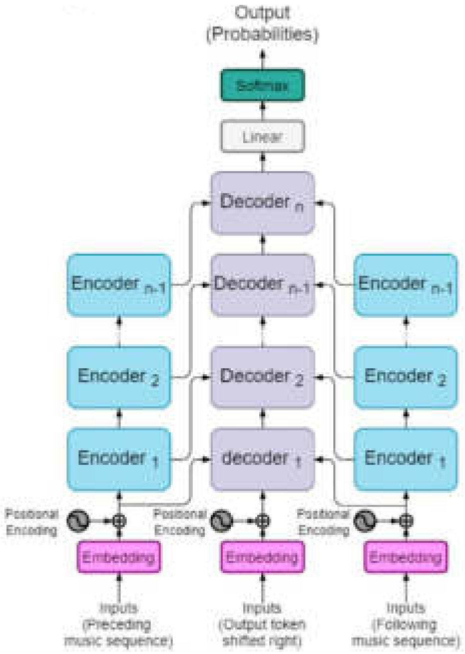


Fig. 3.1: Music candidate segment generation process diagram.

ε_G is the upper limit of the allowed length set by yourself. That's the most significant difference between the two pieces. The following is a more detailed description of the candidate feature generation algorithm for multi-sentence search.

3.1.1. Produce an alternate fragment. The algorithm in Figure 3.1 generates a set of musical candidate pieces G based on the above definition.

3.1.2. Generate alternative feature quantities. The corresponding eigenvalues are generated in the following two ways based on the candidate fragment set:

1) The generation of corresponding tonality profiles, respectively. The method of producing the corresponding melody profile characteristics from the resulting set of alternate music fragments is shown as follows:

(a) The characteristics of the melody profile are generated. When the user hums multiple tunes, the extracted feature quantity is the overall feature quantity composed of multiple tunes [19]. If it is multiple notes, adding one note to the two notes is required. This value is the interval difference between the following sentence's beginning note and the previous sentence's end note. Then the characteristics of two consecutive sentences R_1 and R_2 are:

The first segment of R_1 + the first note interval of R_2.

The interval characteristic of the difference between the end of R_1 and the interval of R_2.

Then, the profile characteristics of the candidate tunes are obtained.

(b) Characteristics of musical rhythm. The tune of the label is characterized by $\Theta(query\ rhythm\ length)$.

$$\Theta = \{\theta_1, \theta_2, \cdots, \theta_{lr}\} \tag{3.3}$$

The musical length field is the musical length of each sentence in a song, $\phi(database\ \ rhythm length)$.

$$\phi = \{\varphi_1, \varphi_2, \cdots, \varphi_n\} \tag{3.4}$$

Let's say $N = \min(msn)$. Let's define $\Delta[N] = \{\frac{\theta_1}{\varphi_1}, \frac{\theta_2}{\varphi_2}, \cdots, \frac{\theta_N}{\varphi_N}\}$. Calculate $dist_{rl} = \sum_{i=2}^{N} |\Delta[i] - \Delta[1]|$. Using the dynamic programming method, the minimum $dist_{rl}$ value can be obtained.

2) Do not consider the specific number of features between sentences. Without considering the eigenvalues between sentences, the eigenvalues are calculated as follows:

For users of multiple speech sentences, the number of features in speech is continuous. These two sentences have an abstract character, but the two coherent sentences make the user hum inaccurate. Multiple features are often extracted from a coherent Chinese character when feature extraction is carried out. When the feature of the armature is set to "?" It can be matched with any number of characters. In this way, you can get the characteristics of multiple sentences that the user is humming.

3.2. Humming multi-sentence retrieval algorithm. A speech recognition method based on DTW (Dynamic Time Warping) is proposed to solve the problem of pitch inaccuracy caused by speech errors in humming recognition. DTW is a nonlinear rule that combines timing rules with distance measurement rules, and it is widely used in speech recognition. This method is used to compare the embedded-remove errors of a certain class of characters. Thus, the optimal matching subsequence is obtained and its similarity is maximized [20]. If the corresponding features of the two substrings are inconsistent, the similarity definition proposed in this paper is used to analyse the similarity of the two substrings. Here is a description of the algorithm:

1) Extract user humming feature string.

$$R = {}'\varepsilon_1\varepsilon_2\cdots\varepsilon_n\cdots\varepsilon_N{}'(N \geq 0) \tag{3.5}$$

2) The corresponding feature sequence is obtained by classifying it based on the length N of the sequence removed by humming.

$$T = {}'\sigma_1\sigma_2\cdots\sigma_l\cdots\sigma_L{}'(L \geq 0), |L - N| < \varepsilon \tag{3.6}$$

3) Search for humming songs according to the DTW algorithm.

The research focus of DTW is to find the time function $l = w(n)$ with certain regularity. The algorithm corresponds a time history n of an input byte to a time history l of a reference byte nonlinearly. And w satisfies$ldis = \min\limits_{w(n)} \sum\limits_{n=1}^{N} d(\sigma_l, \varepsilon_{w(n)})$. Under the best time rule, $d(\sigma_l, \varepsilon_{w(n)})$ is the measure of distance between two fields. The algorithm for approximate string matching based on the stated distance looks like this: Type the string $R = {}'r_1r_2\cdots r_n\cdots r_n{}'(n \geq 0)$. Each$R$ here represents a different record in the database.

Output some data that is highly similar to A.

1) The spacing matrix S of corresponding characters in T and R is recorded in the article i of the calculation library.

2) The optimal route is obtained Through dynamic programming of the model.

3) Starting from the definition of similarity, the minimum and the difference between the two lines and their lengths can be obtained to obtain the similarity of T and R. Store a similar value. Add 1 to the value of i and return to a), and so on—to the last record.

4) When all the similarity values are taken as the result of the closest data among the multiple similarity values. In this way, the user can sing any melody, find the target song, and thus achieve multiple sentence searches.

4. Experimental analysis. Two actual databases verify the proposed method. The MIR-QBSH dataset contains 2,048 songs in MIDI format. A total of 4,431 segments were questioned. The singer sings from the beginning of the song. The IOACAS data contains 298 formatted songs and 759 entries from MIDI. The singer starts singing anywhere in a song. The test data was sampled at 8 kHz and stored in 8-bit resolution Wav format. It is converted to an FM sequence by an FM tracking device. The research work in this paper is carried out on Intel's Q8400 CPU (2.66 GHz) and 2 GB microcomputer. The 64-bit Ubuntu12 system was used. Write programs in C++.

4.1. Verification and analysis of search results based on segmentation of songs. Firstly, the sentence features are extracted from the tonality sequence. Divide the music into different sections according to the sentence. Take the tonality position where tonality 0 occurs consecutively in the tonality sequence as a sentence. The tunes the user hums are usually higher or lower than the original tunes. The pitch the user

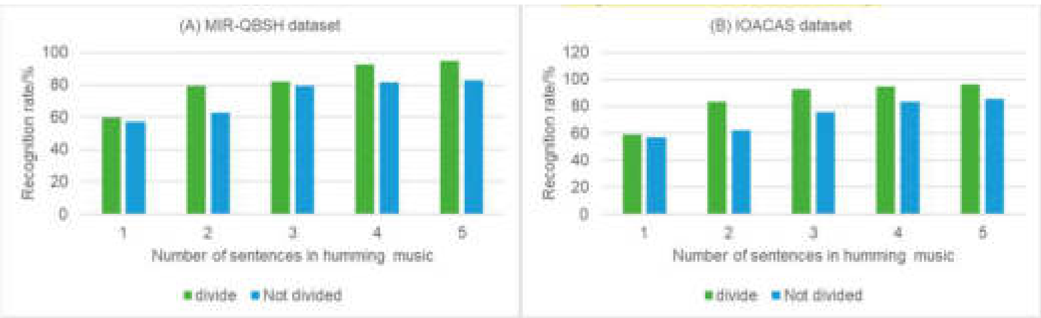


Fig. 4.1: Test effect of music division statement retrieval recognition rate.

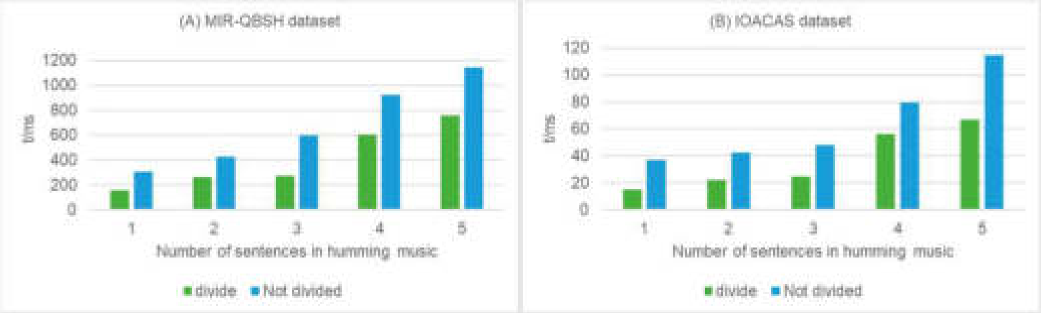


Fig. 4.2: Test effect of music division statement retrieval efficiency.

is humming must be normalized along with the songs in the database. Each point is then modified to be the ratio of its original value to the average height.

We will take MIR-QBSH and IOACAS as research objects to compare BDTW regarding retrieval efficiency and retrieval speed of segmented music sentences. Figuress 4.1 and 4.2 show the identification rate and efficiency of the method. The value of the fault tolerance limit factor should be determined according to the user's feelings about the song. In the experiment, the value of a is 10%. The paper divided the data group into five types based on the number of sounds made by the people in the data group. From Figures 4.1 and 4.2, we can see that the search based on music sentences performs better in all 5 cases regarding efficiency. Especially in the case of a large number of statements, the superiority of this method is more pronounced. This is primarily due to the use of sentences to distinguish, more in line with people's habits of lyrics. The segmentation of the music sentence prevents the mismatching of the front of the unsegmented music sentence from affecting the humming tune. This results in a significant increase in efficiency.

4.2. A method for searching songs by using sentence characteristics. This part of the experiment compares BDTW with similar DTW on a data set. The error constraint factors a is 10%. This paper compares the performance test of the DTW algorithm and DTW algorithm in execution efficiency and execution speed when the return result is top-1, top-5, top-10, top-15 and top-20 under the DIS index structure. The results are shown in Figures 4.3 and 4.4. BDTW search results are better. In DIS, the larger the k value is, the better the retrieval result is. BDTW has better performance than DTW. For top-5 problems, BDTW has obtained a high identification rate. BDTW showed high results in the search of songs, while the total search time spent in the search process remained the same.

5. Conclusion. BDTW algorithm allows users to give fault tolerance limit factors according to their humming level. This algorithm can solve the maximum difference between the restricted database and the query order. This completes a complete sequentially matched musical statement. The exponential structure of

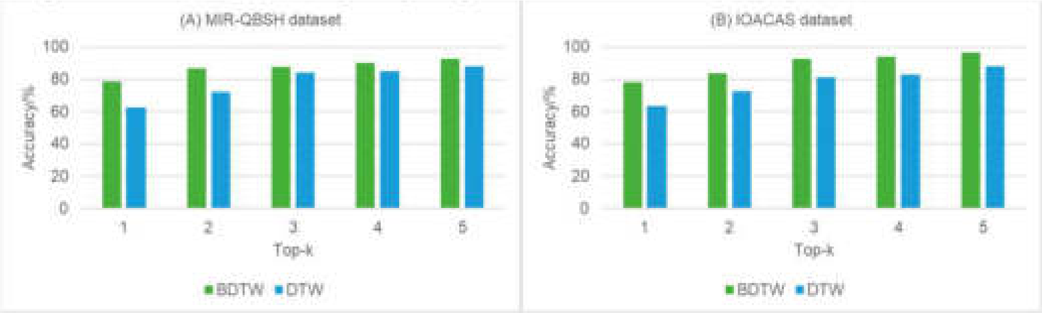


Fig. 4.3: Retrieval accuracy test effect of BDTW algorithm.

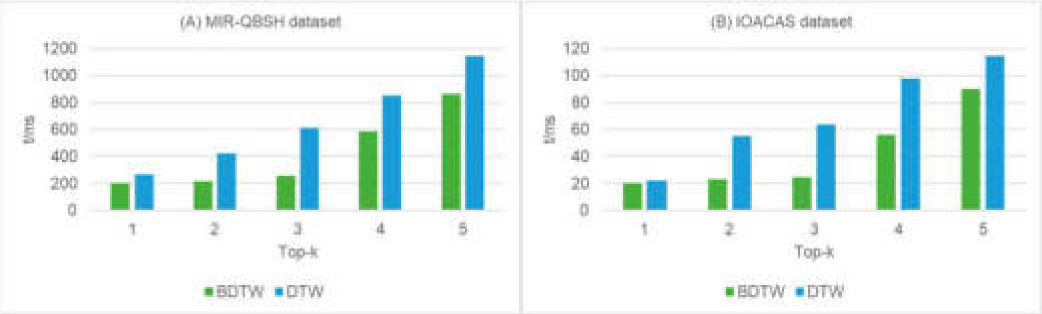


Fig. 4.4: BDTW retrieval efficiency test effect.

DIS is also given. This method is to construct an index structure in the database to reduce the query speed and achieve fast data retrieval. Experimental results show that the algorithm can accurately retrieve the humming problem quickly and efficiently.

REFERENCES

[1] Lee, K. Y., & Hu, C. M. Research on the development of music information retrieval and fuzzy search. Scientific and Social Research, 2022; 4(4):1-10.

[2] Kappen, P. R., Beshay, T., Vincent, A. J., Satoer, D., Dirven, C. M., Jeekel, J., & Klimek, M. The feasibility and added value of mapping music during awake craniotomy: A systematic review. European Journal of Neuroscience, 2022;55(2):388-404.

[3] Muthoifin, M., Ali, A. B. E., Al-Mutawakkil, T., Fadli, N., & Adzim, A. A. Sharia Views on Music and Songs: Perspective Study of Muhammadiyah and Madzhab Four. Demak Universal Journal of Islam and Sharia, 2023;1(01): 10-17.

[4] Carvalho, M. E. S., de Miranda Justo, J. M. R., Sá, C., Gratier, M., & Rodrigues, H. F. Melodic contours of maternal humming to preterm infants in kangaroo care and infants' overlapping vocalizations: A microanalytical study. Psychology of Music, 2022;50(6):1910-1924.

[5] Carvalho, M. E., Justo, J. M., Gratier, M., & Ferreira Rodrigues, H. Infants' overlapping vocalizations during maternal humming: Contributions to the synchronization of preterm dyads. Psychology of Music, 2021;49(6):1654-1670.

[6] Rezaei Oshaghi, N., Baradar, R., & Ghaebi, A. Methods of searching music information sources in search engines. Librarianship and Information Organization Studies, 2023; 33(4): 38-58.

[7] Bakouros, S., Rarey, K., & Evered, J. Retinopathy of Prematurity Screening Exams, Adverse Events, and Music Therapy: A Case Series. Music Therapy Perspectives, 2023; 41(1): 47-53.

[8] Rushton, R., Kossyvaki, L., & Terlektsi, E. Music-based interventions for people with profound and multiple learning disabilities: A systematic review of the literature. Journal of Intellectual Disabilities, 2023; 27(2): 370-387.

[9] Pridy, C. B., Watt, M. C., Romero-Sanchiz, P., Lively, C. J., & Stewart, S. H. Reasons for listening to music vary by listeners' anxiety sensitivity levels. Journal of music therapy, 2021;58(4): 463-492.

[10] Baptista, A., & da Silva, C. G. Organization and Representation of Musical Information (ORMI) in Portugal: a literature review. Boletim do Arquivo da Universidade de Coimbra, 2021; 34(2): 11-26.

[11] Zalkow, F., Brandner, J., & Müller, M. Efficient retrieval of music recordings using graph-based index structures. Signals,

2021; 2(2): 336-352.

[12] Tamboli, A. I., & Kokate, R. D. Query based relevant music genre retrieval using adaptive artificial neural network for multimedia applications. Multimedia Tools and Applications, 2022; 81(22): 31603-31629.

[13] Lee, B. H., & Kim, M. 2021; Algorithm to Search for the Original Song from a Cover Song Using Inflection Points of the Melody Line. KIPS Transactions on Software and Data Engineering, 10(5): 195-200.

[14] Bosher, H. Sheeran succeeds in 'Shape of You'music copyright infringement claim. Journal Of Intellectual Property Law and Practice, 2022;17(7): 544-546.

[15] Fisher, M., & Rafferty, P. Current Issues with Cataloging Printed Music: Challenges Facing Staff and Systems. Cataloging & Classification Quarterly, 2023; 61(1): 91-117.

[16] Velankar, M., & Kulkarni, P. Melodic Pattern Recognition and Similarity Modelling: A Systematic Survey in Music Computing. Journal of Trends in Computer Science and Smart Technology, 2022; 4(4): 272-290.

[17] Kreimer, S. This Neuromuscular Specialist Keeps Life Humming with Guitar Playing, Songwriting, and Board Game Development. Neurology Today, 2021;21(21): 17-18.

[18] Kennedy-Macfoy, M. Everything Must Change. European Journal of Women's Studies, 2023; 30(1): 3-6.

[19] Setyaningsih, E., Chandra, I., & William, W. Aplikasi Music Streaming Menggunakan Flutter dilengkapi Music Recognizer. Jurnal Inovasi Teknologi dan Edukasi Teknik, 2021; 1(9): 707-714.

[20] Lee, K. Y., & Hu, C. M. Research on the development of music information retrieval and fuzzy search. Scientific and Social Research, 2022; 4(4): 1-10.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Oct 18, 2023
Accepted: Nov 6, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1807–1817, DOI 10.12694/scpe.v25i3.2714

INSTINCT, SUPRRESION AND CATHARSIS: THE PSYCHOLOGICAL SOURCE AND GUIDANCE OF TEENAGERS' NETWORK IDEOLOGY IN THE ERA OF BIG DATA

JIAO CHEN[*] AND FENG DU[†]

Abstract. The unique educational attributes of the big data era can not only improve the audience's acceptance effect, but also accelerate the overall process of achieving ideological security education goals, playing a certain role in maintaining the unity of youth's thinking and daily behavior. Youth are the ”core indigenous people” in the field of the Internet. Big data and mobile communication have brought new changes to the lifestyle and values of young people. In an era full of data, linking big data with life, work, and education is a contemporary issue that we should pay attention to and think about. The purpose of this study is to explore the psychological sources and guidance methods of adolescent online ideology in the era of big data. Firstly, through literature analysis and empirical research, we found that psychological factors such as instincts, repression, and venting among adolescents have a significant impact on their online ideology. Secondly, we utilized methods such as questionnaire surveys and in-depth interviews to explore the relationship between different psychological sources and online ideology. Finally, we propose a series of targeted guidance strategies, including enhancing self-awareness, cultivating a healthy mindset, and improving network literacy, to help teenagers establish the correct network ideology. It is necessary to strengthen the self-management awareness of teenagers, cultivate targeted opinion leaders, strengthen online supervision, establish and improve a risk prevention mechanism for online ideology, and promote mainstream ideology.

Key words: Big data era; Teenagers; Network ideology; Risk guidance

1. Introduction. There is a connection between data and the emergence of human society. The role and influence of data on society is very great and cannot be estimated. It is the forefront of the ideological work of teenagers in the Communist Party of China, and the ideological safety work of teenagers is an important part of the ideological work of the Communist Party of China [1]. Youth ideological education is a strategic and basic project of current ideological and political education for teenagers, which is of great significance. The power of data is infinite, and the speed of big data integrating information resources is very fast, the efficiency is extremely high, and it involves a wide range, so you can get the desired results soon. Internet, a global survival concept, is impacting multiculturalism and pluralistic world. It is changing people's daily life, interpersonal communication and even redefining people's social value with amazing power. It puts forward new deployment and theoretical guidance for the core Party Central Committee to strengthen and improve the party's ideological work in the era of big data from the strategic height and theoretical orientation, and forms a socialist ideology with Chinese characteristics in the era of big data [2]. Internet ideological security is an important part of ”overall national security”. However, as an ideological theory with rich academic rationality, ideological security is difficult for Chinese teenagers to make in-depth and accurate cognition. The existence of youth network ideology is due to the high degree of fusion between youth network virtual individuals and real social individuals. It is an intermediary system of digital, symbolic and informative network platform that young people use for their own needs. A system of beliefs and values with symbolic meaning formed in the symbiotic sharing activities of information, knowledge, and spirit in the network society [3]. Teenagers play an important role in studying, researching and publicizing Marxism, cultivating and promoting socialist core values, and cultivating successors and builders. Strive to make the ideological education of young people play a multiplier effect with the help of big data through research. Big data is spreading to every corner of the world in a frenzy, its development speed is amazing, and it permeates our life, work and thinking. Today, when human society is marching towards the information age, we must attach great importance to network moral education,

[*] School of Marxism, Chongqing Youth Vocational & Technical College, Chongqing, 400712, China

[†] School of General Education, Chongqing Youth Vocational & Technical College, Chongqing, 400712, China (Corresponding author, 13637920376@163.com, dufeng@cqyu.edu.cn)

and vigorously strengthen it through various channels and forms, so that teenagers can consciously restrain their behavior. Strive to minimize the impact of ideas and behavioral anomie that the Internet may bring to the younger generation. With the development of global informatization and Internet technology, the Internet has become a communication ideology in some countries. Then the main way to attack the ideology of other countries is even the main battlefield of ideological struggle [4]. The mixed quality, speed, breadth and depth of the network information make "we must know the law of network communication scientifically. Improve the level of network governance with the Internet, and make the biggest variable of the Internet become the biggest increment of career development" [5]. As the "participants" and "attendees" of the network, the majority of the youth are the key to resolve the network ideological risk.

In a certain sense, we should prevent and handle the network ideological risk of the youth. Ideological security is a necessary condition for China's overall national security and an important guarantee for the great rejuvenation of the Chinese nation. "Whether the ideological work can be done well is related to the future and destiny of the party, the long-term stability of the country, and the national cohesion and centripetal force." Starting from the instinctive feelings of teenagers using the Internet, it provides a new perspective for the study of teenagers' network ideology in the era of big data [6]. At present, the Internet and big data technology are infiltrating all aspects of social life, and big data is entangled with mixed and indistinguishable information sources, which has eliminated the mainstream ideological discourse power. This makes the situation of struggle in the domestic ideological field more and more complicated, especially the ideological security construction of young people is facing an extremely severe test. Young people are the "core aborigines" of the Internet space, and big data and mobile communications have brought new changes to the lifestyles and values of young people. Of course, this is also an issue that every citizen should think about and pay attention to, and it is also a major historical task faced by relevant researchers.The virtuality and anonymity of online socializing may lead to trust crises, online bullying, privacy breaches, and other issues for teenagers during the social process, affecting their mental health and social adaptability. Some teenagers may become overly addicted to online games and virtual worlds, leading to problems such as impaired academic performance, decreased social skills, and impaired physical and mental health. Implementing the online real name system can increase the transparency of online social interaction, reduce the problems caused by anonymity, and enhance the self-control awareness of teenagers. Parents and schools should guide teenagers to establish a healthy online lifestyle, arrange their online time reasonably, and avoid excessive addiction to online games and virtual worlds. Schools and families should strengthen cybersecurity education, educating teenagers on how to use the internet correctly, protect personal privacy, and prevent online fraud.

The Communist Party of China attaches great importance to the construction of ideology, which is related to economic development, social harmony and political stability. Adhere to the bottom-line thinking and focus on preventing and resolving major risks in the seminar, taking ideological risk as the second biggest risk, with special emphasis on the important position of network and youth [7]. The unique educational attributes in the era of big data can not only improve the acceptance effect of the audience, but also accelerate the overall process of realizing the goal of ideological safety education, which plays a role in maintaining the unity of teenagers' ideology and daily behavior. In the era of big data, the Internet provides a special expression environment and catharsis space for teenagers, which also increases the difficulty of guiding and controlling teenagers' network ideological risks to a certain extent. Scientifically judge the macro development trend of youth ideological security construction under the background of big data, and clearly understand the realistic dilemma and path choice faced by youth ideological security construction. Nowadays, the development speed of the Internet is beyond imagination, and the utilization rate of big data is getting higher and higher. In the era of data flooding, linking big data with life, work and education is an issue that we should pay attention to and consider at present. Ideological education for young people should embed big data technology into the whole process of ideological education for young people, and inject the vitality of the times and innovation into ideological education for young people [8]. On the basis of in-depth research on the concepts, characteristics and relationship between big data and ideology, we deeply analyze and master the many challenges in security currently facing, and propose corresponding solutions. This novelty, born during the Cold War, has become the busiest, most challenging, and most dynamic system in today's information society. It has become the fourth most dynamic and open media after newspapers, radio and television. The application of intelligent algorithms and big data analysis in film and television creation has become an undeniable trend. The introduction of these technologies not only changes the way film and television production is done, but also provides more possibilities for creators, greatly improving creative efficiency. In the script creation stage, intelligent algorithms can extract possible plot clues and character relationships through the analysis of a large amount of text data, providing inspiration for screenwriters. Meanwhile, through emotional analysis technology, intelligent algorithms can also perform emotional orientation analysis on specific texts, helping screenwriters better grasp the emotional direction of the story. The Internet is a big platform for social information. Hundreds of millions of netizens obtain and exchange information on it, which will have an important impact on their ways of seeking knowledge, ways of thinking, and values [9]. In particular, it will have an important impact on their views on the country, on society, on work, and on life. While strictly anticipating risks in the field of network ideology and striving to prevent them from happening, it is also necessary to be prepared, strengthen response, and effectively improve risk management capabilities. Specifically, one is to achieve division and rule. Faced with the frequent and complex risks in the field of online ideology, it is necessary to be able to clear the clouds and simplify the complexity. Risks in the field of online ideology should be classified and graded based on factors such as content, nature, characteristics, and impact. The core issues that cause risks should be identified, and targeted and targeted measures should be taken in sequence and separately. We should pay attention to distinguishing the root causes of different types and degrees of cyber ideological risks, such as provocation and attacks by hostile forces both domestically and internationally, the influence of erroneous social ideologies, livelihood issues, and the widespread politicization of social events, and solve problems with a targeted approach. And in the process of risk management, it is necessary to effectively avoid the complexity of simple issues and the simplification of principle issues.

2. Related Work. Foreign scholars have different opinions on the existence of ideological problems in the era of big data. Daniel Estey established a decision-making form driven by data, that is, using new scientific and technological means to change the decision-making process of things in the past [10]. Western anti China forces use their control over the internet and information dissemination rights to spread Western values to China, further threatening China's ideology, culture, and national security. Ideology plays an extremely important role in maintaining the stable and orderly operation of a country's society and facing international power struggles, and can serve as a political platform for unifying people's thoughts and actions. Only by returning to the "practical" foundation of ideology can we have a scientific understanding and summary of ideology, and truly understand the theoretical consciousness and practice of Marx's important discourse on ideology. Marx's ideology is similar to opium, and over time, his ability to judge will gradually lose, showing his helplessness [14]. The characteristics of big data include "high capacity", "processing capabilities beyond common software and hardware environments", and "changing human society" [15]. The development of the Internet has led to the emergence of a large number of virtual communities, but there is an objective trend of "de ideology" in virtual communities [16]. The intelligent distribution technology of information formed by the application of algorithms in the field of information dissemination poses enormous risks and impacts on network ideology [17]. The West not only utilizes the technological advantages of the all media era and the influence of online communication to consolidate the mainstream position of Western countries in ideology, but also utilizes the weakening and lack of information rights in other countries in the all media era to raise awareness. The dominance and output of forms have brought new challenges to China's ideological security [18]. Foreign scholars have conducted extensive and in-depth research on ideological security issues caused by the era of big data, and obtained corresponding theoretical results.

3. The Realistic Relationship Between the Era of Big Data and Ideological Safety Education for Teenagers.

3.1. Opportunities Provided by the Era of Big Data for the Optimization of Ideological Safety Education for Chinese Youth. Teenagers should be good at utilizing online resources for learning, communication, and innovation, rather than using the internet for entertainment and leisure. They should actively acquire scientific and cultural knowledge, enhance their internet literacy, and avoid falling into negative information and cultural traps on the internet. Teenagers should develop good internet habits, including regularly and quantitatively surfing the internet, not indulging in online games, and not randomly clicking on unknown

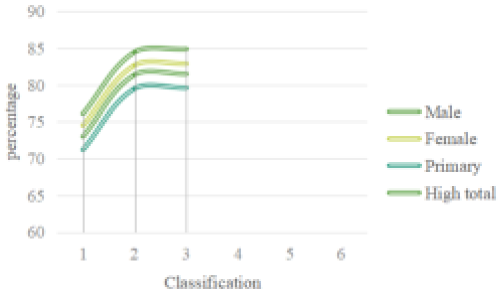


Fig. 3.1: Comparison of the will quality of adolescents of different genders and grades

links. They should pay attention to their online time and avoid excessive use of the internet causing damage to their physical and mental health. The government and society should strengthen the supervision and management of the online environment, combat cybercrime and the dissemination of harmful information, and provide a healthy online environment for teenagers. The government and society should advocate the Convention on Cyber Civilization, guide teenagers to establish correct network values and moral concepts, and cultivate their good habits of civilized and healthy internet use. With the change of people's social life style and the further development of the era of big data, mobile phones, Internet, digital TV and so on have become indispensable tools in modern people's daily life. In reality, people are always bound by various social rules and secular morals. Under the influence of social and cultural reasons, the original impulse and instinctive desire of individuals are suppressed to the threshold of consciousness, and become the subconscious that is not realized by the subject [19]. Therefore, the so-called ”social man” often presents a relatively ”hidden” state of existence. Big data, as a new factor of production in the era of Internet economy, has greatly promoted the profound change of production relations, and will completely change the production mode, lifestyle, working mode and thinking mode of human beings. Imperceptibly promote the development of social ideology, but at the same time it inevitably becomes the coveted object of the spread and penetration of western capitalist ideology [20]. Today is the ”big data era”. Under the background of this era, the impact and change of the diversified dissemination methods and channels of data information on the public opinion environment is unprecedented and historic. Some professional teachers believe that the main task of teenagers is to teach knowledge, and ideological safety education is optional. Comparison of the will quality of adolescents of different genders and grades. As shown in Figure 3.1. The data sources for the willpower characteristics of adolescents of different genders and grades in Figure 3.1 may come from multiple channels. These data may be based on various surveys and studies, including observation and evaluation of adolescents, as well as a review and analysis of relevant literature. In addition, these data may also come from statistical data and research reports provided by some public institutions, non-governmental organizations, or private companies.

As long as it does not violate the line, principles and policies of the party and the state, teaching is the key. Therefore, the focus of all work is placed on teaching and student management. The so-called Youth Ideological Education in the era of big data refers to the whole process of embedding the ideas, technologies and methods of big data into the daily life education and practice of youth ideology. Big data brings convenience to teenagers' work, rapid information transmission, and the specific objectives of psychological education are more detailed and targeted. As shown in Tables 3.1, 3.2 and 3.3. The big data for adolescents in Tables 3.1 and 3.2 uses data mining techniques to process and analyze a large amount of data. These data mainly come from publicly available data from the internet and related institutions, as well as the data we collected in questionnaire surveys and in-depth interviews. We used data mining techniques to clean, integrate, and classify these data, ultimately resulting in big data related to youth work, information transmission, and psychological education.

Promote the comprehensive optimization of the collection, implementation, inspection, evaluation, mediation, and research of youth ideological education information, and realize the innovation of youth ideological education paradigm and the optimization of results. They regard neutral world outlook and pluralistic thinking

Table 3.1: What is your favorite way to strengthen youth psychological education?

Way	Special Report	Traditional teaching	Case analysis	Theme activities	Social practice	Network interaction
Frequency	97	89	156	184	305	293
Effective percentage	16.76%	14.56%	26.34%	32.37%	51.74%	50.50%

Table 3.2: Do you like psychological education in courses or online psychological education best?

Type of psychoeducation	Frequency	Effective percentage
Curriculum Psychological Education	170	29.46%
Online psychoeducation	301	52.24%
Neither like	69	11.63%
Like it all	40	6.68%

as their own, denying the guiding role of the scientific nature and subjectivity of Marxist thought. They insist that mainstream Marxism is democratic socialism, and their views are extreme. First of all, the network culture, as a social norm, forms and regulates various network relationships among network individuals in human life, which is a programmed and institutionalized culture. Grading comparison of students' physical quality. As shown in Figure 3.2. The test group in Figure 3.2 represents the current physical fitness scores of young students. The control group represents the level of physical fitness score standards for young students.

Ideology shows various phenomena of human social thought and civilization, while the concept of ideology and the direct discussion of ideology have only a very short historical record. In a diverse and diverse social trend of thought, the mainstream ideology does not respond timely and adequately to major hot issues, and the discourse content and discourse methods are not vivid and diverse in the dissemination of mainstream values. To some extent, it has delayed the "entry" of mainstream ideology in the field of public opinion. Feed back to the publisher of information in the form of interaction, deepen the young people's individual cognition and internalization of the mainstream social thought from the original level, and form the ability to deal with the diversified pattern of domestic social thought and resist the invasion of Western bad social thoughts. The calculation formula of support is:

$$\sup p\left(X\right)=\frac{occur\left(X\right)}{count\left(D\right)}=P\left(X\right) \tag{3.1}$$

The calculation formula of confidence is:

$$conf\left(X->Y\right)=\frac{\sup p\left(X\cup Y\right)}{\sup p\left(X\right)}=P\left(Y\left|X\right.\right) \tag{3.2}$$

The calculation formula of the ratio provided is:

$$lift\left(X->Y\right)=lift\left(X->Y\right)=\frac{conf\left(X->Y\right)}{\sup p\left(Y\right)} \tag{3.3}$$

$$D=\left|D_1D_2\cdots D_n\right| \tag{3.4}$$

$$D_{ij}=D_i\wedge D_j=\left|d_{1i}\wedge d_{1j}\cdots d_{ni}\wedge d_{nj}\right| \tag{3.5}$$

$$I_1=L_1,L_2,\cdots,L_{k2},L_{k1} \tag{3.6}$$

$$I_2=L_1,L_2,\cdots,L_{k-1},L_k \tag{3.7}$$

$$I=I_1\infty I_2=L_1,L_2,\cdots L_{k-1},L_k\in L_k \tag{3.8}$$

Table 3.3: what are the difficulties in carrying out adolescent psychological education.

Difficulty	Frequency	Effective percentage
Students do not cooperate	33	35.2%
Stressed and no time	60	64.87%
Insufficient relevant professional knowledge	29	34.09%
The form of work is monotonous and difficult to attract students	41	43.90%
Insufficient external support	18	21.93%

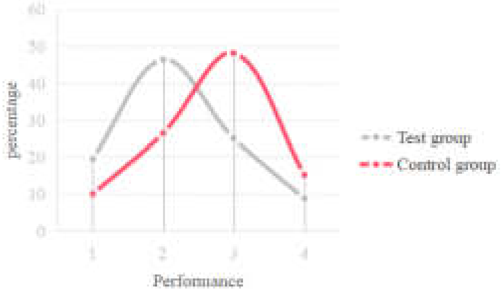


Fig. 3.2: Comparison of young students' physical quality grades

3.2. New challenges to teenagers' Ideological Security Education in the era of big data. Information dissemination in the era of big data is diversified and interactive, and everyone in the virtual platform is not only the receiver of language but also the creator of language. In short, in order to avoid the anxiety and uneasiness caused by violating the rules, teenagers can't blindly indulge the demand of the principle of happiness and do whatever they want. Although the era of big data is a brand-new historical stage in which social productive forces have developed to a certain level, it has not fundamentally changed the social nature of the present era. At present, the whole world is still in the historical trend of transition from capitalist society to socialist society. With the popularization of network media, people's thoughts and voices spread more freely, and data information is also diversified. In class, the teacher talks more, but the students ask less. Some teachers avoid answering students' questions, or don't answer from the front, or give false guidance, so that they can't get students' satisfaction and recognition. In the era of big data, youth ideological education is in the "sea of information", and all data can be collected and analyzed to realize the "transformation from small sampling to big data". The antagonistic situation between capitalism and socialist countries has become more and more serious and cannot be eliminated. This manifestation in the ideological field is essentially the confrontation and exclusion between liberalism and Marxism. Education, all its activities cannot be separated from symbols, it must be the activity of using symbols, and a large part of its function is also manifested in making educated people learn to master and use symbols, including language and other various such as mathematics, science, etc. etc. symbols. In the historical environment of the development of the times, people have a certain tendency to a certain point of view generated in a certain historical environment, and are summed up as truth and value. Ideology, on the other hand, is the product of non-dispositional consciousness. The phenomenon of aphasia in mainstream ideology is more prominent. With the development of mass media, the influence of mass ideology is gradually increasing. However, the content of some popular ideologies is not consistent with that of mainstream ideologies, and there may even be conflicts, which further leads to the phenomenon of "aphasia" in mainstream ideologies. The openness and anonymity of the online environment make information dissemination more free, but at the same time, it is also more prone to information flooding and misleading. This challenges the dissemination of mainstream ideology in the online environment, resulting in the phenomenon of "aphasia".Discourses with main themes and positive energy are regarded as "hypocrisy" and "decent", while online information that is taken out of context, fabricated indiscriminately, and clipped and grafted is easy to attract attention. The authority and dominance of the guiding position of Marxism brings value invasion and

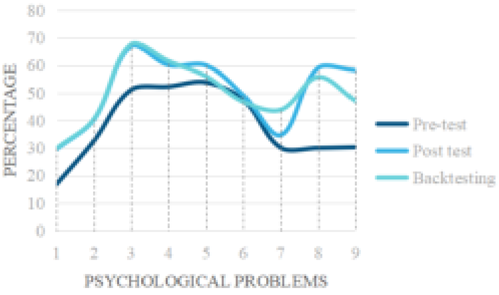


Fig. 4.1: Comparison of changes in students without problems in the backtesting experimental group before and after the intervention

cultural penetration, which leads to the fact that in the process of national ideological security popularization education, education disseminators at all levels of ideological security education are no longer the authoritative dominance.

4. Result Analysis.

4.1. Psychological Education Methods for Teenagers in the Environment of Big Data Era. As a new thing, the era of big data has been widely used by people, but the consciousness of consciously using the era of big data for ideological security education has not really been deeply rooted in the hearts of the people. The main purpose of education is to "help an individual to be himself freely, instead of forcing uniformity". Among them, ego is the central system in Freud's personality structure, the controller and mediator of people's behavior and thoughts, and the social personality of "I". College students are the main force of netizens and the most active group in the Internet space. The characteristics of ideological discourse in cyberspace, such as the penetration of time and space, complexity, diversity and free interaction, easily lead to the chaotic transmission of virtual society to the real world. The widespread and real-time nature of online communication enables the rapid dissemination of information, making it easier for social risks to be generalized. The openness of the cyberspace allows anyone to express their views and opinions online, which has led to the emergence of multiple voices and diversified ideological discourse. The symbiotic effect of Internet communication and social risk generalization leads to the pluralistic division of ideological discourse. People will look at everything from a holistic perspective, acknowledging the diversity and difference of the world, recognizing that all kinds of data are of equal importance, all data will be open to people, and everyone has an equal right to data. In teaching, full education has not been fully formed. The work of ideological education should be undertaken by teachers of ideological and political courses, and ideological education should not be infiltrated into the daily management of students and the education of professional courses. In addition, some teachers of ideological and political theory courses do not understand, understand and understand the content of the textbooks, but preach in a simple way, and the practical teaching and theoretical teaching are out of touch. The students in the experimental group who received the intervention education were not significantly different from the control group in the two dimensions of anxiety and learning pressure after the intervention. As shown in Figure 4.1 and 4.2.

In the Internet era, western countries have changed the previous mode of communication of direct ideological confrontation and packaged it with simple and lively life-oriented language elements. Using Internet technology to describe and model western values, and imperceptibly infiltrate ideology, is more hidden and harmful. The new left's thinking methods, ideas and views come from the west, but once they find the problem, they will point to China. The real world is unequal, and the network can do this. The Internet is a natural product of equality. For example, both men and women may achieve equality in the virtual world. It can be said that in just a few years, it has achieved greater achievements than the women's Liberation Movement in a hundred years. There will be no difference between men and women in enjoying and using the Internet in the future. Therefore, ideology is the external manifestation of the will of the ruling class, so ideology is equated with

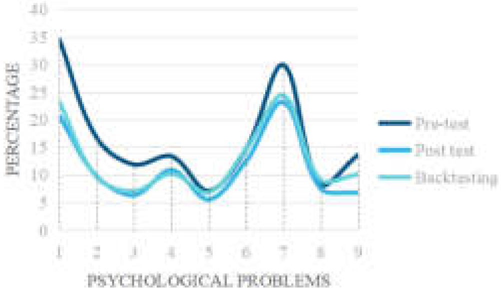


Fig. 4.2: Comparison of changes in students with moderate or above problems in the back-test experimental group before and after the intervention

class interests, goals and ideals. At this time, ideology becomes a false ideology. However, it also brings a lot of problems to the network supervision, including online fraud, online violence, online pornography, online loans, online games and other issues, which are characterized by strong concealment, deep penetration and wide coverage. Experts, such as forums, which are flexible and meet the development needs of teenagers, should strive to cover the contents, objectives, methods and methods of ideological safety education for teenagers in an all-round way, so as to realize the effectiveness of virtual education resources education.

4.2. Build a New Media Practice Education Platform. The material form of ideology determines the importance of life-oriented practical education. For example, ideological safety education can be integrated into people's "micro-life". Then it is possible for educators to influence the ideological consciousness of teenagers in a subtle form and realize the behavior of externalization into individual reality, and finally achieve the expected goal of ideological safety education. Therefore, in addition to the subconscious "showing" and the unconscious tendency to express desire. Adolescents' online activities will still be "masked" by self-consciousness and indirectly suppressed by superego spirit to a certain extent, instead of being in a state of absolute freedom and no will. With the subversive changes in the audiences, media and scenarios faced by ideological work in the era of big data, the authority of mainstream ideological discourse represented by teachers has been weakened. Some communication functions of educators have been replaced by social media, which strengthens students' participation in the process of self-cognition, and weakens the effect of traditional ideological education. On the other hand, China is now in a period of social transformation, coupled with the increasing strength of China's reform and opening up, the value system disorder caused by the diversification of values has spawned many undesirable social phenomena. All kinds of non-mainstream social thoughts and bad information affecting ideological security spread frequently in the society, weakening the influence and dominant position of mainstream ideology. Multiple comparisons show that the changes of students' physical fitness in the experimental group are different in different test periods. As shown in Figures 4.3, 4.4 and 4.5.Figure 4.3 shows the comparison of psychological quality levels among adolescents before intervention. The experimental group consisted of adolescents' psychological quality levels before intervention. The control group is the standard adolescent psychological quality level. Figure 4.4 shows the comparison of psychological quality levels among adolescents before intervention. The experimental group is the psychological quality level of adolescents after intervention. The control group is the standard mainstream ideological and psychological quality level.

Therefore, we must firmly grasp the discourse and leadership of Marxist ideology, and constantly innovate the incentive mechanism and operation mechanism of Ideological and political education curriculum. In the era of big data, all online activities of young people can be transformed into data, but these huge data information are generally mastered by shopping websites, news and social networking websites, government agencies and so on. Due to the needs of business and technical confidentiality, it is often difficult for ideological educators to obtain these data related to the thinking habits and behavior characteristics of young people. Due to the fundamental opposition of social systems and the different national conditions of each country, it can not be generalized. With the acceleration of network communication, the conflict between universal values and

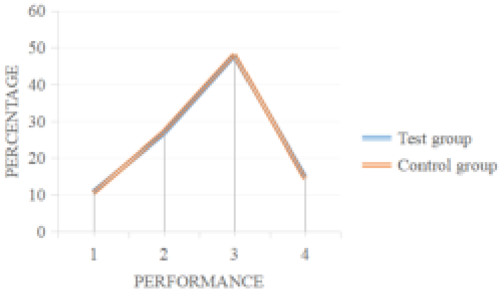


Fig. 4.3: Comparison of psychological quality grades of adolescents before intervention

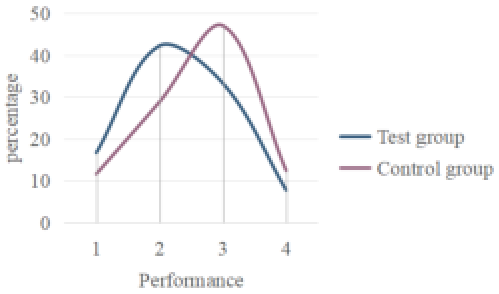


Fig. 4.4: Comparison of psychological quality grades of adolescents after intervention

China's socialist core values is also intensifying. Anyone on the Internet can say or do almost anything they want according to their own wishes, and contact anyone on the Internet all over the world. Free access to various information resources, electronic bulletin boards, newsgroups and electronic forums on different topics in cyberspace, open to anyone who is interested. People are the producers of ideology. The ideological content is rich and numerous, spanning different fields, and the ideological content is also very specific, which is produced by people through practical activities. On the other hand, it is to construct the value position, ideological attitude and social mood based on the supremacy of capital interests, and then obtain the right of spiritual production and control in cyberspace. It influences the trend of Chinese public opinion field and social value orientation, and provides convenience for the spread and practice of capitalist ideology in China. Cooperate with the practical measures of ideological safety educators to carry out more effective practical education in daily life, integrate it into the "micro-life" of teenagers, and strive to touch the field of teenagers' consciousness and emotions to the greatest extent, so as to enhance the effectiveness of education.

5. Conclusions. In addition to strengthening the guidance of adolescents' online ideological risks in the era of big data, we should focus on the research on adolescents' individual online behaviors and their psychology. Big data technology is a double-edged sword. We should correctly understand the advantages and disadvantages of big data, dialectically and comprehensively look at the current opportunities and challenges, and the impact of big data on ideological education. On the basis of adhering to the correct guiding ideology, reflecting the times and novelty, a multi-level and multi-dimensional independent and interconnected ideological safety work system for young college students is formed. Due to the different systems of socialism and capitalism, Western hostile forces will not give up their infiltration and disintegration of mainstream consciousness in China. The Internet has increasingly become an important position for current ideological and cultural dissemination, and online media and online education have had a significant impact on people, especially young people's ideological understanding. As a reflection of the superstructure of the concept of real society, ideology acts on the economic basis of real society and reflects the ideological system of the class political party it represents. It exists in many

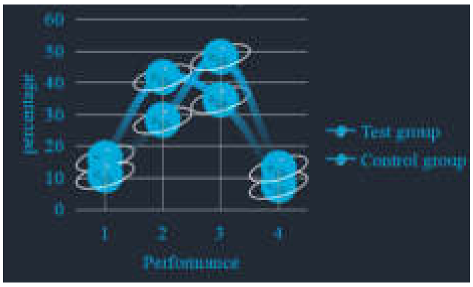


Fig. 4.5: Comparison of adolescents' psychological quality during back test

forms, and ideology has the characteristics of class, social practice and historical inheritance. With the increase of learning content and difficulty, students' motivation level will weaken, and teachers should appropriately reduce the difficulty of practice. The setting of training questions should be simple, enhance their learning motivation, and let all students experience the happiness of learning. We need to study the advantages and values of human nature while taking into account the impact of the environment on individual emotion and personality.

REFERENCES

[1] Sturmey, P., *Psychological and Sociological Theories of Violence and Aggression. Violence and Aggression: Integrating Theory, Research, and Practice, 215-232, 2022.*

[2] Jia, G., *A New Exploration of Strengthening Ideological Safety Education for Teenagers in the Age of Big Data—Comment on "Research on Ideological Safety Education of Teenagers in the Age of Big Data. Journal of Chongqing University of Posts and Telecommunications: Social Science Edition, 33, 1, 1, 2021.*

[3] Huang, Y., *Internet of Everything, Security-based: Challenges and Countermeasures for National Information Security in the Era of Big Data. China New Communication, 22, 21, 155-156, 2020.*

[4] Chen, T., *Research on the dilemma and breakthrough path of ideological and political education in colleges and universities in the era of big data. Journal of Higher Education Research, 3(2), 203-206, 2022.*

[5] Wang, H., Ran, B., *Network governance and collaborative governance: A thematic analysis on their similarities, differences, and entanglements. Public management review, 25(6), 1187-1211, 2023.*

[6] Guo, C., *The triple dimension of network ideological security and precise governance in the era of big data. Journal of Chongqing University of Posts and Telecommunications: Social Science Edition, 33, 5, 7, 2021.*

[7] Chen, T., *Research on the dilemma and breakthrough path of ideological and political education in colleges and universities in the era of big data. Journal of Higher Education Research, 3(2): 203-206, 2022.*

[8] Wang, Y. S., Zhu Y. T., *Challenges to mainstream ideological identity in the age of intelligence and its countermeasures. Journal of Dalian University of Technology: Social Science Edition, 43, 2, 6, 2022.*

[9] Ferreira, M., Martinsone, B., Talić, S., *Promoting sustainable social emotional learning at school through relationship-centered learning environment, teaching methods and formative assessment. Journal of Teacher Education for Sustainability, 22(1): 21-36, 2020.*

[10] Yang, Z. L., *Research on promoting the construction of adolescent ideological risk prevention mechanism in the era of big data. Reform and Opening, 10, 7, 2021.*

[11] Liu, W. X., *Risk and Adjustment: Ideological Security in the Era of Big Data. Journal of the Party School of the CPC Yili Prefecture Committee, 4, 6, 2021.*

[12] Zhong, J., *Research on ideological safety education of college students in the era of big data. University: Ideological and Political Education and Research, 9, 3, 2021.*

[13] Guan, L., *An Analysis of the Way to Improve China's Ideological Discourse Power in the Era of Big Data. New West, 12, 2, 2020.*

[14] Yu, H. R., *Research on China's ideological security in the era of big data. Industry and Technology Forum, 21, 1, 2, 2022.*

[15] Shi, X. N., Lin, L., *The construction of Marxist ideological discourse power in the era of big data. Theory and Modernization, 4, 10, 2021.*

[16] Zhu, Z. L., *Research on ideological safety education of teenagers in the era of big data. Science Education Journal, 26, 2, 2020.*

[17] Yang, Z. L., Zhang, Y. Y., *Prevention of ideological security risks for teenagers in the era of big data. School Party Building and Ideological Education, 8, 3, 2021.*

[18] Chen, J. M., *Big data technology assists the new development of ideological safety education for adolescents: Commentary on "Research on Ideological Safety Education for Adolescents in the Age of Big Data". Journal of Three Gorges University:*

Humanities and Social Sciences Edition, 43, 5, 1, 2021.

[19] Qian, D., *On the ideological safety education of teenagers in the era of big data. Reference for middle school political teaching, 20, 1, 2021.*

[20] Wan, Y., *Risks and prevention of adolescent ideological security in the era of big data. School Party Building and Ideological Education, 3, 3, 2021.*

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Oct 19, 2023
Accepted: Nov 28, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1818–1825, DOI 10.12694/scpe.v25i3.2723

APPLICATION OF BIG DATA ANALYSIS AND INTELLIGENT ALGORITHM IN POWER SYSTEM OPERATION OPTIMIZATION

HUICHAO JIN, JUNYI HUO, QINGFEN WANG, AND DEXIONG LI*

Abstract. The power communication system provides powerful technical support for realizing the intelligent operation and information management of the power grid and improving the operation efficiency and power supply quality of the power grid. Quantum key distribution (QKD) is considered one of the most promising technologies for commercialization. QKD uses a single photon to encrypt data to produce a more secure and reliable password. This paper intends to study the hierarchical, centralized control architecture of power dispatching based on quantum essential supply (QKD). The performance indexes of MDI-QKD under symmetric and asymmetric conditions were studied by local optimization. The optimal key formation rate of the algorithm is analyzed. From the perspective of quantum critical utilization, a quantum key utilization scheme for grid backbone dispatching service is proposed. The dynamic adjustment test of multi-node time slot and service key update rate is carried out. Experiments show that the MDI scheme can effectively improve the effectiveness of a multi-node QKD system. Thus, the security of data transmission of the core business of power dispatching data networks can be ensured to the greatest extent. AMDI can effectively reduce the transmission timeout of low-priority data streams because the delay of high-priority data streams reaches the proportion. It can be an excellent solution to the power system and the password requirements.

Key words: Power dispatching; Quantum key; Dynamic regulation; Intelligent algorithm; Power Grid System

1. Introduction.

With the deepening of quantum cryptography research, its promotion and application have attracted the attention of many enterprises that need high-security performance. Communication is a significant auxiliary means in the operation of the power grid. The power communication system provides powerful technical support for realizing the power grid's intelligent operation and information management, improving the operation efficiency and power supply quality. With the expansion of the scale and scale of the power grid, the power communication system as the monitoring and operation information of the power grid has become more and more complex. The safety of power communication is directly related to the safety of the whole power grid. The need for secure and reliable transmission of network information is more urgent, especially for the power network with UHV network as the core and coordinated development of power grids at all levels.

This paper presents a new power communication technology based on quantum cryptography. The use of quantum cryptography technology to achieve high-capacity and high-rate secure transmission is a research hotspot. Therefore, how to effectively use quantum communication technology is currently a hot topic worldwide. Quantum key distribution (QKD) is considered one of the most promising technologies for commercialization. QKD uses a single photon to encrypt data to produce a more secure and reliable password. There are some essential principles in quantum physics, such as the principle that single photons cannot be divided, the Heisenberg uncertainty relation, the principle of measuring collapse, and the principle of non-cloning. This makes quantum cryptography completely secure in theory. Therefore, the algorithm has a high coding rate. Using quantum communication technology to transmit secret information in practical applications is very difficult. It is difficult to achieve the goal of complete secrecy of grid services. It is necessary to select some essential data streams and use quantum cryptography to encrypt them to ensure data security. In this paper [1], Fuzzy logic is used to adjust the weight of the SMDI algorithm dynamically. This method can adjust the weights according to the delay and throughput rate to improve equity and service quality.

Reference [2] proposes an improved SMDI method. The algorithm adjusts each queue's weight by the buffer size so that the delay between each queue can be balanced. Reference [3] adds a strict priority queue

*Department of Electrical Engineering, Shijiazhuang Institute of Railway Technology, Shijiazhuang, Hebei, 050041, China (Corresponding author, LiDexiong200@163.com)

and the Low Delay queue (LLQ) algorithm. This method can prioritize two types of high-priority services, thus improving service quality. This further increases the likelihood of "hunger" in the low-priority cohort. The existing methods cannot control the network delay directly, so it is difficult to effectively guarantee the service quality of each queue in the network. However, the combined hybrid optimization strategy cannot completely solve the "hunger" problem while ensuring the quality of high-priority queuing service.

2. Modelling and performance optimization of communication system based on MDI. QKD is one of the most practical and promising security technologies. This will provide critical performance optimization for MDI protocols currently in the limelight. The difference with the regular QKD protocol is that there are two sending ends in MDI: Alice and Bob. They send a signal that Charles can detect [4]. If Charles is between Alice and Bob, call it symmetric MDI (SMDI). If Charles is offset at the midpoint, call it an asymmetric MDI (AMDI). Most of them are AMDI in real life. Through the research of this project, it will get the key generation rate of MDI protocol in the case of finite and infinite samples. An infinite set of solutions is an ideal finite solution. In addition, from the existing results, the number of decoy states selected in this paper is 2.

The security key rate obtained by quantum state preparation, transmission, Bell state determination, quantum state screening, parameter estimation, error correction, and privacy amplification in MDI is as follows:

$$S \geq g_d[D_{11}^{C,E}(1 - F_2(e_{11}^{X,V})) - \hat{D}_{\eta_\alpha\eta_\beta}^{C} g_e F_2(\hat{K}_{\eta_\alpha\eta_\beta}^{C})] \tag{2.1}$$

Where g_d is the convention coefficient. The protocol coefficient is $g_d = 1$ for infinite sets. C is a finite set. Here $U_{\eta_\alpha}, U_{C|\eta_\alpha}$ and $U_{\eta_\beta}, U_{C|\eta_\beta}$ are the probability that Alice and Bob transmit the signal state, and the probability that $g_d = U_{\eta_\alpha}U_{C|\eta_\alpha}U_{\eta_\beta}U_{C|\eta_\beta}$ password is selected in this state. Where g_e represents the error correction factor [5]. Where F_e is the binary Shannon entropy. $\hat{D}_{\eta_\alpha\eta_\beta}^{C}$ represents the total amount of detection obtained when both Alice and Bob choose the emission state based on C. Where $\hat{K}_{\eta_\alpha\eta_\beta}^{C}$ is the corresponding bit error. When Alice and Bob both choose the C group to emit a single photon state, $D_{11}^{C,E}$ is the lower bound of the probe. Where $e_{11}^{X,V}$ is the upper bound for detecting the bit error rate under the corresponding X base. $\hat{D}_{\eta_\alpha\eta_\beta}^{C}$ and $\hat{K}_{\eta_\alpha\eta_\beta}^{C}$ are determined by test. In this paper, the superscale ĉan be replaced by the theoretical value provided in the linear channel model of MDI system. The absence of îndicates a corrected data error. These two values are the same regardless of statistical fluctuations [6]. The following analytical formula is obtained by using the Gaussian elimination method:

$$\begin{cases} D_{11}^{C,E} = \eta_\alpha\eta_\beta e^{-(\eta_\alpha+\eta_\beta)} Y_{11}^{C,E} \\ e_{11}^{X,V} = \dfrac{1}{(\lambda_\alpha - \kappa_\alpha)(\lambda_\beta - \kappa_\beta)Y_{11}^{X,E}} (\hat{K}_{\lambda_\alpha\lambda_\beta}^{X}\hat{D}_{\lambda_\alpha\lambda_\beta}^{X} e^{\lambda_\alpha+\lambda_\beta} + \\ \hat{K}_{\kappa_\alpha\kappa_\beta}^{X}\hat{D}_{\kappa_\alpha\kappa_\beta}^{X} e^{\kappa_\alpha+\kappa_\beta} - \hat{K}_{\lambda_\alpha\kappa_\beta}^{X}\hat{D}_{\lambda_\alpha\kappa_\beta}^{X} e^{\lambda_\alpha+\kappa_\beta} - \hat{K}_{\kappa_\alpha\lambda_\beta}^{X}\hat{D}_{\kappa_\alpha\lambda_\beta}^{X} e^{\kappa_\alpha+\lambda_\beta}) \end{cases} \tag{2.2}$$

η_α and η_β are the average number of photons in Alice's signal state, C is the average number of particles in Bob's decoy state, and D is the average number of particles in the vacuum state. $\hat{K}_{\lambda_\alpha\lambda_\beta}^{X}, \hat{D}_{\lambda_\alpha\lambda_\beta}^{X}$ is the total bit error when Alice and Bob choose the X base to transmit the deception state. $\hat{K}_{\kappa_\alpha\kappa_\beta}^{X}, \hat{D}_{\kappa_\alpha\kappa_\beta}^{X}$ is the overall bit error for Alice and Bob to choose the X detector.

The optimal solution is obtained by using the above two methods under infinite sets. Then, the key generation rate of the MDI system and its optimal configuration parameters are given for a particular channel length [7]. The optimal parameters contained in SMDI are η, λ and $\kappa = 5e^{-4}$. Because the positions of the two systems are symmetric, the structural parameters are consistent. The best parameter to consider for AMDI is $\eta_\alpha, \eta_\beta, \lambda_\alpha, \lambda_\beta$.

The statistical jitter effect of the measured variable must be taken into account under the finite set condition. φ simple Gaussian variance method is chosen in this paper [8]. Take A as the standard deviation. The limit value $\sigma = 1 - erf(\varphi/\sqrt{2})$ of the safety factor is obtained. Where $erf(\cdot)$ is the error function. The flutter coefficient is determined as $\delta_d = \varphi/\sqrt{\hat{D}_{d_\alpha d_\beta}^{C} N_{d_\alpha d_\beta}^{C}}$, $\delta_{ed} = \varphi/\sqrt{\hat{K}_{d_\alpha d_\beta}^{X}\hat{D}_{d_\alpha d_\beta}^{X} N_{d_\alpha d_\beta}^{X}}$. Then you can get the upper

Table 2.1: Model input characteristics.

Agreement	$\boldsymbol{E}$	$\boldsymbol{D_E}$	$\boldsymbol{K^C_{\eta\eta}}$	$\boldsymbol{D^C_{\eta\eta}}$	$\boldsymbol{Y^C_{11}}$	$\boldsymbol{D^C_{11}}$	$\boldsymbol{e^X_{11}}$	$\boldsymbol{R}$
SMDI	125	0	2.6175×10^{-2}	2.72×10^{-6}	3.1521×10^{-5}	1.3956×10^{-6}	9.2665×10^{-2}	1.152×10^{-7}
AMDI	115	10	2.3328×10^{-2}	5.1479×10^{-6}	4.9473×10^{-5}	2.4778×10^{-6}	9.001×10^{-2}	1.9118×10^{-7}
AMDI	105	20	2.1775×10^{-2}	8.8574×10^{-6}	7.927×10^{-5}	4.2183×10^{-6}	8.8992×10^{-2}	3.7094×10^{-7}
AMDI	95	30	2.0727×10^{-2}	1.55×10^{-5}	1.2843×10^{-4}	7.3297×10^{-6}	9.2938×10^{-2}	6.4358×10^{-7}

and lower bounds of the measurement results needed to calculate the key rate:

$$\begin{cases} \hat{D}^C_{d_\alpha d_\beta}(1-\delta_d) = D^C_{\;\;{}_-\,d_\alpha d_\beta} \leq D^C_{d_\alpha d_\beta} \leq \bar{D}^C_{d_\alpha d_\beta} = \hat{D}^C_{d_\alpha d_\beta}(1+\delta_d) \\ \hat{K}^X_{d_\alpha d_\beta}\hat{D}^X_{d_\alpha d_\beta}(1-\delta_{ed}) = K^X_{\;\;{}_-\,d_\alpha d_\beta} D^X_{\;\;{}_-\,d_\alpha d_\beta} \leq K^X_{d_\alpha d_\beta} D^X_{d_\alpha d_\beta} \leq \\ \bar{K}^X_{d_\alpha d_\beta}\bar{D}^X_{d_\alpha d_\beta} = \hat{K}^X_{d_\alpha d_\beta}\hat{D}^X_{d_\alpha d_\beta}(1+\delta_{ed}) \end{cases} \tag{2.3}$$

The superscript represents the upper bound of the corresponding parameter, and the subscript represents the lower bound of the corresponding parameter. d forη, λ, κ. Replace the upper and lower bounds of (3) with the optimal solution of (2) and (1). The key generation speed of the MDI protocol with a particular channel length is given. The upper and lower bounds are chosen based on the worst-case performance evaluation principle. This is true even when $D^{C,E}_{11}$ becomes smaller and $e^{X,V}_{11}$ becomes larger in the equation (2). At this time, even if $(\kappa, U_{C,\kappa})$ is at rest, SMDI has $\eta, \lambda, U_\eta, U_\lambda, U_{C|\eta}, U_{C|\lambda}$ optimal parameters. However, even if $(\kappa, U_{C,\kappa})$ is constant in AMDI, there are still $\eta_\alpha, \eta_\beta, \lambda_\alpha, \lambda_\beta, U_\eta, U_\lambda, U_{C|\eta}, U_{C|\lambda}$ optimal parameters when the probability coefficients of the two variables are the same.

The optimal analysis of MDI system performance is carried out with $\lambda = 1550nm$ as the light source. The loss of the fiber is $\alpha = 0.2dB/km$. The probability that a photon is projected by the wrong detector is $e_d = 1.5\%$. The dark count of the detector is $Y_0 = 6.02 \times 10^{-6}$. The quantum utilization rate of the device is $\eta_d = 0.145$. The bidirectional error correction factor is $g_e = 1.16$. The safety factor is $\sigma = 5.73 \times 10^{-7}$. Its standard deviation is $\varphi = 5$. The limited data set is $N = 10^{14}$. The light source has a pulse frequency of $g = 10^9 Hz$. Here, the values of N and $g$are large compared to the traditional DS protocol [9]. This is mainly because there are two receivers in the MDI system, and the detection of the Bell state requires the quantum state transmitted by the receiver to conform to a particular entangled state, leading to the MDI system's low-key generation rate. This problem can be effectively solved by increasing N and g. The LSA method is used to optimize its configuration and indexes in detail. The optimized results are compared with those in the literature.

Figure 2.1 shows the best key bit rates for various MDI protocols with different channel lengths. The universal rate value is obtained by multiplying the optimal key rate calculated by formula (1) with the light source pulse frequency g. Table 2.1 and 2.2, respectively, list the optimal features and configurations of each MDI protocol in the specified channel. The optimal index is the detection rate of signal and photon state alone and the bit error rate of detection. The optimal configuration parameters are the signal-to-noise ratio, the number of deception photons and their corresponding configuration parameters. E represents the entire length. D_E represents the distance difference between each end and the detector.

It can be seen from Fig. 2.1 that the calculated security critical distance of SMDI and the three AMDI is 159,157,155,149 km under a finite set of selected security boundary parameters. The maximum at infinite sets is 201 kilometers. When the distance difference between the two transmitting terminals is not very large, AMDI parameters are comprehensively optimized [10]. The critical transfer rate over a distance of 100 km can be obtained, and the result is comparable to SMDI. In addition, this paper also compares the best critical generation speeds obtained by similar algorithms under different device parameters. Compared with DS, the optimal critical generation speed of MDI is about 1/10 of DS. But its crucial generation distance is relatively long.

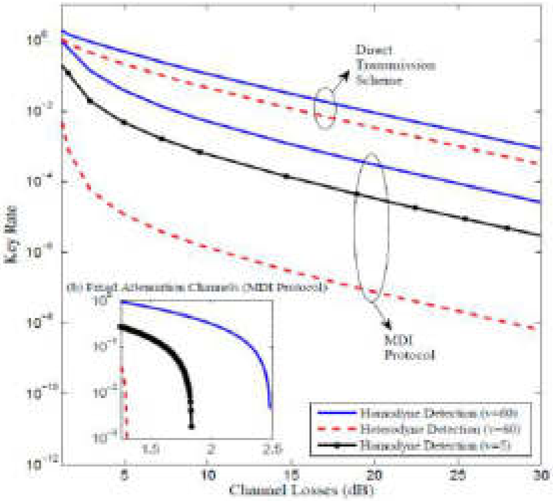


Fig. 2.1: Optimized key bit rate of the MDI protocol.

Table 2.2: Model input characteristics.

Agreement	E	D_E	η_α	λ_α	η_β	λ_β	U_η	U_λ	$U_{C\|\eta}$	$U_{C\|\lambda}$	$U_{C\|\kappa}$
SMDI	125	0	0.289	0.075	0.701	0.235	1.031	0.373	0.253		
AMDI	115	10	0.371	0.083	0.271	0.068	0.642	0.295	1.031	0.414	0.284
AMDI	105	20	0.384	0.084	0.286	0.054	0.655	0.274	1.031	0.420	0.297
AMDI	95	30	0.403	0.063	0.302	0.043	0.657	0.265	1.031	0.425	0.304

3. Research on the application scheme of quantum key based on MDI. Figure 3.1 shows a design unit pattern for abstracting PDDN based on hierarchical centralized control architecture (Picture quoted from Review of Modelling and Simulation Methods for Cyber-Physical Power System). It is cascaded vertically and extended horizontally to form a complex power-dispatching network with multiple structures utilizing reliable Repeaters. The primary station in this table governs two secondary stations and an attached power plant and substation [11]. The solid lines between all the stations are divided into two: one is a quantum line, and the other is a classical line. It is used to transmit and test quantum states in MDI protocol. The dotted line between stations is a typical channel, which realizes typical post-processing and meets the requirements of the protocol. The spacing of stations in this table is determined according to the actual wiring situation of a network province. This paper mainly discusses using the MDI quantum key with a higher security level. If this type of quantum bond is insufficient, it is automatically converted to a lower-level DS or classical bond.

If A type n_s service exists between a pair of QKD nodes, then the number of remaining keys D_t in the corresponding key pool in time window ϕ refers to the difference between the number of keys U_t generated and the number of keys A_t used during this period:

$$D_t = U_t(i) - A_t(i) = R'_t(i)\phi - \sum_{j=1}^{n_s} \frac{N_t(j)}{g_t(j)} V_t(j) \tag{3.1}$$

$N_t(j)$ represents the number of packets to be encrypted in ϕ. $V_t(j)$ represents the number of quantum keys to be used. $g_t(j)$ indicates the update frequency of the service quantum key. Where $R'_t(i)$ is the time-equivalent bit rate of quantum key distribution for the corresponding lines. For n_τ nodes working in TDM mode, the

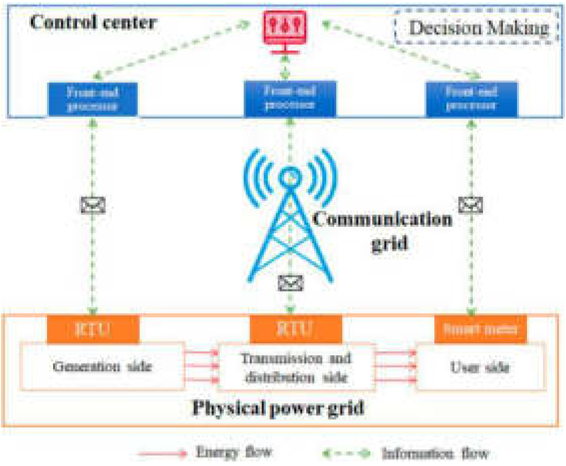


Fig. 3.1: PDDN design unit model.

optimal time interval of each node can be calculated by the following formula:

$$\tau_i = \frac{\frac{A_t(i)}{R_t(i)}}{\sum\limits_{i=1}^{n_\tau} \frac{A_t(i)}{R_t(i)}}\phi \tag{3.2}$$

After using TDM technology, the quantum key distribution rate between each pair of nodes can be equivalent modified as:

$$R_t'(i) = R_t(i)\frac{\tau_i}{\phi} \tag{3.3}$$

Under the condition that the time window after TDM remains unchanged, the key generation and elimination between each node pair obtained by the above method will change [12]. The key generation elimination parameter ζ_{PC} is defined to standardize the characteristics of the key imposed policy. The formula is:

$$\zeta_{PC} = \frac{U_t(i)}{A_t(i)} \tag{3.4}$$

The value of ζ_{PC} should be one or greater. The closer the coefficient is to 1, the more efficient its application in QKD and the better its long-term stability. This design and experimentation will make ζ_{PC} as close to 1 as possible when business data traffic characteristics are determined [13]. The optimal time interval is determined by formula (5). The formula can determine the frequency interval of quantum bond renewal (4) to (6) according to the following process.

4. Simulation analysis. The dynamic adjustment performance is tested in detail by an example. This project aims to test this method's performance under different conditions, especially under abnormal conditions. Based on the Matlab2014a simulation system, four queues are selected for simulation and analysis. Assume that the transmission delay for each queue is $T_1 = 100ms, T_2 = 20ms, T_3 = 300ms, T_4 = 40ms;$. The key length required to encrypt the packet to be encrypted is $L\prime = 128bit$. A quantum critical generation method based on $R = 1\ Mbit/s$ is proposed. Packets queued for encryption obey Poisson distribution, and the arrival rate of packets queued for encryption is the same [14]. The total requirement for the four queue keys is $1Mbit/s$. Because the delay generated by the packet during transmission is much lower than the demand for its transmission delay under normal conditions, this delay can be ignored in the simulation process. Figure 4.1 shows the overall system process, including the application and dynamic adjustment of QKey in practice (Picture quoted from Appl. Sci. 2019, 9(10), 2081).

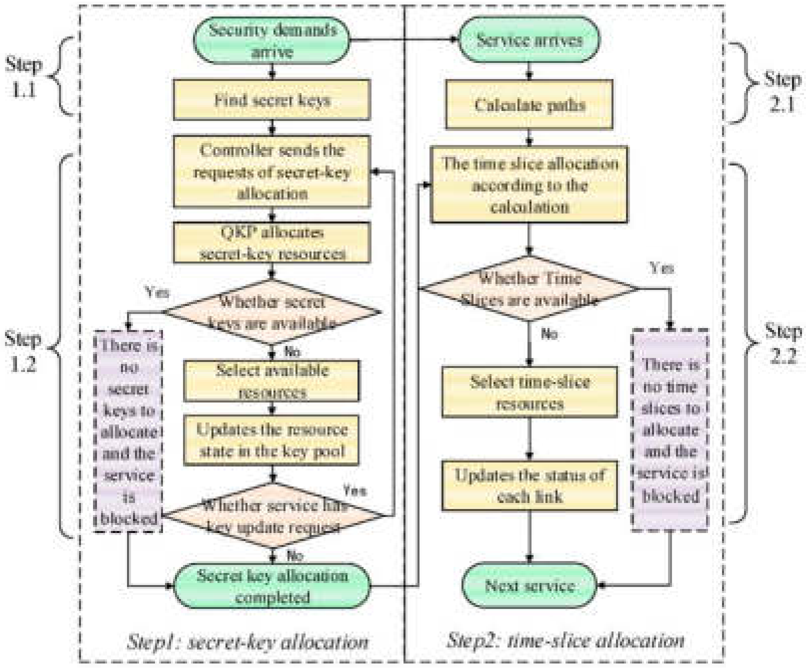

Step 1.1
Security demands arrive
Find secret keys
Controller sends the requests of secret-key allocation
QKP allocates secret-key resources
Whether secret keys are available
Yes
No
There is no secret keys to allocate and the service is blocked
Select available resources
Updates the resource state in the key pool
Whether service has key update request
Yes
No
Secret key allocation completed
Step 1.2
Step1: secret-key allocation
Service arrives
Calculate paths
The time slice allocation according to the calculation
Whether Time Slices are available
Yes
No
Select time-slice resources
Updates the status of each link
There is no time slices to allocate and the service is blocked
Next service
Step 2.1
Step 2.2
Step2: time-slice allocation

Fig. 4.1: Quantum essential application strategy design and dynamic adjustment process.

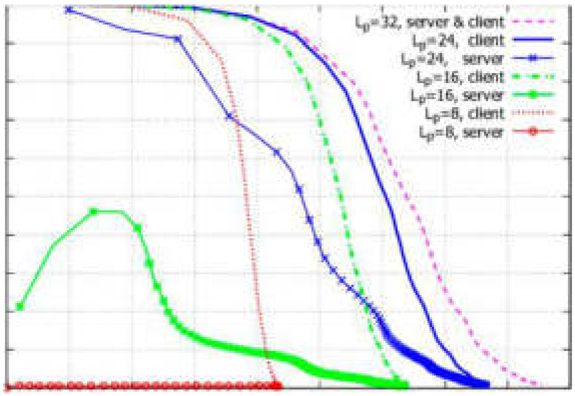


Fig. 4.2: Encryption delay of each queue under SMDI and AMDI algorithms.

The weight $\begin{matrix} w_1 = 0.3258, w_2 = 0.2504, \\ w_3 = 0.2205, w_4 = 0.2033 \end{matrix}$ for each set of columns is obtained from formula (1). Each queue received 3000 packets to be encrypted during the simulation. Due to the complexity of the queuing scheduling system and the strong randomness of packet forwarding delay of the data to be encrypted, the average value of 30 simulation results was selected and compared with the SMDI algorithm and the algorithm proposed in reference (Fig. 4.2).

The SMDI algorithm can obtain more scheduling probability when executing tasks because of the higher-weight queue. Moreover, the shorter the time interval for encrypted packets in the queue, the lower the sending delay. Queues with priority four will have a more fantastic exit time for encrypted packets because they have fewer weights. During the queuing process, the network is blocked because of the lack of effective processing [15].

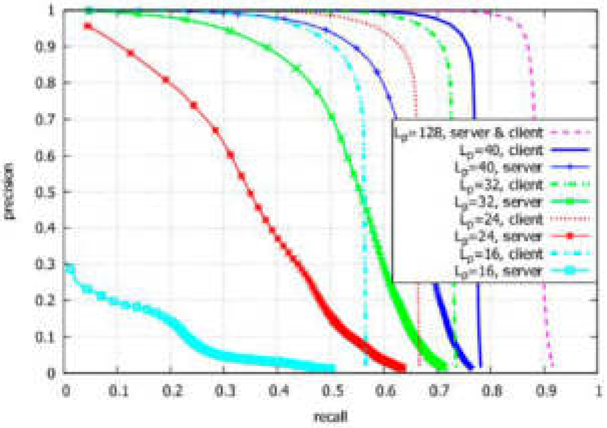


Fig. 4.3: Statistics on the number of timeout data packets.

Finally, the transmission delay of the encrypted information packet exceeds the transmission delay requirement. When the queue length is considerable, the algorithm proposed in reference can appropriately increase the weight of the queue. When queue 4 is congested, it can improve its chances of obtaining a schedule. This algorithm can effectively reduce the queue waiting time but will lead to a slightly longer queue waiting time. Although using this method can make the delay of each queuing system better balanced, it still cannot prevent the phenomenon of a delay exceeding the threshold in the minor queuing system.

The method is simulated by using continuously varying packet arrival rates to explore the optimality of the method to the system performance at different packet arrival rates. The average value of 30 results is calculated to reduce the errors caused by the high randomness of the queuing planning system. In Figure 4.3, the critical requirement rate is the horizontal axis [16]. The aim is to show the effect of the algorithm more directly and reflect the relationship between the critical request rate and the key generation rate. Its value is the number of essential requirements generated at intervals when four queues arrive at a fixed rate.

Because the number of timeout packets in queue 1 is small, it is not listed in item 5. Whereas queue 2 has more time to get more time, the average result of its 30 simulations is still very random. For the rest of the queue, the proportion of timeout packets changes fairly smoothly [17]. Although the algorithm proposed in reference can balance the delay between all queues well, it cannot constrain and optimize the transmission delay with targets effectively. Although the proportion of timeout packets can be reduced by balancing the delay, the optimization results of this method are not significant under the condition of high vital requirements, and its performance is not ideal under the condition of high vital requirements.

The algorithm in this paper can reduce the number of timeout packets. This method can effectively reduce the timeout rate of the system. This method does not significantly affect the timeout rate of high-priority queuing systems while ensuring the transmission delay of queuing systems 3 and 4 to the maximum extent. At the same time, it can effectively reduce the timeout ratio of data packets. Simulation results show that the proposed method has a good delay balance for each queued message waiting for encryption. It can reduce the timeout rate of packets to be encrypted to some extent. The algorithm has good performance for all kinds of crucial requirement rates. In power communication networks, the correctness of the password is more important than the delay. This method can increase the delay under delay tolerance to obtain a higher delay arrival rate. It can solve the cryptographic problem in power communication and improve the efficiency of the quantum key.

5. Conclusion. Due to its limited QKD coding rate, it is only suitable for some essential power services in power communication, so optimizing the quantum cryptographic rate to achieve a reasonable bandwidth allocation is necessary. This paper establishes a quantum communication algorithm based on queueing sort. The method can predict the encrypted packets and schedule the packets that are about to time out preferentially to reduce the timeout of the packets to be encrypted. And it does not significantly reduce the proportion of packets queued for high priority to be encrypted. When the encoding rate is constant, AMDI can better solve the delay problem of encrypted data. It can implement efficient encryption for more packets to be encrypted. The algorithm can also effectively improve the system's low efficiency of quantum key coding. The simulation results show that AMDI can effectively reduce the transmission timeout of low-priority data streams to ensure the delay ratio of high-priority data streams. It can be a good solution for the power system and password requirements.

REFERENCES

[1] Kuang, R., & Perepechaenko, M. Quantum encryption and decryption in IBMQ systems using quantum permutation pad. J. Commun, 2022; 17(12): 972-978.

[2] Sehgal, S. K., & Gupta, R. SOA Based BB84 Protocol for Enhancing Quantum Key Distribution in Cloud Environment. Wireless Personal Communications, 2023; 130(3): 1759-1793.

[3] Domi Caroline, S. Quantum Key Distribution Algorithm for Network Security. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 2021; 12(9): 2277-2284.

[4] Cheng, Z., Ye, F., Cao, X., & Chow, M. Y. A homomorphic encryption-based private collaborative distributed energy management system. IEEE Transactions on Smart Grid, 2021; 12(6): 5233-5243.

[5] Yu, X., Liu, Y., Zou, X., Cao, Y., Zhao, Y., Nag, A., & Zhang, J. Secret-Key Provisioning with Collaborative Routing in Partially-Trusted-Relay-based Quantum-Key-Distribution-Secured Optical Networks. Journal of Lightwave Technology, 2022; 40(12): 3530-3545.

[6] Yan, R., Wang, Y., Dai, J., Xu, Y., & Liu, A. Q. Quantum-key-distribution-based microgrid control for cybersecurity enhancement. IEEE Transactions on Industry Applications, 2022; 58(3): 3076-3086.

[7] Al-Balushi, M. A., Alomairi, S. A., & Okedu, K. E. Power situation in Oman and prospects of integrating smart grid technologies. Int. J. on Smart Grid, 2021; 5(1): 45-62.

[8] Liu, Z. Y., Tseng, Y. F., Tso, R., Mambo, M., & Chen, Y. C. Public-key authenticated encryption with keyword search: A generic construction and its quantum-resistant instantiation. The Computer Journal, 2022; 65(10): 2828-2844.

[9] Nematkhah, F., Aminifar, F., Shahidehpour, M., & Mokhtari, S. Evolution in Computing Paradigms for Internet of Things-Enabled Smart Grid Applications: Their Contributions to Power Systems. IEEE Systems, Man, and Cybernetics Magazine, 2022; 8(3): 8-20.

[10] Yan, Y., Liu, Y., Fang, J., Lu, Y., & Jiang, X. Application status and development trends for intelligent perception of distribution network. High Voltage, 2021; 6(6): 938-954.

[11] Mittal, S., & Ramkumar, K. R. Research perspectives on fully homomorphic encryption models for cloud sector. Journal of Computer Security, 2021; 29(2): 135-160.

[12] Wasumwa, S. A. Safeguarding the future: A comprehensive analysis of security measures for smart grids. World Journal of Advanced Research and Reviews, 2023; 19(1): 847-871.

[13] Wang, J., Hong, Y., Wang, J., Xu, J., Tang, Y., Han, Q. L., & Kurths, J. Cooperative and competitive multi-agent systems: From optimization to games. IEEE/CAA Journal of Automatica Sinica, 2022; 9(5): 763-783.

[14] Cai, B. B., Wu, Y., Dong, J., Qin, S. J., Gao, F., & Wen, Q. Y. Quantum Attacks on 1K-AES and PRINCE. The Computer Journal, 2023; 66(5): 1102-1110.

[15] Paramguru, J., & Barik, S. K. Implementation of β-chaotic mapping to improved elephant herding optimisation to dynamic economic dispatch problem. International Journal of Innovative Computing and Applications, 2022; 13(2): 115-125.

[16] Valdez, F., & Melin, P. A review on quantum computing and deep learning algorithms and their applications. Soft Computing, 2023; 27(18): 13217-13236.

[17] Zhang, C., Wu, J., Huang, Y., Jiang, Y., Dai, M. Z., & Wang, M. Constructive schemes to spacecraft attitude control with low communication frequency using sampled-data and encryption approaches. Aircraft Engineering and Aerospace Technology, 2021; 93(2): 267-274.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Oct 24, 2023
Accepted: Oct 30, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1826–1832, DOI 10.12694/scpe.v25i3.2742

TEXT CLASSIFICATION AND CLUSTER ANALYSIS BASED ON DEEP LEARNING AND NATURAL LANGUAGE PROCESSING

HUA HUANG [*]

Abstract. At present, the commonly used Bag of Words (BOW) expression ignores the semantic information of text and the problems of high dimension and high sparsity of feature extraction. This paper presents a multi-class text representation and classification algorithm. This project is based on the vector expression of keywords and takes the multi-category classification problem as the research object. Then, a hybrid Deep Location network (HDBN) is constructed by combining DBN with Boltzmann (DBM). Then, this paper does a lot of tests on the algorithm and proves the effectiveness of the algorithm. In addition, the 2D visual experiment is carried out with HDBN, and then the high-level text expression based on HDBN is obtained. The expression has strong cohesion and weak coupling.

Key words: Text classification; Deep belief network; Deep learning; Deep Boltzmann machine network

1. Introduction. Under ”information overload,” managing and screening information effectively is an urgent problem. Text is the primary way for people to get information on the Internet. Using the method of word classification can solve various complicated problems well to help users find the information they need better [1]. Text must first be converted into a readable format to realize automatic recognition of text. Text expression is the most essential part of the whole text recognition process, and its correctness directly affects the whole system's performance. Most existing text expressions are based on lexical packages (BOW) and vector Spaces (VSM). The default words are independent of each other, and the correlation between semantics is ignored. However, due to the diversity of text types and the elaboration of topics, such shallow text expression lacks the semantic meaning of the text itself, and it is difficult to cope with the current complex classification problem. The continuous development of deep learning technology provides a new opportunity for the development of character recognition technology. The project research results in this field will provide new ideas and methods for large-scale data analysis [2]. Deep learning has been widely used in many problems, such as data compression, object detection and tracking, information retrieval, machine translation and speech recognition. Deep learning technology can better relate to specific questions to uncover the complex semantic connections hidden in the text [3]. At the same time, massive data training and processing capabilities have been greatly improved with the expansion of network scale and the rapid development of multimedia networks. This opens up new opportunities for deep learning.

2. Overview of text classification.

2.1. Concept Analysis.

Text classification is classifying and marking a text according to a specific system and criteria [4]. The so-called ”text characteristics” refers to words closely related to the text and can express the work.

2.2. Development of text classification.

Word classification is a typical problem in natural language processing. In the early 1950s, character recognition research mainly used expert judgment. This requires human intervention [5]. This inevitably affects the efficiency of retrieval. With the rapid development of network technology since the 1980s, a large amount of text data has been used for processing. While statistical and computer-aided algorithms have emerged to deal with these problems, these algorithms still stay in the traditional manual processing and single modes. KNN, Naive Bayes, neural networks, decision trees, support

[*] School of Computer and Artificial Intelligence, Henan Finance University. Zhengzhou, Henan, 450046, China (Corresponding author, hafuhuanghua@163.com)

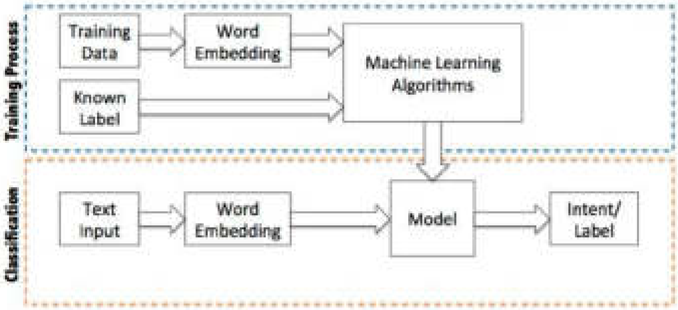


Fig. 2.1: Training flow of training text classifier.

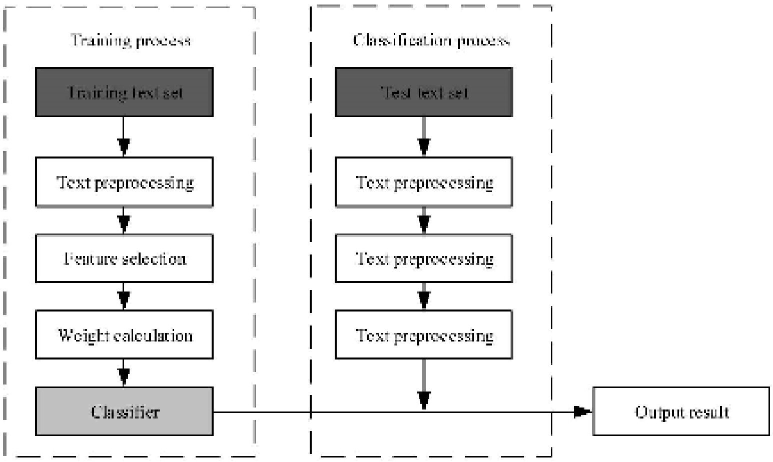


Fig. 2.2: Flow chart of text classification.

vector machines (SVM), etc. The problem is then decomposed into two main steps: one is featuring extraction, and the other is classifier design [6]. The process of training the text classifier is shown in Figure 2.1. The core of text classification lies in the selection of feature values and the design of algorithms, in which the feature quantities used include information gain, text frequency, mutual information, CHI ($\chi 2$) and so on. It is compared with the data to be tested to determine the text category. These algorithms are prone to problems such as small samples, local overfitting and local optimality, and need dimensionality reduction. This leads to the loss of text and the reduction of recognition accuracy.

2.3. Process of text classification. Text classification is divided into two stages: one is the learning stage, and the other is the classification stage. A machine learning algorithm based on a deep neural network is proposed. The text classification process is shown in Figure 2.2. The input text is standardized data. It is a computer-expressible form, that is, the text vector. A text-based automatic recognition method is proposed [7]. The recognition model is learned and trained to get the model parameters using the text training set method. Experimental results show that this method has good learning performance. An adaptive learning algorithm based on a neural network is proposed and dynamically adjusted to improve the learning effect of the classifier.

3. Natural language text processing model based on deep learning. Firstly, ICTCLAS software is used to slice the original text and eliminate invalid words to obtain the characters needed for the experiment. The traditional TF-IDF algorithm is used to solve the weight of each characteristic word [8]. Construct the original feature matrix of text. Assume that each text has n properties. In this way, an n-dimensional vector space is formed, and a characteristic vector of n-dimensions can represent each literal s:

$$U(s) = (R_1, E_1(s); R_2, E_2(s); \cdots\cdots; R_n, E_n(s)) \tag{3.1}$$

R_i is a segmented word of text. Where$E_i(s)$ is the weight of R_i in the text D. Use the TF-IDF formula to calculate the weight of text segmentation:

$$e_i(s) = \frac{TF(t_i) \times IDF(t_i)}{\sqrt{\sum\limits_{i=1}^{n} (TF(t_i) \times IDF(t_i))^2}} = \frac{TF(t_i) \times \log(\frac{N}{n_i} + D)}{\sqrt{\sum\limits_{i=1}^{n} (TF(t_i) \times \log(\frac{N}{n_i} + D))^2}} \tag{3.2}$$

where $e_i(s)$ is the weight of eigen term R_i. $TF(t_i)$is the frequency with which the eigen b is used in the s sentence. Where N represents the total number of samples. n_i is the number of samples that occur in R_i.

3.1. Automatic recognition of text. The existing SVM algorithm and BP neural network algorithm have significant differences in the recognition accuracy of different samples due to the interference of sampling data. Text recognition based on a deep confidence network can be divided into two stages: pre-training artificial neural network and network adjustment [9]. Most existing classification methods use dimensionality reduction to avoid dimensionality disaster, while deep belief networks (DBN) can extract low-dimensional features with strong discrimination ability from massive original features. In this way, the classification model can be built directly without dimensionality reduction. Meanwhile, it fully uses the rich information in the text. The weights of each BP neural network level are initialized using DBN network weights [10]. This method does not need to initialize any initial value of DBN, nor does it need to extend the BP neural network. BP neural network is used for global optimization to solve the local extreme value problem caused by DBN's randomness of weight parameters.

3.2. DBN Pre-Learning. A deep confidence Network (DNN) is a deep nonlinear network. The underlying information is fused by constructing the learning mode of multiple implicit levels. This creates more abstract high-level features to recognize text effectively. Suppose $F$is a system that includes n layer $(F_1, F_2, \cdots, F_n)$, if G is used to represent the input and P is used to represent the output. It can be expressed in $G \geq F_1 \geq F_2 \geq \cdots \geq F_n \geq P$ to continuously adjust the parameters in the system [11]. The result of the system is still input G, and then we can automatically obtain the hierarchical property of input G, which is $F_1, F_2, \cdots, F_n$. DBN is a probability-based modeling method that assigns observed samples and tags jointly. The DBN is formed by stacking layer upon layer of constrained Boltzmann machines (RBMS). RBM is a representative neural network (Figure 3.1 cited in Deep neural Networks (Part IV). Creating, training and testing a model of Neural Networks).

The RBM model is divided into two levels: one is the visual layer, usually the input layer, and the second is the implicit layer, usually called feature extraction [12]. Learning the neurons of the hidden layer in the visual-hidden layer can capture the higher-order association information presented by the video layer. Where is the weight of the visible layer and the hidden layer. is the displacement of the nodes of the visible layer. is the displacement of the node of the hidden layer. Where is the state vector of the node of the visual layer. Where is the state vector of the node of the hidden layer. In the process of BP network learning, the greedy algorithm is used for hierarchical learning of each layer of RBM. After learning the RBM of the previous layer, it is used to learn the RBM of the next layer, and so on, finally forming a complete DBN network (Figure 3.2).

RBM is an energy-based model that combines the visible layer variable u and the hidden layer variable l in RBM. Its energy expression is shown in Figure 3.1:

$$Q(u, l|\beta) = \frac{1}{2}(u^T el + \varepsilon^T u + \sigma^T l) \tag{3.3}$$

$\beta = (e, \varepsilon, \sigma)$ is a parameter combination. After the values of each parameter are given, the standardized coefficient of RBM is $C(\beta) = \sum_{u,l} e^{-Q(u,l|\beta)}$. According to this energy equation, the joint probability distribution of (u, l) can be obtained as follows:

$$p(u, l|\beta) = \frac{e^{-Q(u,l|\beta)}}{C(\beta)} \tag{3.4}$$

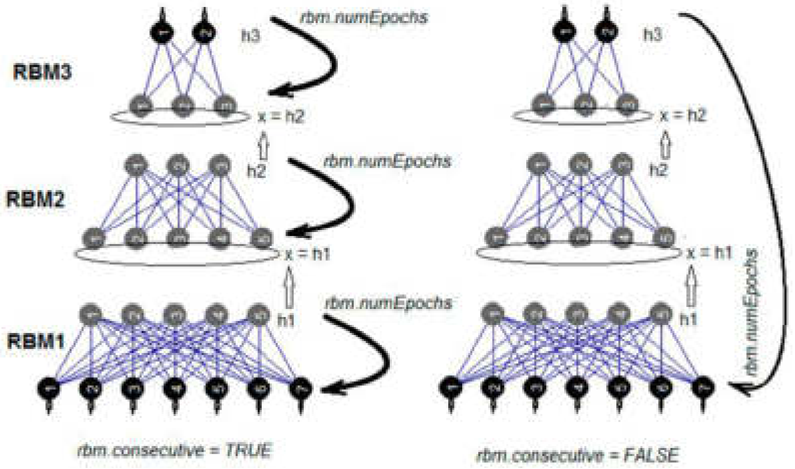


Fig. 3.1: Neural structure of RBM.

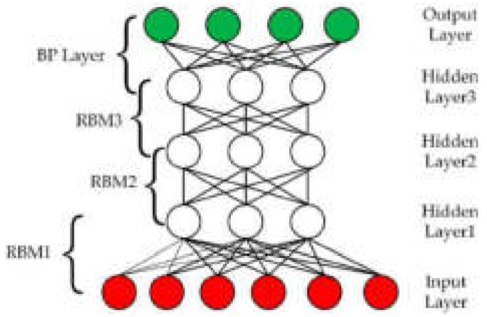


Fig. 3.2: DBN network structure.

There are several possibilities for the nodes of the hidden layer:

$$p(l_j = 1|u) = \varphi(\sigma_j + \sum{}_i u_i e_{ij}) \tag{3.5}$$

There are the following possibilities for visual layer nodes:

$$p(u_i = 1|l) = \varphi(\varepsilon_i + \sum{}_j e_{ij} l_j) \tag{3.6}$$

The learning essence of the RBM method is to find a probability distribution that can generate training samples to the greatest extent [13]. In other words, you need a distribution that produces the most significant number of possibilities. Because weight A is the key to influencing the probability distribution, we will learn weights based on probability graphs to learn the underlying model. This paper presents a fast algorithm called ”contrast branching.” This method only repeats the cycle in B cycles and gets a model estimate of 1. CD method first uses training samples to initialize the visual layer and then uses the conditional probability method to find the hidden layer [14]. The visual layer is obtained from the hidden layer by the conditional distribution. The result is a reconstruction of the input. The visual layer generates a vector C and transmits the value to the hidden layer through this vector. Inputs at the corresponding visual level are randomly selected to recover the original input. Finally, the visualized neuron reconstructs the neuronal activity unit l in the network through forward conduction [15]. The adjustment of weights is determined according to the degree of correlation between the hidden layer's active cells and the visible layer's input end. The model is solved according to the CD algorithm

$$\Delta e_{ij} = \zeta(\langle u_i l_j \rangle_{data} - \langle u_i l_j \rangle_{recon}) \tag{3.7}$$

ζ is how much students learn. $\langle u_i l_j \rangle_{data}$ represents the expected value of the sampled data. $\langle u_i l_j \rangle_{recon}$ represents the expected value of the reconstructed visualized data. The pre-training procedure for DBN follows the following steps:

1) The greedy algorithm is used to learn the first RBM.

2) Determine the weight and bias of the first RBM and use the calculated results as input to the upper RBM;

3) Repeat the above process several times until the reconstruction errors are minimized. The hidden layer can then become input to the visual layer.

E, a, b The specific steps of the DBN pre-training algorithm are as follows:

The input training: sample x_0, number of visible layer and hidden layer units n, m, learning rate ζ, and maximum training cycle R.

The output training: weight matrix e, visible layer bias a and hidden layer bias b.

Step 1. Initialize the initial state$u_1 = x_0$ of the visibility unit. E, a, b is a small arbitrary number.

Step 2. The iterative training period is t.

Step 3. The hidden layer l_1 is calculated from the visible layer u_1. The value of $P(l_{1j} = 1|u_1)$ is periodic, and probability is used as the probability of hiding the j cell of the layer.

Step 4. The visible layer u_2 is calculated from the hidden layer l_1. The value of $P(u_{2i} = 1|l_1)$ is computed cyclically, and this possibility is given as the possibility that the i unit of the visible layer is set to 1.

Step 5. The visible layer l_2 is calculated from the hidden layer u_2. The value of $P(l_{2j} = 1|u_2)$ is computed cyclically, and this possibility is given as the possibility that the i unit of the visible layer is set to 1.

Step 6. Update parameters

$$E \leftarrow E + \zeta(P(l_1 = 1|u_1)u_1^T - P(l_2 = 1|u_2)u_2^T) \tag{3.8}$$

$$a \leftarrow a + \zeta(u_1 + u_2) \tag{3.9}$$

$$b \leftarrow b + \zeta(P(l_1 = 1|u_1) - P(l_2 = 1|u_2)) \tag{3.10}$$

Step 7. Confirm that the number of iterations has reached the 8th step, not the 2nd step. Step 8: Output parameter

$$e, a, b$$

End.

3.3. Network Tuning. BP neural network is used to achieve one-step training based on each given weight. This process is called optimizing the deep trust net (Figure 3.3).

Set up the BP network at the last layer of the DBN. The feature vector is used as its input for guided learning. Each level of the BP neural network only ensures that the weight of this layer corresponds to the characteristic vector of this layer [16]. BP algorithm adopts a backward neural network to transmit the error message to each RBM layer from top to bottom to adjust the whole DBN. This improves the classification effect of the neural network.

4. Chinese news text classification experiment.

4.1. Introduction to Data Sets. The author obtains information from the Internet and corporate information by sifting financial information on a website. Divide the database into 1000 categories, each representing a business. The performance of the model will be tested with some mainstream classification algorithms.

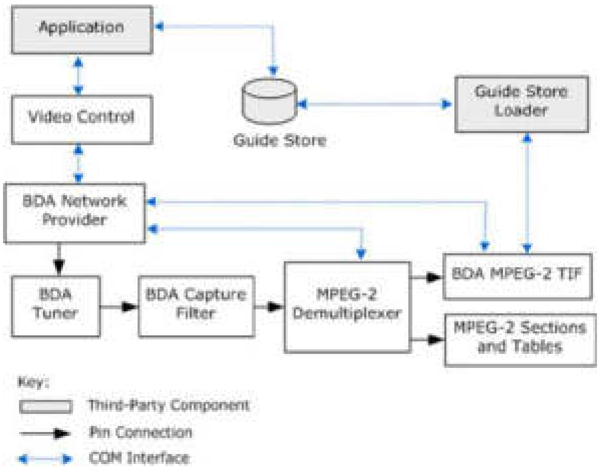


Fig. 3.3: Network tuning.

Table 4.1: Experimental results I.

Classification algorithm	Accuracy rate (%)	Recall rate (%)
HDBN	92.19	92.19
BP	88.54	87.50
SVM	90.63	91.67
ELM	90.10	89.58

Table 4.2: Experimental results II.

Classification algorithm	Accuracy rate (%)	Recall rate (%)
HDBN	90.10	90.10
BP	86.88	84.58
SVM	91.56	92.19
ELM	90.31	90.10

4.2. Test Results. First, the test sample's recognition accuracy should be evaluated. If the results significantly differ from the expected results, returning to the feature screening process and re-screening until the recognition value is in the appropriate range is necessary. The accuracy rate reflects the accuracy of text classification [17]. Only a high accuracy and a low recall rate mean that the label categories that should be predicted are not predicted. In particular, unbalanced samples tend to turn smaller categories into larger ones. Some other multilabel classification methods have problems, such as over-matching between samples. All these problems are worthy of attention. This paper uses HDBN, BP neural network algorithm, support vector machine, ELM and other algorithms to test it. The experimental classification accuracy and recall rate were evaluated (Table 4.1).

THUC News verifies the algorithm. The results are shown in Table 4.2.

5. Conclusion. This paper uses TF-IDF to weigh the text features and obtain the original text feature matrix. The classifier is built and optimized by using the DBN network. Finally, the accurate and fast classification of the text is achieved. Experiments show that the accuracy of using deep neural networks for text classification is significantly higher than BP, SVM, ELM and other classification methods.

REFERENCES

[1] Wu, H., Qin, S., Nie, R., Cao, J., & Gorbachev, S. (2021). Effective collaborative representation learning for multilabel text categorization. IEEE Transactions on Neural Networks and Learning Systems, 33(10), 5200-5214.

[2] Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning–based text classification: a comprehensive review. ACM computing surveys (CSUR), 54(3), 1-40.

[3] Edara, D. C., Vanukuri, L. P., Sistla, V., & Kolli, V. K. K. (2023). Sentiment analysis and text categorization of cancer medical records with LSTM. Journal of Ambient Intelligence and Humanized Computing, 14(5), 5309-5325.

[4] Srilakshmi, V., Anuradha, K., & Shoba Bindu, C. (2021). Incremental text categorization based on hybrid optimization-based deep belief neural network. Journal of High Speed Networks, 27(2), 183-202.

[5] El Rifai, H., Al Qadi, L., & Elnagar, A. (2022). Arabic text classification: the need for multi-labeling systems. Neural Computing and Applications, 34(2), 1135-1159.

[6] Kumar, Y., Koul, A., & Mahajan, S. (2022). A deep learning approaches and fastai text classification to predict 25 medical diseases from medical speech utterances, transcription and intent. Soft computing, 26(17), 8253-8272.

[7] Luo, X. (2021). Efficient English text classification using selected machine learning techniques. Alexandria Engineering Journal, 60(3), 3401-3409.

[8] Moon, S., Kim, M. Y., & Iacobucci, D. (2021). Content analysis of fake consumer reviews by survey-based text categorization. International Journal of Research in Marketing, 38(2), 343-364.

[9] El-Alami, F. Z., El Alaoui, S. O., & Nahnahi, N. E. (2022). Contextual semantic embeddings based on fine-tuned AraBERT model for Arabic text multi-class categorization. Journal of King Saud University-Computer and Information Sciences, 34(10), 8422-8428.

[10] Ibrahim, M. F., Alhakeem, M. A., & Fadhil, N. A. (2021). Evaluation of Naïve Bayes classification in Arabic short text classification. Al-Mustansiriyah J. Sci, 32(4), 42-50.

[11] Wang, Z., Wang, L., Huang, C., Sun, S., & Luo, X. (2023). BERT-based Chinese text classification for emergency management with a novel loss function. Applied Intelligence, 53(9), 10417-10428.

[12] Gurcan, F., & Cagiltay, N. E. (2023). Research trends on distance learning: A text mining-based literature review from 2008 to 2018. Interactive Learning Environments, 31(2), 1007-1028.

[13] Kalra, V., Kashyap, I., & Kaur, H. (2022). Improving document classification using domain-specific vocabulary: hybridization of deep learning approach with TFIDF. International Journal of Information Technology, 14(5), 2451-2457.

[14] Lagrari, F. E., & Elkettani, Y. (2021). Traditional and deep learning approaches for sentiment analysis: A survey. Advances in Science, Technology and Engineering Systems Journal, 6(4), 1-7.

[15] Pintas, J. T., Fernandes, L. A., & Garcia, A. C. B. (2021). Feature selection methods for text classification: a systematic literature review. Artificial Intelligence Review, 54(8), 6149-6200.

[16] Sharma, S., Princy, K. B., & Sharma, R. (2023). A Study on Image Categorization Techniques. International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET), 6(5), 1147-1152.

[17] El-Alami, F. Z., El Alaoui, S. O., & Nahnahi, N. E. (2022). A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model. Journal of King Saud University-Computer and Information Sciences, 34(8), 6048-6056.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Oct 28, 2023
Accepted: Nov 15, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1833–1840, DOI 10.12694/scpe.v25i3.2743

RESEARCH ON HIGH-PERFORMANCE COMPUTING NETWORK SEARCH SYSTEM BASED ON COMPUTER BIG DATA

XIAOGANG CHEN [1*] AND DONGMEI LIU [2†]

Abstract. An efficient computing system for massively parallel systems is established. The stochastic Petri net abstracts and models the high-performance computer work scheduling system. The IB switch is modeled. Stochastic Petri nets are used for performance analysis. Finally, the proposed method is combined with InfiniBand interconnection architecture to evaluate the system's delay. The experimental results prove the feasibility of this algorithm.

Key words: Multi-cluster; Job scheduling; High-performance computing; Task sequencing; Performance evaluation

1. Introduction. The degree of digitalization in various industries is increasing, and the depth and breadth of its applications are gradually increasing. In scheme design, system simulation verification and optimization, digital design and analysis tools are widely used in engineering development. However, majors such as structure, strength and fluid are more likely to use digital simulation methods to solve various technical problems. The increasing size of analytical models, the increasing accuracy of calculations, and the increasing number of multidisciplinary iterations led to an explosive increase in the demand for computing power. High-performance computing systems delivering supercomputing power are already an essential digital foundation. It is already a significant indicator of Chinese overall competitiveness. It can effectively support and drive the research and development of China's primary science and technology projects and thus promote the development of science and technology. Because there is no interconnection among different HPC clusters, many computing tasks are challenging to execute in the clusters. As a result, the system's management complexity and resource efficiency are not well utilized. Using multiple HPC clusters to build a platform with logical consistency and fully use computing resources is a problem that needs to be solved.

Literature [1] reviews the research progress of high-performance computing at home and abroad. The research results of the high-performance computing ecosystem built by the Chinese Academy of Sciences are introduced. This lays a foundation for the research of high-performance computing in China. Literature [2] illustrates the challenges and problems faced in building HPC portals and the technical paths taken. Especially for aviation and other industries, the construction of high-performance computing has essential reference value. [3] Building efficient computing architecture. Literature [4] presents new challenges and development directions for high-performance computing in cloud environments. They research performance evaluation of high-performance computers. At present, the commonly used evaluation techniques include measurement method, reference method, simulation method, model evaluation method and so on. This paper presents a new performance evaluation method. This method has significant application value in performance prediction, capacity planning and hardware and software procurement. This project will start by constructing a random Petri net (GSPN) and conducting fine processing. This results in a higher-level random network. Then, the relevant performance evaluation is carried out.

2. System design ideas. This paper makes a detailed analysis of the distribution of multiple high-performance computer clusters in each laboratory. For example, each cluster uses existing scheduling software and storage systems and adopts a hierarchical scheduling mode [5]. Single-layer scheduling in the same room reconstructs a complete high-performance computing platform. Each cluster location is maintained at the same

[*] 1. College of Computer Engineering, Henan Institute of Economics and Trade, Zhengzhou, Henan 450018, China; Corresponding author's e-mail: cxiaogang@126.com

[†] 2. College of Management, Henan Institute of Economics and Trade, Zhengzhou, Henan 450018, China

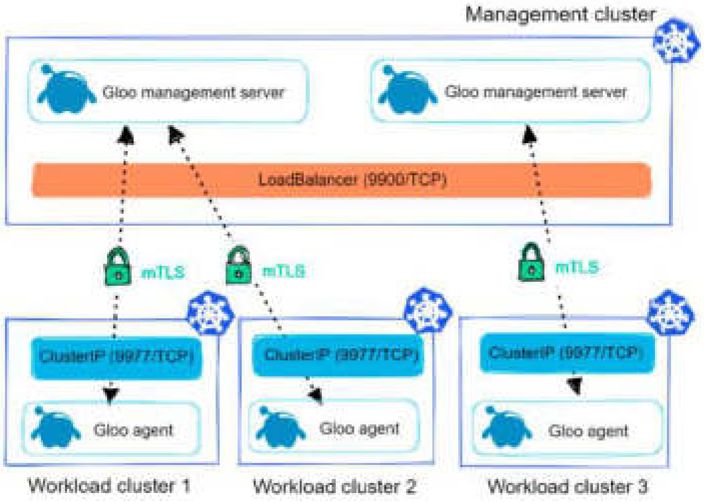


Fig. 3.1: A shared architecture diagram for multiple clusters.

level, considering existing conditions, compute and storage capabilities, and scalability. The following ideas are adopted for construction:

(1) Unified entrance: Users can use efficient computing resources through the unified entrance.

(2) Unified user: log in to the website with a unified ID and perform efficient calculations.

(3) Integrated scheduling: Effective integration and allocation of high-performance computing tasks through a unified scheduling system. In this way, high-performance computing resources can be fully utilized.

(4) Integrated storage: comprehensive integration of scattered storage on each entity. This ensures maximum utilization of storage resources.

(5) Integrated monitoring: Through the comprehensive monitoring and statistics of the operation and utilization of high-performance computing tasks, resource utilization, license, etc., the reasonable allocation of computing resources is realized. This improves the efficiency of operation management.

3. Technical architecture.

3.1. Basic Principles.

(1) High scalability: it can access multiple high-performance computers simultaneously.

(2) High security: The information security in the system can be reliably transmitted and saved after the cluster is networked.

(3) High ease of use: it can quickly and effectively use high-efficiency computing resources.

3.2. System Architecture. Construct an efficient computing system based on a distributed system and integrate and share it. The management center mainly manages user access, user management, unified work arrangement and platform monitoring [6]. Figure 3.1 shows the HPC platform architecture (image referenced in High availability and disaster recovery).

The high-performance computing platform consists of several functional modules:

(1) Access portal: Users and administrators can access, use and manage high-performance computers through this portal.

(2) User management: Authentication of user credentials by integrating with the existing certificate issuance and verification system. Through the integration with the central database to achieve the collection of enterprise-related information [7]. The active table of the College Network Administration Center is used to authenticate operating system users.

(3) Task allocation: Each computing center builds efficient task clusters to complete task allocation. In a distributed environment, the computing tasks of each node are transmitted in real-time [8]. The

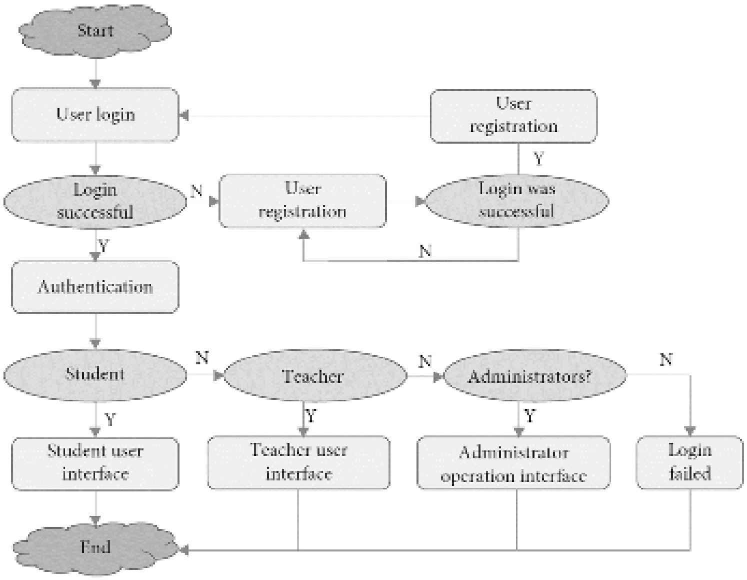


Fig. 3.2: User management diagram.

computing resources of other nodes are used to compute, and the corresponding input and output results are forwarded in real-time.

(4) File system: Use a unified file system to save all kinds of intermediate data and calculation results generated by each operation center during the operation process.

(5) Resource monitoring: various machine-generated information is collected by each computing center. The dispatching center comprehensively processes the operation center, and the overall statistics, analysis and charges are carried out.

3.3. Portal Design. The access portal is configured in the management center of the system to realize the efficient utilization of each system. Access portals are configured in clusters [9]. This can balance the network load and ensure the system's high availability. The implementation of the system includes task management, data management, graphic interaction, compilation and debugging, third-party system integration, web page customization and so on. Through the portal, administrators can manage clusters, tasks, users, permissions, projects, etc. This portal allows Users to submit, monitor, and manage work and data. This system is based on B/S architecture. Access the entry using a browser.

3.4. Multi-User Cluster Management and Scheduling.

3.4.1. User and License Management. The recognition and control of users are realized through system control. The authentication of user credentials is realized by integrating with the existing authentication system [10]. Through the integration with the central database to achieve the collection of enterprise-related information. The paper uses LDAP technology to authenticate computing resources. Figure 3.2 shows a schematic of user management (image cited in Wireless Communications and Mobile Computing, 2022, 2022.).

Through the hierarchical authorization method, the unified management of all kinds of users is realized, while the computer system administrator can only manage the corresponding permissions of users.

3.4.2. Job Scheduling and Software Management. The administrative center is responsible for coordinating and arranging the work. The manager can set an upper limit for CPU time, memory size, runtime, etc., required to perform the task. It can adjust the priority of tasks and perform operations such as pause and resume. You can configure the task schedule according to the following scheduling strategy:

(1) First come, then calculate: the calculation task distribution method is "first come, then calculate." By its

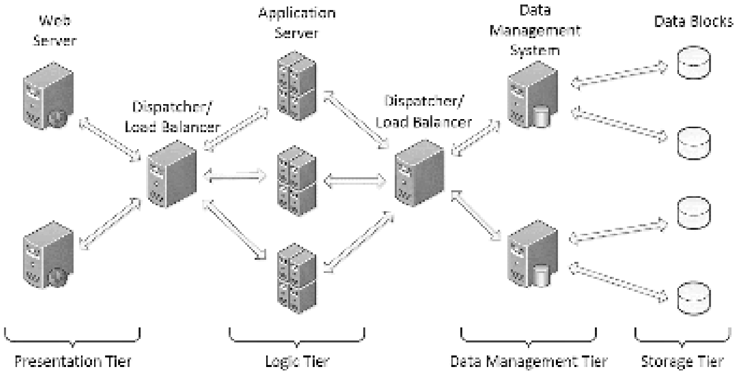


Fig. 3.3: Installation diagram of multi-node simulation computing software.

order in the queue to determine. Users or managers can change the order of issuance by modifying the priority of computation work.

(2) Fairness: Distribute different data to different user groups for different needs. This enables fair access to different types of data streams. The fair sharing mechanism can ensure the fair and reasonable use of the system by distributing the resources allocated to each person or a particular group. If the workload is insufficient, other human computing tasks can use the extra resources for other people's computing [11]. This makes full use of the system. If a user submits more computing tasks, its computing tasks will be completed with higher priority. A method based on fair sharing is proposed. In this way, a reasonable allocation is made to specific users.

(3) Limited time constraint: Resource restriction Scheduling policies can restrict the use of resources. When the number of resources occupied by a computing task exceeds the specified number, it is labeled, or its priority is reduced. The queue parameter is set to limit the resources available for the computation job. Resource constraints determine the number of resources that an arithmetic task can use.

(4) Preemptive scheduling: This method allows high-priority tasks to occupy a smaller space and be executed immediately under tight conditions. When two arithmetic tasks compete for the same arithmetic resources, the arithmetic task in execution is suspended. Currently, the scheduling of work parts is mainly based on the combination of first access calculation and limited resource constraints. By default, the first commit computation task has a higher priority and terminates the configuration of the user's resources if the user reaches a limit. In this way, the dynamic adjustment of the emergency operation task is realized. The simulation analysis software is uniformly installed and configured (Figure 3.3).

Integrate access points and task plans. Use a variety of simulation analysis software to complete user tasks.

3.4.3. multi-computing center planning and monitoring. The schedule for the cross-cluster is shown in Figure 3.4.

(1) Computing network connection: the existing computing network is divided into a management network, computing network and monitoring network. The network management system implements cluster management. The computing network realizes the interconnection of each computing server. The monitoring network can monitor and control the hardware. The network management, computing, and monitoring network are interconnected through networking. The task scheduler is executed on the management server. The task is executed on the computing server. Computing tasks are then assigned to specific computing centers.

(2) File transfer: It is responsible for transmitting working data. And enter the file into the data buffer directory of the working server. Set parameters and compress them before sending. In the data transfer section, the paper added the function of resumable breakpoint. It can prevent the retransmission of big data due to network failure. To ensure that the communication between the client and the client is not interfered with by the outside world, the access of the data between the client and the server must be authenticated. API is used to verify the data in the system and ensure the correctness of the system information. When data is

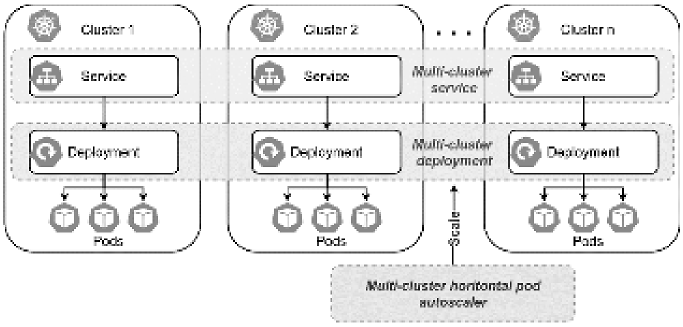


Fig. 3.4: Schematic diagram of cross-cluster scheduling.

transferred, it is transferred in blocks and finally integrated into the overall file. At the same time, it provides the function of a resumable breakpoint [12]. It can effectively prevent data resending caused by network failure. Authorization differential management is carried out on operation data to ensure the security of operation data. Users can only access and manage their operation data but cannot see the operation information of other users.

The cluster monitoring module retrieves each cluster periodically to obtain the related data of each cluster. The specific implementation method of cluster monitoring is:

(1) Real-time fine-cluster load monitoring and analysis: monitor and analyze the CPU core ratio, task ratio, CPU core usage, memory usage, etc., at each stage.
(2) File system load and health monitoring: monitoring file system space utilization, IOPS, etc.
(3) Load analysis of multi-dimensional spatial clusters: The data is studied from multiple levels, such as units.
(4) Real-time monitoring of the license: the usage of each component in the license is monitored.
(5) Statistics and early warning of system records: analysis and early warning of abnormal data.
(6) Email alert: In any abnormal situation, the email is sent to the user's mobile phone to play a role in prompting. When no task is executed, it will be merged into the queue. This allows multi-dimensional monitoring of the queue for computing tasks.

4. Abstract model of the job scheduling system. This paper abstracts it based on the analysis of LSF work plan theory. The corresponding random Petri net model is established [13]. The model contains only one queue g in the LSF where the default is reached. A separate CPU does each task. The workpiece is assigned to a specific arithmetic node (a_i) for the first time upon arrival. A specific $CPU(s_{ij})$ is then assigned to that arithmetic node to complete the job. Repositories and changes in this pattern include the following (Fig.4. 1):

g indicates a work queue for temporary storage tasks that have not been specified. β_i represents the task that has been temporarily assigned to compute node i. g_i represents the task waiting queue for computing node i. A task temporarily stored in arithmetic node i that is not assigned to a processor. v_{ij} represents the work wait queue, the processor$j$ used to compute node i. It is used to store the artifacts that the processor processes. z represents the process in which the task is hosted by the client. a_i indicates that the workpiece is assigned to the arithmetic node i according to a particular scheduling strategy. w_i indicates that the operation node i adds the planned work program to its wait queue. s_{ij}represents a processor j that assigns a job to an arithmetic node i according to a scheduling strategy. r_{ij} represents the working process in the processor.

4.1. Mode refinement and analysis. The original mathematical modeling method is divided into several independent sub-models. Each submodule represents $Y/Y/1$ queue system, in which transition s_{ij}, library v_{ij} and transition r_{ij} respectively represent the workpiece arrival process (arrival rate μ_{ij}), the workpiece waiting queue (queue length d_{ij}) and the workpiece processing process (processing rate η_{ij}) of the queue system of the node i processor j.

4.2. Scheduling method of operation nodes. This project aims at a global minimum average latency. The task is assigned to the nodes with the global minimum mean delay by the comprehensive minimum mean

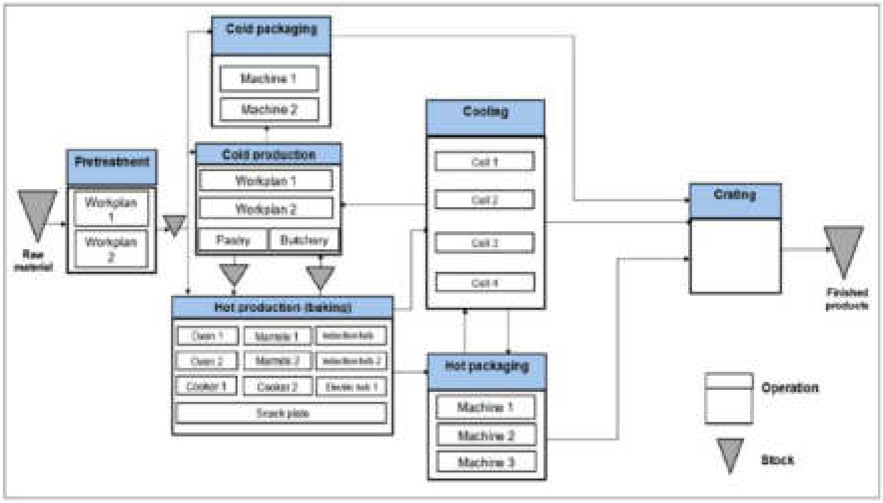


Fig. 4.1: Abstract model of job scheduling.

delay optimization algorithm [14]. The executable predicate y_{a_i} of change a_i is used to restrict change $a_i$and determine whether it can be executed. If the processor queue for the compute node i is not satisfied for the current task to be dispatched, then either the processor i requires the slightest delay or the other compute node's processors are full. Currently, tasks are assigned to the compute node$i$ according to the scheduling algorithm. The conditions that can be implemented can be expressed in the following expressions:

$$\begin{aligned} y_{a_i} : (\sum_{y=1}^{n} Y(v_{iy}) < \sum_{y=1}^{n} d_{iy}) \Lambda((\sum_{y=1}^{n} \frac{Y(v_{iy}}{\eta_{iy}} = \min(\sum_{y=1}^{n} \frac{Y(v_{1y})}{\eta_{1y}}, \cdots, \sum_{y=1}^{n} \frac{Y(v_{my})}{\eta_{my}})) \\ \vee (\forall \zeta \neq i, \sum_{y=1}^{n} Y(v_{\zeta y}) = \sum_{y=1}^{n} d_{\zeta y})) \end{aligned} \tag{4.1}$$

Type a_i random switching u_{a_i} represents the possibility of changing the realization of a_i, that is, the possibility of changing the realization of a_i if it can be realized at multiple operation nodes [15]. The following expression can express this possibility

$$u_{a_i}(Y) = \begin{cases} \frac{1}{|OSEDR(Y)|}, & if\, i \in OSEDR(Y) \\ 0, & otherwise \end{cases} \tag{4.2}$$

Among them,

$$\begin{aligned} OSEDR(Y) = \{\zeta | \sum_{y=1}^{n} \frac{Y(v_{\zeta y})}{\eta_{\zeta y}} = \\ \min(\sum_{y=1}^{n} \frac{Y(v_{1y})}{\eta_{1y}}, \cdots, \sum_{y=1}^{n} \frac{Y(v_{my})}{\eta_{my}} \wedge \sum_{y=1}^{n} Y(v_{\zeta y}) < \sum_{y=1}^{n} d_{\zeta y}\} \end{aligned} \tag{4.3}$$

4.3. Processor scheduling algorithm. This project is based on the principle of minimum delay. The task is assigned to the processor with the most minor delay by sorting the given processor. The executable predicate $y_{s_{ij}}$ of change $s_{ij}$is used to limit the change and determine whether it can be executed [16]. The task assigned to processor j can be assigned to processor j if it has the shortest expected delay among the tasks currently to be assigned. What can be achieved can be expressed in the following expressions:

$$\begin{aligned} &y_{s_{ij}}:(Y(v_{ij})<d_{ij})\wedge((\forall\zeta\neq j,\frac{Y(v_{ij})}{\eta_{ij}}\leq\frac{Y(v_{i\zeta})}{\eta_{i\zeta}}\\ &\vee(\forall\zeta\neq j,Y(v_{i\zeta})=d_{i\zeta})) \end{aligned} \tag{4.4}$$

Random switching $u_{s_{ij}}$ in change $s_{ij}$is the possibility of changing the implementation of s_{ij}, that is, the possibility of changing the implementation of s_{ij} when multiple processors can implement it.

$$u_{s_{ij}}(Y)=\begin{cases}\frac{1}{|SEDR(Y)|}, & if\, j\in SEDR(Y)\\ 0, & otherwise\end{cases}$$

Among them, $SEDR(Y)=\{\zeta|\frac{Y(v_{i\zeta})}{\eta_{i\zeta}}=\min(\frac{Y(v_{i1})}{\eta_{i1}}),\cdots,\frac{Y(v_{in})}{\eta_{id}})\wedge Y(v_{i\zeta})<d_{i\zeta}\}$.

4.4. Perform a scheduling system. Let $U(Y)$ represent the steady-state probability of identifying Y. Suppose that the warehouse v is a queue with capacity$d$, then the average number of tokens in this queue represents the average number of artifacts $S(v)$ in the queue [17]. It can be expressed in terms of $S(v)=\sum_{y=1}^{d}y*U(Y(v)=y)$. The usage degree $A(t)$ of the change t is equal to the sum of the stable probabilities of all the identifiers that make the change executable. If the change is the processor running, then the change in usage is the processor usage. It can be expressed by $A(t)=\sum_{Y\in E}U(Y)$. Here is the entire accessible identification set that can execute t. The productivity of change $T(t)$ is the product of the efficiency with which the change is used and the efficiency with which it is executed, and can be expressed by $T(t)=A(t)*\mu$. Here μ is the execution rate t. According to the queuing theory, the waiting time ST_{ij} is equal to the number of waiting workpieces/represents the number of transferred processes that exit the queue. It can be expressed as $ST_{ij}=S(v_{ij})/T(r_{ij})$, where $S(v_{ij})$ is the average token number of the warehouse v_{ij}. Where $T(r_{ij})$ is the output of r_{ij} in the transfer process. The lag time $ST_{scheduling}$ of the whole plan can be expressed by $ST_{scheduling}=(\sum_{i=1}^{m}\sum_{j=1}^{n}S(v_{ij}))/\sum_{i=1}^{m}\sum_{j=1}^{n}T(r_{ij})$. The paper takes the probability of queuing up to the maximum time as the task loss rate LR. During the queuing process, the workpiece is discarded, and the loss rate LR of the workpiece can be expressed by $LR=U(Y(v)=d)$. Using the above mathematical expression, we can use SPNP to calculate the system's performance.

5. Conclusion.

This paper studies a design scheme of multi-cluster high-performance computing on cluster architecture. Build a high-performance computing platform with unified access and resource sharing to provide users with efficient computing and software support, thereby improving resource utilization efficiency.

REFERENCES

[1] Chen, X., Zhang, J., Lin, B., Chen, Z., Wolter, K., & Min, G. (2021). Energy-efficient offloading for DNN-based intelligent IoT systems in cloud-edge environments. IEEE Transactions on Parallel and Distributed Systems, 33(3), 683-697.

[2] Lv, Z., Lou, R., Li, J., Singh, A. K., & Song, H. (2021). Big data analytics for 6G-enabled massive internet of things. IEEE Internet of Things Journal, 8(7), 5350-5359.

[3] Abualigah, L., Diabat, A., Sumari, P., & Gandomi, A. H. (2021). Applications, deployments, and integration of internet of drones (iod): a review. IEEE Sensors Journal, 21(22), 25532-25546.

[4] Luo, Q., Hu, S., Li, C., Li, G., & Shi, W. (2021). Resource scheduling in edge computing: A survey. IEEE Communications Surveys & Tutorials, 23(4), 2131-2165.

[5] Ghosh, A., Edwards, D. J., & Hosseini, M. R. (2021). Patterns and trends in Internet of Things (IoT) research: future applications in the construction industry. Engineering, Construction and Architectural Management, 28(2), 457-481.

[6] Ding, Y., Jin, M., Li, S., & Feng, D. (2021). Smart logistics based on the internet of things technology: an overview. International Journal of Logistics Research and Applications, 24(4), 323-345.

[7] Yazdeen, A. A., Zeebaree, S. R., Sadeeq, M. M., Kak, S. F., Ahmed, O. M., & Zebari, R. R. (2021). FPGA implementations for data encryption and decryption via concurrent and parallel computation: A review. Qubahan Academic Journal, 1(2), 8-16.

[8] Cao, K., Hu, S., Shi, Y., Colombo, A. W., Karnouskos, S., & Li, X. (2021). A survey on edge and edge-cloud computing assisted cyber-physical systems. IEEE Transactions on Industrial Informatics, 17(11), 7806-7819.

[9] Ouyang, F., Zheng, L., & Jiao, P. (2022). Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020. Education and Information Technologies, 27(6), 7893-7925.

[10] Humayun, M., Jhanjhi, N. Z., Alsayat, A., & Ponnusamy, V. (2021). Internet of things and ransomware: Evolution, mitigation and prevention. Egyptian Informatics Journal, 22(1), 105-117.

[11] Sun, Y., Liu, J., Yu, K., Alazab, M., & Lin, K. (2021). PMRSS: privacy-preserving medical record searching scheme for intelligent diagnosis in IoT healthcare. IEEE Transactions on Industrial Informatics, 18(3), 1981-1990.

[12] Chen, J. I. Z., & Lai, K. L. (2021). Deep convolution neural network model for credit-card fraud detection and alert. Journal of Artificial Intelligence and Capsule Networks, 3(2), 101-112.

[13] 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.

[14] Yang, L., Moubayed, A., & Shami, A. (2021). MTH-IDS: A multitiered hybrid intrusion detection system for internet of vehicles. IEEE Internet of Things Journal, 9(1), 616-632.

[15] Ageed, Z. S., Zeebaree, S. R., Sadeeq, M. M., Kak, S. F., Yahia, H. S., Mahmood, M. R., & Ibrahim, I. M. (2021). Comprehensive survey of big data mining approaches in cloud systems. Qubahan Academic Journal, 1(2), 29-38.

[16] Chen, W., Qiu, X., Cai, T., Dai, H. N., Zheng, Z., & Zhang, Y. (2021). Deep reinforcement learning for Internet of Things: A comprehensive survey. IEEE Communications Surveys & Tutorials, 23(3), 1659-1692.

[17] Sarker, I. H., Khan, A. I., Abushark, Y. B., & Alsolami, F. (2023). Internet of things (iot) security intelligence: a comprehensive overview, machine learning solutions and research directions. Mobile Networks and Applications, 28(1), 296-312.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Oct 28, 2023
Accepted: Nov 15, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1841–1851, DOI 10.12694/scpe.v25i3.2762

DATABASE ACCESS INFORMATION SECURITY MANAGEMENT SIMULATION UNDER BIG DATA PLATFORM

ZHAOCUI LI *AND DAN WANG †

Abstract. When people perform database access information security management, the traditional method cannot accurately verify the identity of the visitor, the credibility of the identity information, and the security management of the access information. With the widespread application of big data technology, the amount of data in databases is rapidly increasing, which brings new challenges to information security management. The main purpose of this study is to explore how to more effectively manage the security of database access information on big data platforms.Therefore, the trusted computing platform is established to implement database access information security management under the data platform. The method determines the user behavior is credible by establishing a behavior chain of behavior based on the user identity and measuring user operation behavior. For the user's private data, the encryption/decryption module is used for security protection, preventing data from being leaked through illegal copying. A trusted metric model based on the USB Key user identity is established and a trusted platform is established. By improving the ELGamal algorithm, the IMC/IMV metrics architecture is utilized to measure platform security attributes. In the first round of anonymous authentication, the identity authentication of the platform is completely completed, and the database access information security management under the big data platform is completed. The simulation results show that in 10 experiments, the transmission time delay of TCP/IP protocol is less than 200ms, and the security of database access information is enhanced after the encryption system is established in the database. This has certain theoretical enlightenment for the improvement of database security and the optimization of information security management.

Key words: Big Data, Database, Information Security, Simulation Research

1. Introduction.

Under the rapid development of computer network technology, various network security incidents continue to occur, seriously affecting the security of user information. Advanced technology has brought convenience to people, and it has also brought various network hazards to them. Code attacks, illegal destruction of systems and data, and illegal information theft are the three most prominent security risks [1]. Traditional network protection methods, such as intrusion detection and virus protection, are implemented in software. Most security methods need to be based on the operating system to operate. Entering the system with a relatively low level of security will not meet the high security requirements of a particular region [2]. Therefore, how to realize security protection from computer architecture and realize the safe and reliable operation of computer system platform has become a core problem to be solved [3].

Trusted computing technology is one of the main technologies to solve computer security problems. By establishing a behavioral trust chain based on user identity, the user behavior is judged to determine whether the user behavior is trustworthy. For the private data of the user, the encryption/decryption module is used for security protection to prevent data from being leaked through illegal copying, and it is proved by experiments. The program can effectively protect the system from the illegal behavior of users and prevent the private data from being illegally stolen [4]. Therefore, to build the security of computer terminals, the basic hardware and software of the terminal must be improved [5]. From the computer terminal core chip, hardware architecture, operating system security protection and other aspects to comprehensively take security measures to ensure that computer terminals are less affected by security issues, which is the basic idea of trusted computing [6]. In summary, trusted computing is used to study the title of database access information security management simulation under the big data platform.Databases have become the main tool for enterprises and organizations to store and process a large amount of information. However, with the expansion of database usage, data

[*] Department of Senior Technician, Shandong Labor Vocational and Technical College, Jinan, 250022, China (Corresponding author, lee009086@126.com)

[†] Department of Senior Technician, Shandong Labor Vocational and Technical College, Jinan, 250022, China (wangdan_sdlvtc@163.com)

security and privacy protection issues are becoming increasingly prominent. Therefore, how to implement security protection on computer architecture to ensure the security and privacy of data in databases has become an urgent problem to be solved. This article will introduce a database access information security management solution based on simulation methods on big data platforms, aiming to improve data security and privacy protection levels. Computer architecture is the organizational structure and behavior of a computer system, and its security protection involves various levels such as hardware, operating systems, and application programs.

This article uses real datasets for simulation experiments. This simulation scheme can effectively evaluate the effectiveness of existing security policies and identify potential security risks. This scheme can also be tested and analyzed for different attack scenarios, providing strong support for the formulation of security strategies. In 10 experiments, the transmission delay of TCP/IP protocol was less than 200ms. After establishing an encryption system in the database, the security of database access information was enhanced. This has certain theoretical implications for improving database security and optimizing information security management.

It is mainly divided into three parts: The first part introduces the Trusted Computing Platform, the Trusted Platform Module and the Trusted Metrics Mechanism, and a trusted platform is built for database access information security management. In the second part, a user behavior measurement method based on trusted platform module is proposed. The trusted metric model is constructed based on the trusted platform module and USBKey two-factor authentication mechanism. In the third part, a trusted security terminal management system is established to verify the proposed trusted model.

2. Background of the Study. With the continuous innovation and development of modern information technology, the most commonly used and widely used in people's life is information. The management of database access information security under the big data platform has attracted the attention of many scholars. Alkida B et al. pointed out that computer operating systems and network platform systems constitute an information database and form a network information platform system [7].

Ramos G et al. mentioned that access control was an important mechanism to ensure that the database system was not invaded and information was protected. And various access control mechanism methods were proposed by them, such as autonomous access control, mandatory access control, etc. [8]. Christodoulou N A et al. proposed autonomous access control method based on the attributes of user access data information. The object access rights are defined by the different attributes of the subject and the subject, and the research proves that the control method is autonomous control [9]. Chuan-Yu L V et al. proposed a database access information security management method under the big data platform. The attribute-based multi-authorization encryption system was constructed to reduce the number of matching operations and improve the efficiency of password utilization. The results show that the encryption of the database enhances the security of database access information and realizes the security management of database access information under the big data platform [10]. Aulkemeier F and others believed that while using computers, a large amount of data information was be generated. At this time, computer database technology should be used to efficiently accomplish the task of information security management. The results show that this can improve the accuracy and reliability of data information transmission [11]. Konstantelos I and others analyzed the security risks and management status of hospital information systems, proposing effective measures to achieve the security of hospital information databases, which is of great significance to ensure the stable, efficient and safe operation of hospital information systems [12]. Zhang F et al. established a real-time monitoring system based on WEB query server, which realizes the real-time monitoring and management functions of online query users, thus improving the security of back-end database information [13]. Example F used the reference table of the network asset refinement table, threat list, and network security threat risk factor matrix for bank security risk, information assets, network security management, and secure time management to analyze the threats and vulnerability of bank through qualitative and quantitative risk analysis, which is of universal significance for the study of bank information security [14]. Hirose K et al. explored its application by analyzing the security access and backup management technology of Oracle database, providing enterprises with more comprehensive and accurate ideas to ensure the system is reliable and secure [15].

According to the above research by China and other foreign scholars, the level of confidentiality of information involved in each level of database access security management is different, and the requirements for users to view content are also different. If confidential information is disclosed, it will have unpredictable consequences. Therefore, ensuring the reliability and completeness of data information in the database is an important topic in the field of information security. The database access information security management under the big data platform can solve the above problems and ensure the information security in the database, which has important practical significance.

3. Application of Trusted Computing in Database Access Information Security Management.

3.1. Trusted platform construction based on database access information security management. The trusted computing platform is computer hardware and software integrated entity constructed by a hardware security chip and its supporting software plus some functional components, and provides trusted computing services externally. TCG believes that starting from an initial ”trust root”, in every state transition of the trusted computing platform, this trust state can remain unchanged through delivery [16,17]. Then the trusted computing environment will not be destroyed, and the trusted state will remain. The trusted operation of the trusted platform will not cause damage to the platform, so the trusted state of the trusted platform will be maintained. This mechanism is called the trust delivery mechanism. The trusted platform for both local users and remote users is always a trusted platform. In order to convince users that the platform is trustworthy, it is often to let users believe that a trusted password security module has been configured in the computing platform, and a series of user-selected security protection software is correctly installed and correctly operated in the system, so that a trust relationship between the user and the computing platform can be established.

The Trusted Computing Platform architecture consists of three parts, TPM, TSS, and user programs. TPM has its core part to serve the upper layer applications. The main process of the metric is as follows: establish metrics and rules, implement metrics on metrics, collect metrics and process them to generate metric sets, and compare the generated metrics set with the expected metrics given by the producer or trusted set [18,19]. Finally, the measurement results are obtained. Trusted measurement technology mainly includes three main parts: trusted metrics, trusted storage and trusted remote reporting. The trusted computer uses the trusted computing metric technology to measure whether a certain program running by the verification system is secure and reliable, and ensures the trusted state of the trusted platform. In the research field of trusted computing, trusted metric technology can be divided into multiple types due to different measurement time or measurement objects. For example, according to different measurement time, it can be divided into static measurement and dynamic measurement [20]. A static metric is a measure that is measured only once when the object is being measured. A dynamic metric is a measure of the behavior of a metric object or the behavior of an object's behavior during object execution. According to different measurement objects, it can be divided into platform-based integrity metrics, platform-based attribute metrics, and semantic-based metrics.Platform integrity metrics, platform based attribute metrics, and semantic based metrics are all used to evaluate the integrity of information or data, but the differences between them are mainly reflected in the application scenarios and methods. Platform integrity measurement is mainly used to evaluate the integrity of a platform. Platform based attribute measurement mainly evaluates the reliability, stability, and security of a platform based on its attributes. Semantic based measurement mainly evaluates the integrity of data based on its semantic content. It focuses on whether the data expresses its semantic meaning truthfully, accurately, and completely, such as whether the text data expresses the correct meaning, and whether the image data is clear and complete.

For data integrity metrics, the TPM provides a hash function and a grouping key for calculating the digest value, while also providing a secure storage unit platform configuration register (PCR, Platform Configuration Register) for storing and updating the metric results. When the platform collects the expected value of the object, the expected value of the collected metric object is stored in the platform configuration register, so that the PCR and the hash function interface provided by the TPM can extend the trust chain established by the trusted root. The hash function provided by the TPM is combined with the platform register PCR to update the PCR value as shown in equation 3.1:

$$PCR[i] = SHA - 1(PCR[i]|new\_Value \tag{3.1}$$

Use a hash function, such as SHA-256 (secure hash algorithm 256 bits), to hash the old PCR value as input. This will generate a new hash value. Add new data to the old PCR values after hashing. This may involve performing bitwise AND operations, bitwise OR operations, addition operations, or other forms of combination between new and old data. Hash the results after adding new data again. This will generate a new PCR value. The old PCR value and the added new data are hashed in the (1) manner to obtain a new PCR value, so that the PCR value generated by the update is related to the old value and the order of the newly added data, and the old value of the SHA-1 algorithm is performed. The sequence of PCR and new data is not interchangeable. That is, the update PCR is not possible if the value of the old value PCR is not based. Integrity metrics and integrity verification are the basic functions of trusted metrics. The metric first measures the operational state of the system in real time and provides a metric reference. A trusted report is a metric that a trusted platform generates to measure different components when creating a trusted environment. It is not only needed to measure its own components, but also to provide external trusted reporting information when it proves that its platform is trusted. Therefore, the trusted report trust root and trusted report are the core of the trusted platform integrity measurement model, and are the necessary conditions for mutual authentication and measurement between the requester and the authenticator.

3.2. Trustworthy Metric Model Based on USBKey User Identity. In a computer system, user behavior is generally defined as the user's access or operation of system resources after authentication, including the behavior of the process as the principal. Therefore, the user identity is determined by combining the USBKey and the TPM for different users, and the different users are measured according to the basis. At present, the security operating system mainly starts from the three aspects of authentication, authorization and auditing, and designs three levels of different users (administrators, ordinary users, audit users), and sets different permissions for level 3 users [21,22]. By binding different security policies through trusted mechanisms, users' access to system resources and trusted use are controlled. The administrator has the highest level of access control permission of the system. The ordinary user can only perform customized operations on the system (such as adding/decrypting based on user roles). The audit user can only view the audit log. The root password is reset to the default password after initialization. Before using the key, the root password needs to be used for authentication, and the authentication is passed to authorize the use of the key. Therefore, in order to prevent malicious access by other programs, it is only necessary to modify the root password to a system-specific password. Therefore, the authentication of the user identity uses the USBKey two-factor authentication mechanism to authenticate the user identity based on whether the USBKey is connected to the operating system, which effectively improves the security of the information stored in the USBKey, and the USBKey user authentication information and TPM certification and authorization are combined to achieve trusted authorization for different users [23,24]. The USBKey is used to store the unique identifier UUID of the key stored in the TPM, and the UUID can be used to use the TPM key. The USBKey is held by the user. Before using the USBKey, the user needs to input the password of the USBKey and verify its integrity. After passing the algorithm, the TPM key is used.

When the user logs in to the system with the correct USBKey, the decryption key is obtained through the TPM. The user database is decrypted by the TPM, and the user identity information is provided for the behavior measurement module. The behavior measurement model measures the user behavior according to the user identity and the specified identity access policy. For the entire measurement system, a key part of the system is dynamically loadable user identity-based transparent encryption/decryption and user behavior metrics.

3.3. Platform Authenticity Verification Based on User Behavior Trusted Metric Design. When the user logs in to the system, after the encryption/decryption attribute is set to the system resource, the system uses the key in the USBKey to dynamically add/decrypt the specified file resource of the system. The results generated by the system trusted process and user system resource configuration are stored in two special files and protected by an access control mechanism. These two special files are the system trusted process list and the user system resource configuration table, which are the basis of user behavior metrics. After the security kernel module is initialized, the system executable program is dynamically loaded, and the trusted process information is measured, the trusted process is loaded, the security configuration policy information of the user is obtained by acquiring the TPM encryption/decryption database, and a resource configuration file is read to establish a System resource controlled information chain. When the user configures the encryption/decryption attribute on the system resource, when the file is read, it first searches for the executable file in the controlled

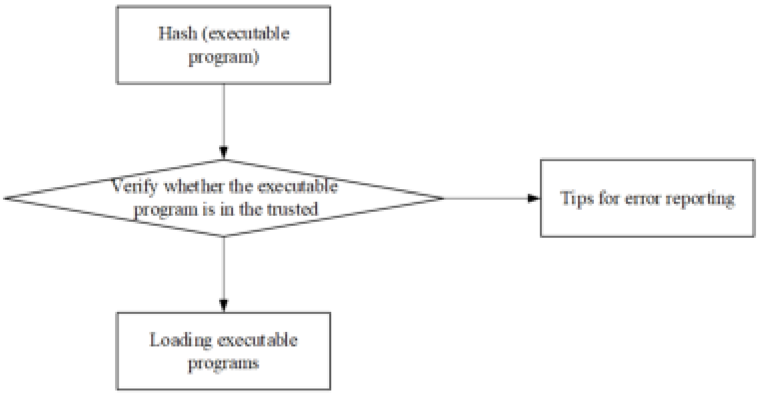

Hash (executable program)
Verify whether the executable program is in the trusted
Tips for error reporting
Loading executable programs

Fig. 3.1: Executable program loading process

file list. If it belongs, the key is obtained by the TPM and the file is decrypted. To protect the security of individual special files, the user identity information is associated with the process information, and the access control policy is performed with the process information. For the system resources, the user can be specified and the specified process can be operated.

When the system logs in, the user's identity role and the information chain based on the user role are established. When a specified user modifies or accesses the specified resource information, it can only be performed by the specified process. The credibility of the user's identity authentication has been described above. The trustworthiness of the process is mainly determined by checking whether the binary code of the executable program is complete, and ensuring that the application executable program is complete and has not been tampered with.

In order to ensure the integrity of the executable program, the "system program whitelist" design is adopted. The system establishes and maintains a trusted list of executable programs. The list stores the path and hash value of the executable program code in the system. When an executable program is started, the TPM hash function is used to calculate the Hash value of the executable program. It is then compared to the stored hash values in the trusted list. If the program path exists in the trusted list and the pre-stored hash value matches the actual metric hash value, the program is allowed to run. If the executable program is not in the list or the pre-stored value does not match the actual value, the operation is prohibited. Since the trusted list is also damaged, in order to ensure that the trusted list is not modified by the illegal program, the access restriction is restricted by the control. Only the specified user can accept the operation request by using the specified program in the specified format. The loading process of its executable program is shown in Figure 3.1.The hash values stored in the trust list are typically used to verify the integrity and trustworthiness of data. By comparing the data with the hash values in the trust list, it can be determined whether the data has been tampered with or damaged. If the data matches the hash value in the trust list, it means that the integrity of the data has been verified, as the hash value is generated by converting the data into a fixed length string. If the data does not match the hash value in the trust list, it indicates that the data may have been tampered with or damaged, and corresponding measures should be taken, such as re obtaining the data or further processing.

In trusted computing, an important aspect of inter-platform security authentication is identity authentication between platforms. Trusted authentication between platforms can only be completed on the basis of proving the identity of the platform. The authentication of the platform identity is based on the identity authentication of the TPM. The verification of the platform authenticity is the verification of the platform identity, mainly to verify whether the platform at both ends of the communication is a real trusted platform. In trusted computing, each TPM has a unique endorsement key (EK), and each endorsement key uniquely represents a TPM. In order to prevent the identity of the endorsement key leakage platform, the TCG specification uses the ATM (attestation identity key) generated by the TPM first to replace the endorsement key to prove that

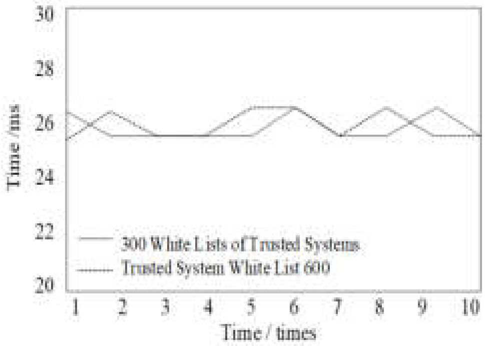


Fig. 4.1: Time comparison of system running program

it has a real trusted platform.

4. Experimental Design and Analysis.

4.1. Experimental environment. In view of the above scheme, the whole system is realized in this paper. The implementation environment is a domestic machine based on the trusted platform module. The experimental environment is configured as: operating system: the winning standard 32-bit operating system. Processor chip: Godson 3A, clocked at 1GHz. Memory and hard disk capacity: 2G memory, 500G hard disk. Develop IDE: QtCreator, NetBeans. Development language: QT, LinuxC. In the experimental environment, the software used for the trusted platform module of Chinese made machines is Kirin V3.0. For the developed system, in order to ensure the availability of the system without affecting the operation and user experience of the operating system, the performance of the developed software system is tested. The test terminal environment is configured as follows: the operating system is a winning 32-bit operating system of Kirin, the processor chip uses Godson 3A, the main frequency is 1 GHz, the memory is 2G, and the hard disk is 500G.

The experimental terminal is configured as follows: The processor chip adopts Godson 3A, the main frequency is 1 GHz, the memory is 2G, and the hard disk 500G compares the user login consumption time, and performs 10 tests. It takes an average of 16 seconds to log in to the system desktop using this system, and it takes an average of 16 seconds to log in with a trusted system without user authentication. Analysis Because the USBKey two-factor authentication replaces the original system user password authentication during the system login process, although the USBKey communication and the TPM communication consume time compared to the original system authentication, the time consumption can be neglected.

4.2. Model Performance Analysis. This method is implemented in the system of winning the standard Qilin + Godson 3A. The system performance is mainly analyzed from the following two aspects. The first is the user identity authentication measurement time, the second is the system whitelist query and measurement time, and the third is the file encryption/decryption time. For the system whitelist metrics, the following tests were made respectively, when the system whitelist was 300 and when the system whitelist was 600. A 1M size executable program is queried and the digest value is calculated and tested 10 times. The required time comparison is shown in Figure 4.1.

According to the experimental data, when the system trusted whitelist is basically the same in the query and measurement time required for 300 and 500. Therefore, it can be known that the query time in the whitelist is relatively short, mainly because the executable program summary is worth calculating. System whitelist queries and metrics have little impact on user experience and system performance at the ms level, and are within acceptable limits. For the time consumption of file transparent encryption/decryption, 10 groups of experiments are designed to consume time for reading and writing operations on 1M files and 2M files respectively. The time required to read and write 1M files and 2M files is shown in Figure 4.2.

The time to write the 1M file and the 2M file separately is as shown in Fig. 4.3.

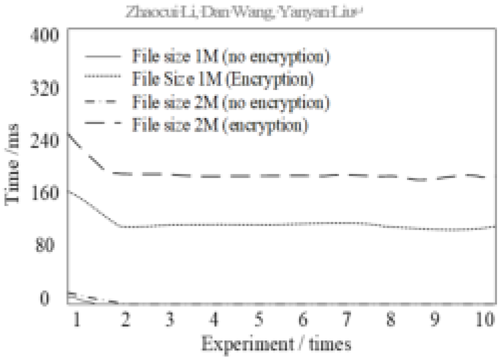


Fig. 4.2: Comparisons of the time required to read files

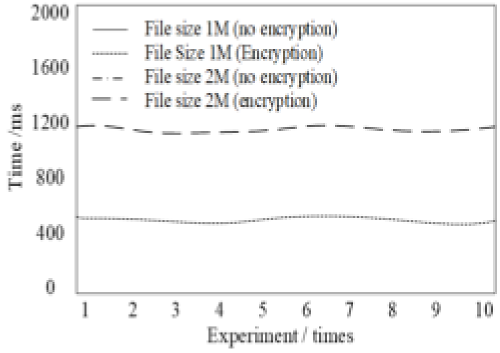


Fig. 4.3: Time required for file writing

Combined with the data of the file reading and writing experiment, it can be known that during the read operation, due to the caching mechanism of the system, the first reading time is greater than the time required to read the file later, the reading and writing time is compared, and the writing operation time is significantly larger than the reading operation time. This is because file writes require lock protection to prevent data consistency from being compromised. Comprehensive experimental data when the transparent addition and decryption functions are added, the file read and write operations take time to increase, but the increase time is still within the acceptable range of the user.

4.3. Platform performance test and result analysis. Three calculator programs are added to the dynamic measurement system for testing. Compared with the dynamic measurement system metric program, the CPU resource consumption and memory resource consumption are shown in Figures 4.4 and 4.5.

From the comparison experiments of Figs. 4.4 and 4.5, the CPU consumption rate of the enabled measurement system is 15% higher than that of the non-enabled measurement system. In terms of memory space usage, enabling metrics is about 8% more than enabling metrics. Since system consumption is related to the metrics process, it can be seen that there are certain effects on system performance when measuring three calculator processes. But, in order to ensure high reliability, the consumption of system resources is acceptable. The test terminal environment configuration of this paper is as follows: The processor chip adopts Godson 3A, the main frequency is 1GHz, the memory is 2G, and the hard disk is 500G. In order to test the performance of the ETAAP protocol, an experimental comparison is made with the protocol (TCP/IP) not used. The experiment is carried out 10 times. The 200KB data test is transmitted under two trusted computing platforms to check the comparison between the system CPU usage and the system memory usage during the transmission process.

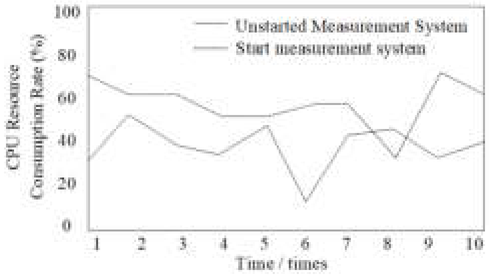


Fig. 4.4: CPU resource consumption ratio comparison

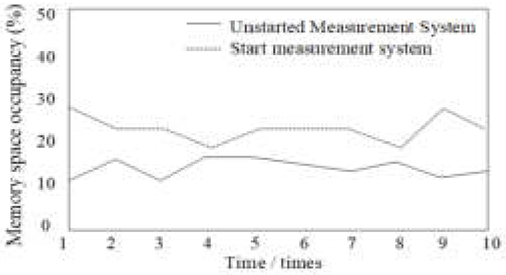


Fig. 4.5: Comparisons of memory occupancy

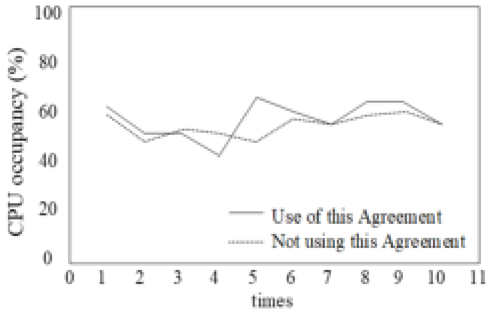


Fig. 4.6: CPU resource consumption

The experimental results are shown in Figure 4.6 and Figure 4.7.

It can be seen from the comparison between Fig. 4.6 and Fig. 4.7 that the CPU usage and memory usage of this protocol are basically different from those of the unused protocol. In response to the data delay caused by the adoption of this protocol, 200KB and 400KB files are tested and sent in the experimental environment, and 10 experiments are performed. The experimental results are shown in Figure 4.8 and Figure 4.9.

From the comparison implementation of Figs. 4.8 and 4.9, it can be seen that when transmitting the same size data file, the transmission time delay of adopting the end-to-end trusted anonymous authentication protocol and directly adopting the TCP/IP protocol is within 200ms. The delay of time mainly occurs in the verification phase of the identity authentication and extended system security attributes, and will not affect the system performance in the future data transmission. From the analysis of the experimental results, it can be seen that the impact of this protocol on the performance of the system is acceptable.In today's information age, computer terminals play an important role in various industries and fields. However, with the improvement

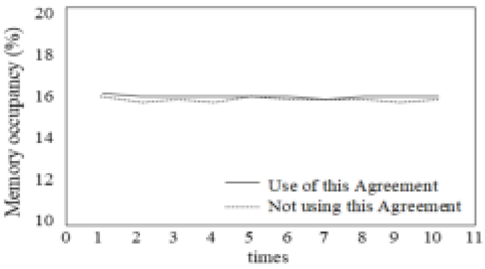


Fig. 4.7: Memory resource consumption

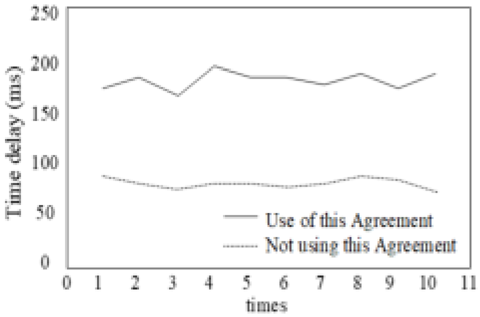


Fig. 4.8: Time Delay Contrast Diagram for Transferring 200KB Files

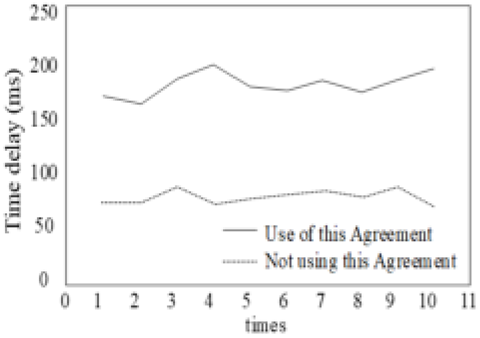


Fig. 4.9: Time Delay Contrast Diagram for Transferring 400KB Files

of its popularity, the security issues of computer terminals are becoming increasingly prominent. For critical or sensitive computer terminals, hardware level security measures such as security chips and trusted execution environments can be adopted to prevent physical attacks and illegal access. Timely update and upgrade the hardware components of computer terminals, such as CPU, memory, hard disk, etc., to improve the system's processing power and operational efficiency, while also reducing security vulnerabilities. For computer terminals involving sensitive information, measures such as encrypted storage and encrypted transmission can be adopted to ensure data security.

5. Conclusions. The research background and theoretical knowledge of trusted computing are introduced, and the platform structure, basic composition and functional mechanism of trusted computing are analyzed.

The basic components include a trusted platform module and a software protocol stack, which introduces the password support platform for trusted computing and the security function and password mechanism for password support. Based on the service provided by the trusted computing platform, the internal user behavior measurement model of the main text platform is studied. A USB Key user identity measurement model is proposed. In this model, based on the mutual authentication of USBKey and Trusted Platform Module, user identity based authentication and authorization are implemented. User identity based behavioral trust chain and data security encryption/decryption functions are established. Experimental analysis shows that the proposed method implements user behavior metrics and data security guarantee based on user identity grading. For the problem of security authentication between platforms, based on trusted computing, the zero-knowledge authentication protocol is used to complete the mutual authentication of the inter-platform identity, which further improves the security performance of the platform. Under this premise, mutual authentication is performed between the integrity of the platform and the security attributes of the platform to ensure that the security of the platform conforms to the access policy while the identity of the platform is correct. However, there are still deficiencies. In future research, it is necessary to further improve the applicability of the system and develop a security system in different environments.The simulation of database access information security management under big data platforms relies on a large amount of data for simulation and training. However, these data may have quality issues, such as incomplete or inconsistent data, which can have a negative impact on simulation results. Therefore, it is necessary to strengthen the evaluation and cleaning of data quality to ensure the accuracy of simulation. In the future, it is necessary to use artificial intelligence technology to intelligently upgrade simulation systems, such as adaptive optimization and automatic decision-making, in order to improve the efficiency and accuracy of simulation.

REFERENCES

[1] Chehri, A., Fofana, I., Yang, X., *Security risk modeling in smart grid critical infrastructures in the era of big data and artificial intelligence. Sustainability, 13(6): 3196, 2021.*

[2] Awaysheh, F. M., Aladwan, M. N., Alazab, M., *Security by design for big data frameworks over cloud computing. IEEE Transactions on Engineering Management, 69(6):3676-3693, 2021.*

[3] Smys, D. S., Wang, D. H., Basar, D. A., *5G network simulation in smart cities using neural network algorithm. Journal of Artificial Intelligence and capsule networks, 3(1): 43-52, 2021.*

[4] Gong, Y., Liao, J., *Blockchain technology and simulation case analysis to construct a big data platform for urban intelligent transportation. Journal of Highway and Transportation Research and Development (English Edition), 13(4): 77-87, 2019.*

[5] Zhou, Z., Wang, M., Huang, J., *Blockchain in big data security for intelligent transportation with 6G. IEEE Transactions on Intelligent Transportation Systems, 23(7): 9736-9746, 2021.*

[6] Gladun, A. Y., Khala, K. A., *Ontology-based semantic similarity to metadata analysis in the information security domain. Problems in Programming, (2): 34-41, 2021.*

[7] Garg, S., Singh, A., Kaur, K., et al. *Edge computing-based security framework for big data analytics in VANETs. IEEE Network, 33(2): 72-81, 2019.*

[8] Wandji, P. Y. B., Charrier, C., Di, M. J.,*Deep features fusion for user authentication based on human activity. IET Biometrics, 12(4): 222-234, 2023.*

[9] Liu, X., Ding, N., Shi, J., *An Identity Recognition Model Based on RF-RFE: Utilizing Eye-Movement Data. Behavioral Sciences, 13(8): 620, 2023.*

[10] Stergiadis, C., Kostaridou, V. D., Veloudis, S., *A Personalized User Authentication System Based on EEG Signals. Sensors, 22(18): 6929, 2022.*

[11] Son. S., Park, Y., Park. Y., *A secure, lightweight, and anonymous user authentication protocol for IoT environments. Sustainability, 13(16): 9241, 2021.*

[12] Konstantelos, I., Jamgotchian, G., Tindemans, S., *Implementation of a Massively Parallel Dynamic Security Assessment Platform for Large-Scale Grids. IEEE Transactions on Smart Grid, PP(99): 1-1, 2016.*

[13] Sun, J., Khan, F., Li, J., *Mutual authentication scheme for the device-to-server communication in the Internet of medical things. IEEE Internet of Things Journal, 8(21): 15663-15671, 2021.*

[14] Janjanam, L., Saha, S. K., Kar, R., *Optimal Design of Hammerstein Cubic Spline Filter for Nonlinear System Modeling Based on Snake Optimizer. IEEE Transactions on Industrial Electronics, 70(8): 8457-8467, 2022..*

[15] Hirose, K., Kim, S., Kano, Y., et al. *Full information maximum likelihood estimation in factor analysis with a large number of missing values. Journal of Statistical Computation & Simulation, 86(1): 91-104, 2016.*

[16] Liang, J., Zhang, M., Leung, V. C. M., *A reliable trust computing mechanism based on multisource feedback and fog computing in social sensor cloud. IEEE Internet of Things Journal, 7(6): 5481-5490, 2020.*

[17] Teisserenc, B., Sepasgozar, S., *Adoption of blockchain technology through digital twins in the construction industry 4.0: A PESTELS approach. Buildings, 11(12): 670, 2021.*

[18] Kiu, M. S., Chia, F. C., Wong, P. F., *Exploring the potentials of blockchain application in construction industry: a systematic review. International journal of construction management, 22(15): 2931-2940, 2022.*

[19] Sukhija, N., Bautista, E., *Towards a framework for monitoring and analyzing high performance computing environments using kubernetes and prometheus, 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). IEEE, 257-262, 2019.*

[20] Liu, Z., Chi, Z., Osmani, M., et al. *Blockchain and building information management (BIM) for sustainable building development within the context of smart cities. Sustainability, 13(4): 2090, 2021.*

[21] Rahman, A., Nasir, M. K., Rahman, Z., *Distblockbuilding: A distributed blockchain-based sdn-iot network for smart building management. IEEE Access, 8: 140008-140018, 2020.*

[22] Li, W., Wu, J., Cao, J., *Blockchain-based trust management in cloud computing systems: a taxonomy, review and future directions. Journal of Cloud Computing, 10(1): 1-34, 2021.*

[23] Zhong, B., Wu, H., Ding, L., et al. *Hyperledger fabric-based consortium blockchain for construction quality information management. Frontiers of engineering management, 7(4): 512-527, 2020.*

[24] Thapa, C., Camtepe, S., *Precision health data: Requirements, challenges and existing techniques for data security and privacy. Computers in biology and medicine, 129: 104130, 2021.*

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Oct 30, 2023
Accepted: Nov 24, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1852–1862, DOI 10.12694/scpe.v25i3.2764

HOSPITAL MEDICAL BEHAVIOUR SUPERVISION AND OPERATIONAL EFFICIENCY EVALUATION METHOD BASED BASED ON BIG DATA PLATFORM

YI LIU*AND YI ZHANG†

Abstract. The application of cloud computing and big data core technologies and concepts to medical informatization can improve its flexibility and efficiency, and achieve the overall deployment and intensive management of the system. In the era of big data, the use of information management platform can optimize and standardize clinical diagnosis and treatment process, improve the quality and efficiency of diagnosis and treatment services, and improve the quality and level of scientific research, which meets the requirements of medical reform for fine management of hospitals and precise medical treatment in the era of evidence-based medicine. This paper introduces the development and application of big data analysis in the field of information technology. Through the research on the supervision of hospital medical service behavior, the top-level design is used to standardize the content of medical behavior supervision, and the effectiveness of supervision is discussed to achieve the purpose of reducing clinical paperwork. Based on the needs, the medical service behavior supervision system is proposed and constructed. Strengthen hospital medical behavior supervision through hospital big data analysis and knowledge base system support.This system can provide management with an effective monitoring and management tool, enabling them to promptly identify and solve problems that arise in medical services. By analyzing a large amount of medical data, hospitals can help predict the trend of disease occurrence in advance, thereby taking preventive and control measures in advance, and reducing the occurrence and spread of diseases. Through deep learning and analysis of patient data, personalized treatment plans can be provided for each patient to improve treatment effectiveness.

Key words: Big Data; Medical Behavior; Efficiency Evaluation

1. Introduction. In recent years, with the advancement of new medical reform, medical informatization, one of the "four pillars" and "eight pillars", is playing an increasingly important supporting role and has become the key development direction of medical system reform [1]. While medical informatization has received unprecedented attention and development, due to the arrival of the era of big data and the relatively backward development of traditional medical information technology, medical informatization also has many problems [2]. Big data and other modern computer technologies are widely applied in the construction of hospital informatization, which is the only way to change the current construction and application status of "chimney" hospital information system with high investment, low efficiency and difficult management and to establish a new type of digital hospital information platform architecture system [3]. The application of big data analysis technology in the field of medical and health care, the use of data mining and analysis technology to analyze medical data, and the combination of traditional medical data, can achieve accurate and personalized health care services [4]. Making full use of large medical data can reduce the infection rate of infectious diseases and improve public health monitoring. In terms of management, large medical data can be used for disease classification, resource management, quality control and other operations [5]. In a word, making full use of big medical data is an important way to promote medical informatization and improve the efficiency and quality of medical industry. With the gradual advancement of medical and health reform in recent years, the medical environment has been greatly improved, and more people have access to quality medical services. However, for medical institutions, due to the introduction of relevant policies and systems in the process of medical reform, hospitals have policies. The implementation and implementation basically rely on traditional meetings, training and missions, and post-intervention. It is difficult to know the implementation effect of the policy in the first time [6].The new regulatory policy on medical service behavior may have an impact on the behavior of medical

*Department of Medical Technology, Chongqing Medical and Pharmaceutical College, Chongqing, 401331, China (18996139677@163.com)

†Department of Medical Technology, Chongqing Medical and Pharmaceutical College, Chongqing, 401331, China (Corresponding author, 10758@cqmpc.edu.cn)

personnel. For example, stricter regulatory policies may limit doctors' behavior, which may lead to doctors being more conservative when dealing with patients, thereby affecting their treatment effectiveness. In addition, if regulatory policies lack fairness and transparency, it may lead to dissatisfaction and resistance from doctors, thereby affecting the quality of medical services [7]. Pareto efficiency describes the effectiveness of resource allocation in a perfectly competitive market, which is an enjoyable way of resource allocation. However, it has nothing to do with fairness. Even if all medical resources are monopolized, they can still be pareto efficient [8]. Medical and health resources are limited, non-competitive and non-exclusive, and medical service is a quasi-public product. Therefore, the principle of profit maximization and cost minimization cannot be used by hospitals to participate in market competition. Therefore, the evaluation of hospital efficiency cannot be simply measured by pareto efficiency [9]. Information-based big data analysis and medical service behavior supervision are complementary to each other. In medical behavior supervision, the process supervision of implementing medical technical norms is the basis and also the means [10]. From the perspective of hospital administrators, big data analyzes all the violations and problems of the whole hospital to analyze big data, provide a basis for managers' decision-making, monitor the work of all levels of the hospital, decompose the data at different levels, and implement the indicators to specific The executive department or medical staff achieves the purpose of multi-dimensional regulation [13].

Medical behavior supervision should integrate scattered distribution with business data of various systems, integrate and analyze these business data, and realize self-configuration of data acquisition according to actual business, interface with various applications or application systems, and subject-oriented analysis of historical data [14]. When we view the external environment of a hospital as a determined environment, such as a stable political situation, good policies, universally followed laws, and appropriate public opinion guidance. Then change the internal factors of the hospital, such as increasing or reducing beds, increasing or decreasing the number of medical staff [15]. Similarly, assuming that the internal factors of a hospital are fixed within a certain period of time, and the external factors change, such as the adjustment of medical insurance policy or the national policy adjustment, which makes foreign capital and private capital enter the medical market in a large scale, whether endogenous or exogenous factors change, can affect the operational efficiency of a hospital [16].

In this paper, we propose an algorithm based on the big data platform, which is a new algorithm for the supervision of medical behavior and the evaluation of operational efficiency in hospitals. In summary, our contributions are as follow:

1. This algorithm is a new technology based on big data platform for the problem of hospital medical behavior supervision and operational efficiency evaluation methods.
2. This technology is widely applicable in the big data platform environment, and it has high applicability for most of the solid hospital medical behavior supervision and operational efficiency evaluation methods.
3. Higher precision, wider applicability and higher recognition.

2. Related Work.. As medicine enters the era of big data, the mining and utilization of clinical data will inevitably improve clinical decision-making and management levels, improve service efficiency, reduce medical errors, and deepen medical reform. This also puts forward higher requirements for hospital refined management [17]. IBM has creatively proposed the concepts of earth intelligence, intelligence, and medical technology as one of its important aspects of rapid development worldwide. Currently, the academic community has reached the following consensus in several stages of implementing intelligent medicine: 1. Medical data collection and analysis; 2. Big data information analysis and processing; 3. Intelligent medical knowledge learning [18]. Based on medical diagnosis and treatment support based on big data processing, some scholars believe that introducing key technologies and core concepts of cloud computing and big data into medical informatization construction is in line with the inevitable trend of modern medical development. In the era of big data, cloud computing and big data technology have become key technological support for achieving the transformation of medical informatization and promoting medical system reform [19].

Cloud computing, as a system engineering in the era of big data, has unparalleled advantages in the research and application of big data. Some scholars have introduced hospital informatization construction into cloud computing. Through overall deployment and intensive management, equipment investment has been reduced, resource utilization has been improved, system operating costs have been reduced, green hospital IT system architecture has been achieved, and a new type of hospital information construction has been established. In the process of big data processing in hospitals, it is necessary to consider the security and sharing of data. Develop corresponding data security policies to ensure data privacy and security. Meanwhile, for valuable analysis results, data sharing and exchange can be achieved through shared platforms, promoting the development of medical research [20]. Some scholars believe that big data has potential in the medical field: around how to reshape the medical system, based on the belief that medical big data contains huge wealth, medical big data has many applications, and for hospital informatization construction, meaningful information extracted from medical big data has also been greatly promoted [21]. Some scholars believe that traditional computer architectures have limited processing power for big data based on the symbiotic impact evaluation method of topological reduction of big datasets. Cloud computing provides an effective way for big data processing [22].

3. Materials and Methods. With the rapid development of information technology, especially the rise of cloud computing and big data technology in recent years, more and more medical institutions in China have begun to accelerate the change of the traditional mode of hospital informatization, to realize the transformation and upgrading of digital hospital construction, in order to improve their own medical service level and core competitiveness. Based on cloud computing medical cloud service mode, a new architecture of digital hospital information platform is established, which enables medical institutions to improve "patient-centered" clinical medical services and meet the needs of medical resources [23]. Is not a simple computerized medical informatization, but information sharing as the core, including internal medical institutions, medical institutions, medical institutions and community, the health administrative department, medical insurance agency information sharing between each other, maximum convenient patient medical treatment process, improve the efficiency of medical work, and the convenience of various kinds of management personnel management work of analysis and decision, give full play to the information technology application in the medical industry value, improve the effective utilization ratio of health resources. Will be the key technology and the concept of cloud computing and big data used in hospital information construction, can improve the flexibility and efficiency of medical informatization, to realize the overall deployment, intensive management system, on-demand configuration, satisfy its large-scale application in scalability, reliability and on-demand services, at the same time can make the elasticity of large data storage, centralized management and effective utilization.

The evaluation of hospital efficiency should also be discussed separately from the short-term and long-term perspectives; the short-term external environment is relatively stable, and the internal factors directly affect the hospital's economic efficiency; using a single-time panel data to measure the relative relationship between different hospitals. In the long run, because internal factors and external factors are simultaneously changing, such as the introduction of a policy, it may be good news for a certain type of hospital, and for a hospital that does not have this attribute, it may be bad news. Therefore, the long-term efficiency of the hospital cannot be measured by the accumulation of panel data. So how do you judge whether the hospital has maintained high efficiency for a long period of time, 10 or 20 years, or in which years is more efficient, and in other years is less efficient? When evaluating the technical efficiency of different hospitals, especially when the technical efficiency is related to many different factors, the multiple linear regression model is a better solution. Multivariate linear regression analysis solves the correlation between a phenomenon (interpreted variables) and multiple influencing factors (interpreted variables) in market competition activities, and can more intuitively describe what are the most significant factors affecting a phenomenon and the degree of such influence.

Firstly, the large data collection module is used to collect patient's visiting information from various hospital information systems to form user preference data, and these historical records are transformed into a simple triple:

$$dF_r = \tau b dx \tag{3.1}$$

Then use several similarity measures to calculate the similarity between users, the formula is:

$$x_{ij}^{'} = \frac{x_{ij}}{S_J} \tag{3.2}$$

In essence, the Euclidean distance represents the true distance of two points in a multidimensional space,

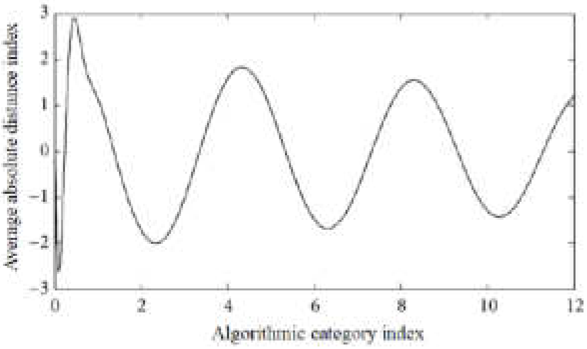


Fig. 3.1: Average absolute distance

and its formula is as follows:

$$\rho(x,y)=\begin{cases} k/(M_1 * M_2)(x,y)\in S \\ 0(x,y)\notin S, \end{cases} \tag{3.3}$$

The similarity expressed by Euclidean distance is formula 6. It can be shown in Figure 3.1.

$$F(x)=\frac{1}{1+e^{-ax}} \tag{3.4}$$

Pearson correlation coefficient represents the ratio between skew variance and standard deviation of two triples, and its calculation formula is as follows:

$$q_f=-\frac{1}{1+e^{-ax}} \tag{3.5}$$

Cosine similarity represents the cosine value of the Angle between two images of two triples in the vector space, which is used to measure the difference between them. The calculation formula is as follows:

$$S_1=R_1=[G^1,G^2,...,G^k] \tag{3.6}$$

According to the similarity measure calculated by the above similarity calculation method, two types of methods are used in our system to obtain adjacent users or items, that is, neighbors based on similarity threshold and a fixed number of neighbors, as shown in Figure 3.2 and Figure 3.3.

Accuracy and recall are two measures used to evaluate the efficiency of behavior supervision and operation. Firstly, relevant data needs to be collected, including behavioral data of medical personnel, patient feedback, and reports of medical accidents. These data can be obtained through the hospital's information system, survey questionnaires, surveillance cameras, and other means. Classify the behavior of medical personnel based on the collected data. For example, behaviors can be divided into "compliant" and "non compliant", or classified based on specific types of behaviors. Use machine learning or statistical methods to establish models based on collected data to predict the behavior of healthcare workers. Formulas such as Formula 3.7 and Formula 3.8 can be used to establish the model. Use a test set to evaluate the accuracy and recall of the model. Accuracy refers to the proportion of the correct number of samples predicted by the model to the total number of samples, while recall refers to the proportion of the correct number of positive samples predicted by the model to the total number of positive samples. Apply the model to actual data to monitor the behavior of medical personnel. If the accuracy and recall of the model are high, the model can effectively identify and predict the behavior of medical personnel, thereby assisting hospitals in behavioral supervision. By collecting and analyzing this data, comprehensive supervision of doctors' medical behavior can be carried out, including their prescription behavior,

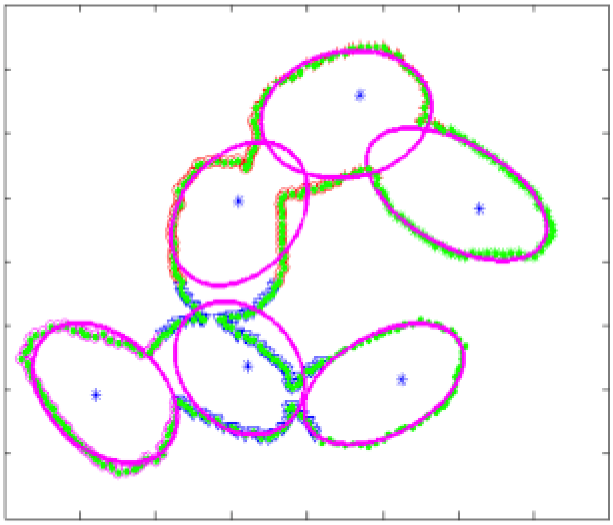


Fig. 3.2: The neighboring method based on the threshold of similarity threshold

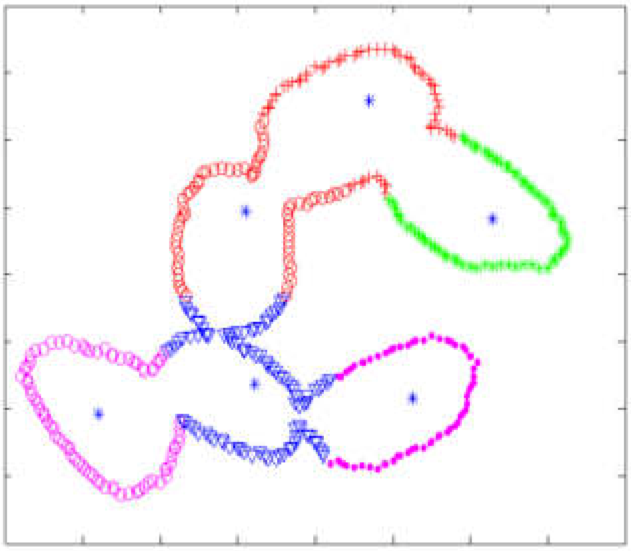


Fig. 3.3: Fixed number of neighboring methods

diagnostic behavior, and treatment behavior. These data can help us understand whether doctors' behavior complies with regulations and whether they can provide high-quality medical services. The measurement methods defined by them are Formula 7 and Form 8. The accuracy recall index is shown in Table 3.1. The types of supervision are shown in Table 3.2 and the growth of operational efficiency is shown in Table 3.3.

$$a_i = (\tau_i - \tau_{i-1})/(\rho_i h_i) \tag{3.7}$$

$$v_D = \frac{d_w}{d_t} \tag{3.8}$$

Table 3.1: Accuracy recall rate

Actual Class		Positive	Negative	Total
Actual Class	Positive	Ture Positive	Ture Negative	TP+FN
Actual Class	Negative	Fasle Positive	Fasle Negative	FP+TN
Actual Class	TP+FP	TN+FN	TP+FN+FP+TN	

Table 3.2: Accuracy recall rate

	No profit	Profit seeking
Non regulation	0 r	-s,r
Supervise	-c,r	f-c-s,r-f

In large-scale data processing, these two separate components are often used together in the following format:

$$\Delta y = M(t_0 + \Delta t) - M(t_0) \tag{3.9}$$

Multivariate linear regression analysis is used to evaluate the technical efficiency of different hospitals, which can solve the problem of hospital public welfare. The methods and results are in the following equations:

$$R_i(i) = P(q_t = s_i | y_t) \tag{3.10}$$

$$\lambda = f(x_1, x_2, x_3, x_4) \tag{3.11}$$

In conclusion, the performance formula of medical behavior supervision and operational efficiency evaluation methods in hospitals under the big data platform is as follows:

$$\begin{array}{l} \qquad\qquad min\theta_{j_0} \\ s.t. \begin{cases} \sum_{J=1}^{n} X_j\lambda_j + S^+ = \theta_{j_0} X_{j_0} \\ \sum_{J=1}^{n} Y_j\lambda_j - S^- = Y_{j_0} \\ \sum_{J=1}^{n} \lambda_j = 1 \\ \lambda_j \geq 0, j = 1, ..., n; \theta_{J_0} \in E, \end{cases} \end{array} \tag{3.12}$$

The generation of large data has gone through several stages step by step, from the initial operation of special application groups to the active generation of data by the whole people, that is, users, to the final generation of data automatically and endlessly by human intelligent sensor devices, which together constitute large data. From these data generation methods, W can clearly understand the characteristics of large data introduced in the previous section. Data generation speed is faster and faster, data volume is larger and larger, data types are diverse, and data value density is lower and lower. With the popularization of hospital information system, a large number of medical information will be generated every day, including image information (CT, MR), vital signs, clinical examination, diagnostic information and other information. These resources are valuable resources of the hospital, and have important value in patient diagnosis, clinical comprehensive display and medical research. The effective acquisition and storage of this information is the basis for the use of medical data. Platform-as-a-service provides an operational environment for the creation of application services, integrated development tools and software for medical software vendors, medical institutions at all levels, and medical management departments to support the subsequent development of medical organizations in distributed computing platforms. . By using the various interface calls provided by the platform, software tools, integrated SDKs, data mining engines and other basic services, you can quickly develop a variety of software to meet your own needs and seamlessly integrate into existing platforms.

Table 3.3: Operating efficiency growth

Time	Growth rate
1	7.68%
2	10.37%
3	13.29%

Table 4.1: Economic benefits

Number of users	106	97
Scale	Big	Small
Return on investment	76.7%	45.3%

4. Results.. The purpose is to provide medical information system service by desktop service, so that hospitals and users can obtain the permission of medical information system by "leasing" according to their own needs, and realize the mode of medical information system as a service. At the same time, for providers of medical cloud services, virtual desktop architecture allows them to quickly and cheaply build a mobile hospital information environment without modifying existing medical information systems or developing new mobile medical applications to support large-scale access to medical information systems based on cloud computing. The larger the number of users, the larger the scale of virtual desktop architecture implementation and the greater the economic benefits, as shown in Table 4. The growth rate of hospital assets is shown in Table 5. The providers and owners of medical big data service platform can be hospitals or third-party institutions independent of hospitals. Compared with the third-party medical cloud service platform, the hospital has the inherent advantages of adapting measures to local conditions. First, the hospital has a complete medical information system-level environment, which can be quickly upgraded to the cloud platform. Second, IT can be closely integrated with the hospital process rather than provide a single service, such as the storage of medical images, as a third-party service provider. Third, IT personnel and medical workers in the hospital can quickly communicate with each other, find problems in the cloud platform and solve them.

Establishing hospital data platform is one of the ways to solve the problem of hospital big data processing. Hospital data platform mainly includes data storage layer, business component layer, data interaction layer, hardware network infrastructure layer, and four levels. Two major systems: standard specification system and safety guarantee system. See Figure 4.1 for details. Building a hospital data platform and gradually realizing a new data exchange processing mode of unified and efficient, resource integration, interconnection and information sharing is an important trend in big data processing.

Nonparametric estimation efficiency is mainly the data envelopment analysis, the main ideas of this way of efficiency evaluation is by looking at a large number of actual production data, and based on the production effectiveness standard to find out in efficient point on the surface of the production frontier, is made up of all these efficient point of a surface data envelopment, comparison of the distance between the observed value and the ideal surface, measuring the effectiveness of the technical efficiency and the effectiveness of resource allocation. Evaluate the efficiency of traditional Chinese medicine hospitals in utilizing medical resources, including human, material, and financial resources. For example, is there a problem of insufficient manpower leading to a decline in the quality of medical services, or is there a phenomenon of uneven financial allocation leading to imbalanced development in certain departments. Inspect the investment and achievements of traditional Chinese medicine hospitals in medical technology innovation. For example, can the introduction, research and application of new technologies improve the quality and efficiency of medical services. Understand whether the economic burden of patients is reasonable, as well as the measures and effectiveness of traditional Chinese medicine hospitals in reducing the burden on patients. For example, is there any phenomenon of excessive examination, excessive treatment, etc. that increases the financial burden on patients.It should be noted here that as a non-parametric efficiency evaluation method, data envelopment analysis has the advantages of wide application, simple operation and relatively easy data acquisition. However, this method can only be used

Table 4.2: Asset growth rate

Number of years	Growth rate
2	16.59%
4	18.71%
6	19.58%

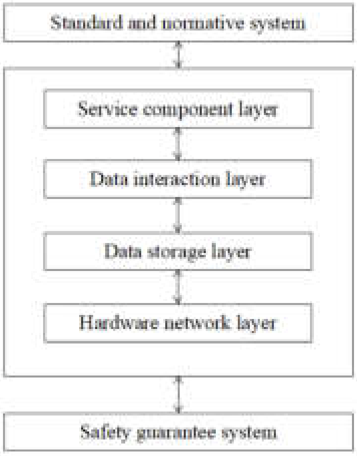


Fig. 4.1: Hospital data platform technology architecture

to evaluate the relative efficiency. Even if the efficiency value of all decision making units is 1, it cannot be concluded that all decision making units are technically effective. Data envelopment analysis can only judge the relative effectiveness of technical efficiency between a group of decision making units. Therefore, when we want to evaluate the absolute level of hospital technical efficiency, data envelopment analysis is inadequate. Using big data analysis technology, and the support of knowledge base and rule base, the medical behavior supervision system conducts big data intelligent analysis on clinicians' violations, automatically captures the problematic medical orders or prescriptions, and displays them to relevant auditing personnel to conduct violations. Review and break down the problem to find the specific root cause of the problem. At the same time, after the system automatically generates the complaint file for the violation problem, the system automatically warns the clinician that there is a violation document, and the clinician can appeal the case. For example, timely monitoring and capturing of the problem of irrational drug use interactions, the medical management department in the event and the post-mortem medical management department to review and capture the violations of the doctors, and the violations automatically captured by the system are displayed to the medical administration. Comments, thus helping the functional departments to change from terminal management to link management and process management.

The expanding medical information data is mixed with a large number of unstructured data, and the analysis data is becoming more and more diversified. The current storage architecture has been unable to meet the needs of large data applications, especially when dealing with and querying large data sets. Massive data storage system must have the corresponding level of expansion capacity. In addition to the huge scale of data, it also has a huge number of files, so how to manage the metadata accumulated at the file system level is also a difficult problem. There are real-time problems in the application of large medical data, which require real-time or quasi-real-time data processing and second-level query response. In recent years, with the development of medical information, cloud computing and the application of large data model, medical data has shown explosive growth, as shown in Figure 4.2.

In the classification and prediction analysis of medical image data, we classify them according to the char-

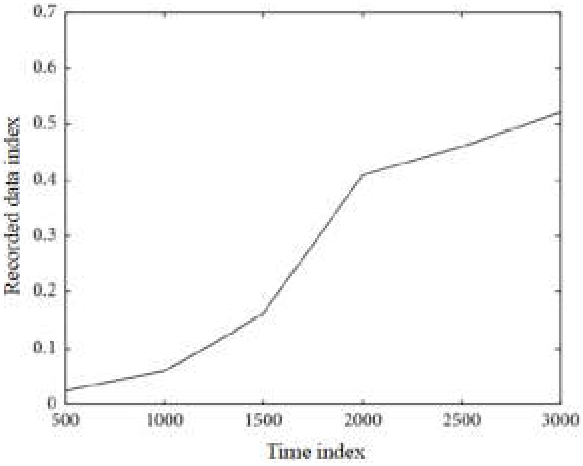


Fig. 4.2: The explosive growth of big medical data

acteristics of images and obtain knowledge and rules to predict future information. In the evaluation method of hospital medical behavior supervision and operational efficiency based on big data platforms, multidimensional analysis of image data can effectively extract and analyze key information in medical images. By constructing multidimensional features such as color, texture, and shape of images, these features can be comprehensively applied for deeper analysis and evaluation. Color feature is one of the basic features of an image, which can reflect the basic attributes of the image, such as brightness, contrast, and saturation. In medical image analysis, the extraction of color features can help doctors better understand and diagnose the condition. For example, in medical imaging, the color differences of different tissues can help doctors identify lesion areas. By extracting and analyzing color features, quantitative descriptions of color distribution, color composition, and color changes in images can be provided, providing important basis for medical behavior supervision and operational efficiency evaluation. In the similarity retrieval of image data, we can use features based on image color histogram, features based on image multi-feature composition, features with regional granularity based on wavelet or features based on image wavelet to conduct image similarity retrieval. In association mining of image data, we also mine association rules according to image features. In a particular cluster configuration, you must find a balance between cluster performance and the number of tasks. In the experiment, except for the network connection to the clinical data center, all the experimentally related hosts use the same hardware configuration and network bandwidth to ensure that there is no other than the system architecture (ie distributed and non-distributed). Differences in system performance. The specific performance is shown in Figure 4.3. When the data scale increases to reach the bottleneck of the protective gear processing, the distributed system platform can increase its capacity by increasing the number of nodes, which has good scalability, as shown in Figure 4.4.

By analyzing the utilization of medical resources, it is possible to understand whether hospitals have problems such as resource waste and insufficient resources. For example, if the bed usage rate in certain departments is too high, it may be necessary to increase the number of beds or adjust the department layout. By analyzing the treatment cycle of patients, we can understand whether there are any problems with the hospital's treatment process and whether optimization is needed. For example, if the treatment cycle of some patients is too long, it may be necessary to identify the cause and make improvements.

5. Conclusions.. This article introduces the development and application of big data analysis in the field of information technology. By studying the regulation of medical service behavior in hospitals, using top-level design to standardize the content of medical behavior regulation, exploring the effectiveness of regulation, and achieving the goal of reducing clinical paperwork. Based on the needs, a regulatory system for medical service behavior has been proposed and constructed. Strengthen the supervision of hospital medical behavior through hospital big data analysis and knowledge base system support. By comparing the distance between observation

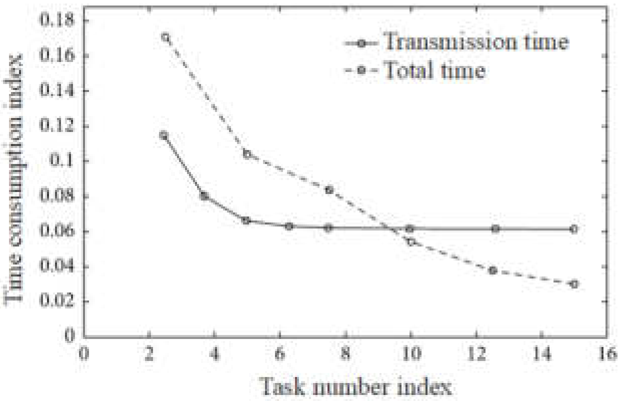


Fig. 4.3: Data collection comparison experiment results

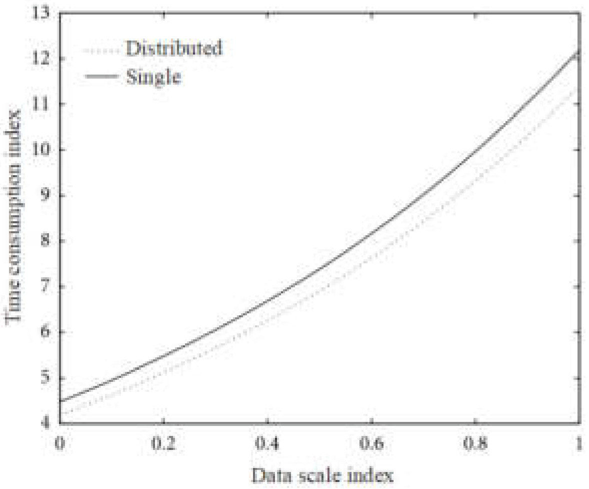


Fig. 4.4: Big data analysis performance results

values and ideal surfaces, as well as the efficiency of measurement techniques and the effectiveness of resource allocation, decision support can be provided for hospital management. Help them better understand the issues of medical behavior norms and operational efficiency, and develop corresponding improvement strategies. At the same time, this method can also provide better medical services for patients, improve medical quality and safety.

REFERENCES

[1] Schüssler, F. R. S. M., Contrepois, K., Moneghetti, K. J., et al. *A longitudinal big data approach for precision health. Nature medicine, 25(5): 792-804, 2019.*

[2] Hernandez, B. T., Bozkurt, S., Ioannidis, J. P. A., et al. *Minimar (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. Journal of the American Medical Informatics Association, 27(12): 2011-2015, 2020.*

[3] Himeur, Y., Elnour, M., Fadli, F., et al. *AI-big data analytics for building automation and management systems: a survey, actual challenges and future perspectives. Artificial Intelligence Review, 56(6): 4929-5021, 2023.*

[4] Serhani, M. A., T. El, Kassabi, H., Ismail, H., et al.*ECG monitoring systems: Review, architecture, processes, and key*

challenges. Sensors, 20(6): 1796, 2020.

[5] Sun, L., Shang, Z., Xia, Y., et al. *Review of bridge structural health monitoring aided by big data and artificial intelligence: From condition assessment to damage detection. Journal of Structural Engineering, 146(5): 04020073, 2020.*

[6] Mader, T. J., Nathanson, B. H., Soares, W. E., et al. *Comparative Effectiveness of Therapeutic Hypothermia After Out-of-Hospital Cardiac Arrest: Insight from a Large Data Registry. Therapeutic Hypothermia and Temperature Management, 4(1): 21-31, 2014.*

[7] Chen, C. M., Jyan, H. W., Chien, S. C., et al. *Containing COVID-19 among 627,386 persons in contact with the diamond princess cruise ship passengers who disembarked in Taiwan: big data analytics. Journal of medical Internet research, 22(5): e19540, 2020.*

[8] Raj, D. J. S., *A novel information processing in IoT based real time health care monitoring system. Journal of Electronics and Informatics, 2(3): 188-196, 2020.*

[9] Palanisamy, V., Thirunavukarasu, R., *Implications of big data analytics in developing healthcare frameworks–A review. Journal of King Saud University-Computer and Information Sciences, 31(4): 415-425, 2019.*

[10] Yang, D., Wu, L., Wang, S., et al. *How big data enriches maritime research–a critical review of Automatic Identification System (AIS) data applications. Transport Reviews, 39(6): 755-773, 2019.*

[11] Hayashida, K., Murakami, G., Matsuda, S., et al. *History and profile of diagnosis procedure combination (DPC): development of a real data collection system for acute inpatient care in Japan. Journal of epidemiology, 31(1): 1-11, 2021.*

[12] Selvaraj, S., Sundaravaradhan, S., *Challenges and opportunities in IoT healthcare systems: a systematic review. SN Applied Sciences, 2(1): 139, 2020.*

[13] Kadhim, K. T., Alsahlany, A. M., Wadi, S. M., et al. *An overview of patient's health status monitoring system based on internet of things (IoT). Wireless Personal Communications, 114(3): 2235-2262, 2020.*

[14] Ageed, Z. S., Zeebaree, S. R. M., Sadeeq, M. M., et al.*Comprehensive survey of big data mining approaches in cloud systems. Qubahan Academic Journal, 1(2): 29-38, 2021.*

[15] Goodday, S. M., Atkinson, L., Goodwin, G., et al. *The true colours remote symptom monitoring system: a decade of evolution. Journal of medical Internet research, 22(1): e15188, 2020.*

[16] Ghazal, T. M., Hasan, M. K., Alshurideh, M. T., et al. *IoT for smart cities: Machine learning approaches in smart healthcare—A review. Future Internet, 13(8): 218, 2021.*

[17] Cheng, W. C., Chiu, M. H. P., *How do medical researchers use open health data? A case study on data reuse behavior of using Nhird in Taiwan. Proceedings of the Association for Information Science and Technology, 54(1): 637-639, 2017.*

[18] Tzanis, G., *Biological and Medical Big Data Mining. International Journal of Knowledge Discovery in Bioinformatics, 4(1): 42-56, 2017.*

[19] Luo, J., Wang, Z., Xu, L., et al. *Flexible and durable wood-based triboelectric nanogenerators for self-powered sensing in athletic big data analytics. Nature communications, 10(1): 5147, 2019.*

[20] Schaeffer, B., Lawrence, et al. *Big Data Management in US Hospitals: Benefits and Barriers. Health Care Manag, 36(1): 87-95, 2017.*

[21] Dong, C. Z., Catbas, F. N., *A review of computer vision–based structural health monitoring at local and global levels. Structural Health Monitoring, 20(2): 692-743, 2021.*

[22] Rahman, S., Montero, M. T. V., Rowe, K., et al. *Epidemiology, pathogenesis, clinical presentations, diagnosis and treatment of COVID-19: a review of current evidence. Expert review of clinical pharmacology, 14(5): 601-621, 2021.*

[23] Hossain, M. S., Muhammad, G., Guizani, N., *Explainable AI and mass surveillance system-based healthcare framework to combat COVID-I9 like pandemics. IEEE Network, 34(4): 126-132, 2020.*

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Oct 30, 2023
Accepted: Nov 24, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1863–1871, DOI 10.12694/scpe.v25i3.2775

OPTIMIZATION OF INTERNAL CONTROL FOR BUDGET OPERATIONS IN PUBLIC INSTITUTIONS BASED ON RANDOM FOREST ALGORITHM

YANG JIN*

Abstract. In order to promote the construction of internal control in public institutions and improve work efficiency, the author proposes an optimization of internal control in public institution budget business based on random forest algorithm. We have constructed a big data audit framework for internal control of A Maritime Bureau based on the financial cloud platform and sorted out its audit process. By using the random forest algorithm to identify suspicious points in the internal control audit of administrative institutions at the data level, an example analysis is conducted using some data from A Maritime Bureau's assets, budget, revenue and expenditure, infrastructure, and contract business. The results indicate that the design of the internal control big data audit plan for administrative institutions will promote the innovation of audit information technology and application in A Maritime Bureau, provide theoretical guidance for the internal control big data audit carried out by administrative institutions, and effectively solve the problems of high workload and low work efficiency when A Maritime Bureau conducts internal control big data audits, thereby improving audit efficiency.

Key words:
Random forest algorithm; Public utility units; Budget business; internal controls

1. Introduction. The main purpose of internal control is to control risks, continuously improve control measures around internal control objectives, and implement them in place to play a role in risk prevention and control. When carrying out internal control work for budget business in public institutions, budget work is carried out in accordance with relevant policies and regulations, in order to achieve internal control objectives for budget business, continuously improve the internal control system for budget business, effectively solve problems in the process of budget work, promote pre planning, in-process control, and post assessment of budget business, and urge public institutions to do a good job in fund management, in order to better fulfill public service functions [1]. The internal control of budget business can play a good promoting role in the budget management work of public institutions [2,3].

The management content involved in internal control of budget business in public institutions mainly includes risk assessment of budget business, separation of incompatible positions, construction of budget and financial information systems, organizational structure setting, internal control environment, supervision and assessment mechanisms, etc.

When carrying out specific budget operations, internal control work involves many links such as budget preparation, approval, execution, adjustment, final accounting, and evaluation. Public institutions need to implement budget work into specific business operations, and internal control is responsible for controlling this process. Public institutions need to conduct regular or irregular risk assessments on budget operations, and timely identify risk points based on the degree and type of risk, and then take effective measures to address them. The orderly implementation of internal control over budget business in public institutions can ensure the efficient use of funds and legal compliance of budget activities, effectively supervise budget revenue and expenditure, prevent fraudulent behavior, ensure the authenticity and completeness of budget information, and promote the stable operation of public institutions to achieve established work goals.

With the development and popularization of new information technologies, emerging technologies such as big data and cloud computing have pushed social informatization to a new peak. The ecological pattern of interaction between the Internet space and real society has deeply influenced social change.

More and more enterprises, national governments, and administrative institutions have established cloud

*School of Economics, Henan Polytechnic Institute, Nanyang, Henan, China, 473000 (2003015@hnpi.edu.cn)

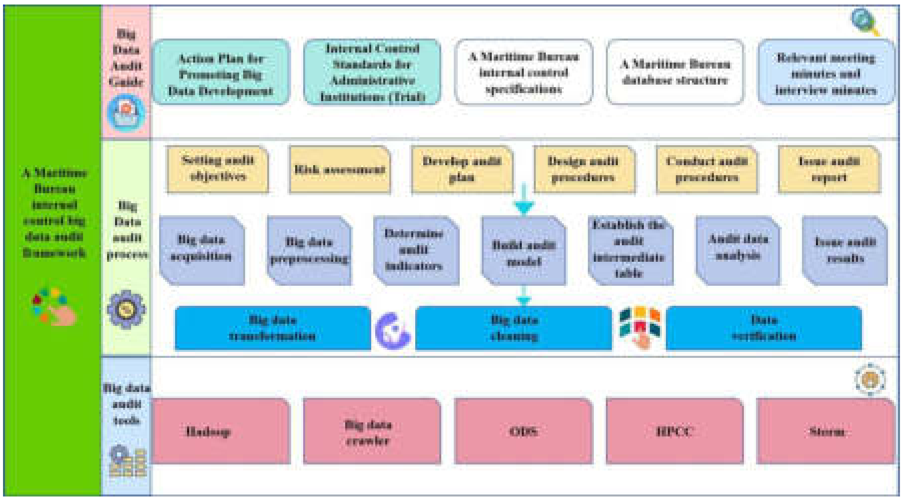


Fig. 2.1: A Maritime Administration's Internal Control Big Data Audit Framework

systems related to their actual business, such as the Golden Tax Project of tax authorities Enterprise financial shared service centers, among which many are financial cloud platforms [4]. In order to strengthen internal control and improve financial work efficiency, A Maritime Safety Bureau has established a financial cloud platform to connect the financial work of all subordinate units. The financial cloud platform summarizes all financial data and other related data from various departments, subordinate maritime departments, public institutions, etc. of the headquarters of A Maritime Bureau, forming big data with characteristics such as large data volume, diverse data types, fast data generation and transmission speed, and low data value density [5]. With the rise of cloud computing and big data technology, it has become a trend for administrative institutions to establish their own financial cloud platforms based on new information technologies in order to improve financial work efficiency. The financial cloud platform can effectively collect, store, and manage relevant economic business data of administrative institutions, but how auditors can effectively leverage the platform and data advantages provided by the financial cloud platform to achieve efficient internal control auditing of administrative institutions from the data level has become an urgent problem to be solved. Therefore, the author's research objective is to explore possible methods for administrative institutions to fully utilize big data audit methods based on financial cloud platforms to achieve efficient internal control auditing by designing corresponding big data audit plans for internal control in administrative institutions.

2. Methods.

2.1. Internal Control Big Data Audit Framework.

The A Maritime Bureau's financial cloud platform utilizes big data technology to integrate Kingdee K3 financial accounting software and other business systems, achieving effective collection, storage, preprocessing, and analysis of A Maritime Bureau's internal control big data, which is helpful for the subsequent audit work. The internal control big data audit framework of A Maritime Administration should include three dimensions: Big data audit guidelines, big data audit processes, and big data audit tools, as shown in Figure 2.1 [6].

2.1.1. Big Data Audit Guide Dimension. The Big Data Audit Guidelines refer to the documents and materials that need to be referenced when conducting big data audits. Among them, the "Action Outline for Promoting the Development of Big Data" issued by the State Council, the "Internal Control Standards for Administrative Institutions (Trial)" issued by the Ministry of Finance, and the internal control standards issued by the A Maritime Administration and the Maritime System are policy guidelines and important reference basis for the application of big data [7,8].

The various databases of A Maritime Administration are important data sources for conducting big data audits, and their data structure is the foundation for implementing big data audits. The relevant meeting minutes and interview minutes reflect the specific implementation of the internal control system of A Maritime Bureau on the financial cloud platform, and are an important basis for conducting big data audit implementation.

2.1.2. Big Data Audit Process Dimension. The internal control big data audit process of A Maritime Safety Administration based on the financial cloud platform reflects the specific implementation process of audit work, including six steps: determining audit objectives, risk assessment, developing audit plans, designing audit procedures, executing audit procedures, and issuing audit reports, the design of audit procedures can be further divided into steps such as big data collection, big data preprocessing, determination of audit indicators, construction of audit models, establishment of intermediate tables, analysis of audit data, and issuance of audit results. These are specific methods for carrying out big data audit implementation work.

Through this process, a large amount of data with a wide variety can be obtained the internal control big data with fast generation and transmission speed and low value density can be transformed into audit evidence that can provide support for audit doubts.

2.2. A Maritime Bureau's Internal Control Big Data Audit Process.

2.2.1. Determine audit objectives. Clarifying the audit objectives of internal control in administrative institutions is a prerequisite for discovering audit doubts. In the cloud accounting environment, auditors not only need to consider the compliance of the internal control system design and the effectiveness of internal control execution in administrative institutions, but also need to consider the integrity of internal control under the cloud platform when determining audit objectives [9].

2.2.2. Risk assessment. The audit risk of A Maritime Bureau's internal control big data audit under cloud accounting is not only negatively related to the degree of importance, but also closely related to whether the process and results of big data preprocessing meet certain processing rules. If the preprocessing process of big data does not comply with relevant regulations or the processing results do not meet audit requirements, audit risks will increase [10,11]. At the same time, the credibility of A Maritime Bureau's financial cloud system is also related to audit risks. The low credibility of A Maritime Bureau's financial cloud system not only affects the final audit results, but also greatly increases audit risks. In order to control audit risks, auditors should monitor the entire process of big data preprocessing, at the same time, the credibility evaluation results of A Maritime Bureau's financial cloud system and other business systems were obtained through third-party experts.

On the basis of considering the importance of the audit business itself, combined with the evaluation results of A Maritime Bureau's financial cloud system provided by third-party experts, the possible audit risks are ultimately identified and evaluated qualitatively based on the types and indicators of risks.

2.2.3. Develop audit plan. In the cloud accounting environment, the audit plan for big data audits should plan an audit schedule that includes audit time, audit scope, and audit human resources. In particular, the time and manpower required for the big data preprocessing process in big data audits should be reflected in the audit schedule [12]. Among them, the audit scope of big data auditing has become larger compared to traditional audit methods.

After big data preprocessing, various types of data stored in the audit data warehouse, including business data and external data of A Maritime Bureau, can be selected by auditors as long as they are related to internal control of A Maritime Bureau, rather than only focusing on the causal relationship between data and business.

2.2.4. Design audit procedures. The audit procedure for the internal control big data audit of A Maritime Safety Administration should include steps such as big data collection, big data preprocessing, determining audit indicators, constructing audit models, establishing intermediate tables, analyzing audit data, and issuing audit results. In the process of designing the internal control audit program for A Maritime Bureau, algorithmic tools such as SQL statements, database technology, semi Markov chain models, and random forest algorithms can be used to implement big data audits [13].

Data collection: In order to meet the internal control big data audit needs of A Maritime Bureau, auditors should collect the actual business process data of A Maritime Bureau on the cloud accounting platform, including business data and financial data. A Maritime Bureau's financial cloud platform can store data from different format business systems such as DBMS, Filc, Excel, etc. in the business synchronous replication database, then use ETL (extraction, transformation, loading) tools to preprocess the data in the business synchronous replication database. At the same time, auditors should also obtain the current internal control system of administrative institutions, audit knowledge base, and standard documents such as the ”Internal Control Standards for Administrative Institutions (Trial)”, and uniformly store them in the audit database for management.

Building an audit model: Auditors can build an audit model based on audit indicators. The audit model is an important tool for implementing data auditing, which can be constructed in various ways. The author intends to use a semi Markov chain model to achieve mathematical expression of internal control in administrative institutions, in order to structure internal control data, and then use random forest algorithm to construct a specific audit model, and use the algorithm to discover audit doubts in internal control of administrative institutions.

There are many algorithms that can be used in the process of discovering audit doubts, such as BP neural network algorithm, naive Bayesian algorithm, random forest algorithm, etc. The random forest algorithm has the advantages of fast training speed, good noise resistance, and simple implementation. Therefore, the author chooses to use the random forest algorithm to discover suspicious points in internal control audits of administrative institutions [14]. The principle of the random forest algorithm is to randomly extract a portion of data samples from the target dataset to construct multiple decision tree models, which are training sets. Multiple decision trees composed of training set data samples are independent of each other, forming a random forest. After completing the construction of the random forest model, the samples from the original dataset need to be judged separately by each decision tree in the forest to determine which category the sample should belong to. The final judgment result is voted by the judgment results of all decision trees in the random forest, and the majority of votes obtained is the final judgment result. The use of random forest algorithm in the discovery of audit doubts is to extract training sets from the thematic data mart to construct a random forest model, and then make judgments. After data collection, data preprocessing, and setting audit indicators, the raw data collected by the auditor should be converted into a dataset mathematically expressed by a semi Markov chain model, which includes eigenvalues used for decision tree judgment, such as state transition probability matrix, dwell time matrix, etc., and stored in a thematic data mart according to different business processes and audit indicators required by internal control, then, a random forest algorithm can be used to construct an audit model and identify audit doubts.

Constructing a training set using an audit knowledge base to construct a random forest model: The following assumptions need to be made when constructing a training set using an audit knowledge base to construct a random forest model.

Assumption 1: The data samples without audit doubts in the training set are normal data samples, and the data samples with audit doubts are doubtful data samples.

Assumption 2: When the judgment result of the decision tree model shows that there are doubts, the output result is y=0, if it is determined that there are no audit doubts, the output result is y=1.

Assuming that the final judgment result of the random forest algorithm is determined by the voting of each decision tree, the final judgment made by the random forest model can be expressed as:

$$z = \begin{cases} 1, count(y=1) > count(y=0) \\ 0, count(y=1) < count(y=0) \end{cases} \tag{2.1}$$

Using the random forest model for auditing, it was found that based on the previous assumption, when the sample size of the training set is large enough, the voting results of the random forest can divide the original dataset into two subsets: z=1 and z=0, usually, the data samples in the z=1 subset are normal data samples, while the data samples in the z=0 subset have audit doubts [15].

By constructing an audit model and audit intermediate table, preliminary audit doubts can be obtained. For the sake of audit prudence, further analysis of audit doubts still requires professional judgment or other means from auditors. After analyzing the audit data, the final audit results can be obtained. The audit doubts need to be sorted and summarized by the auditors to form a summary table of audit doubts, and communicated with the leaders of administrative institutions to conduct preliminary verification of the doubts

Table 3.1: Internal Control of Asset Business

<table>
<tr><td>classification</td><td>Number of samples</td><td colspan="3" rowspan="3">-</td></tr>
<tr><td>Normal Data Samples</td><td>1420</td></tr>
<tr><td>Doubtful data samples</td><td>3</td></tr>
<tr><td>Business number</td><td>Accounting Month</td><td>Accounting voucher number</td><td>amount of money</td><td>Asset nature</td></tr>
<tr><td>470</td><td>05</td><td>33</td><td>16000</td><td rowspan="2">Science and Technology Information Department</td></tr>
<tr><td>1184</td><td>12</td><td>02</td><td>50000</td></tr>
</table>

and clarify whether there are audit doubts caused by force majeure. These doubts and subsequent verification need to be separately listed in the internal control audit report.

2.2.5. Execute audit procedures. Execute audit procedures according to the designed audit plan, monitor the entire process of obtaining, cleaning, and storing big data, and combine third-party experts or institutions' comprehensive evaluation of A Maritime Bureau's financial cloud system and other business systems to fully leverage the advantages of combining financial and business data brought by A Maritime Bureau's financial cloud system with other business systems of the enterprise, forming audit doubts, search for audit evidence and obtain audit results. Any issues discovered during the big data audit process should be promptly reported to the relevant leaders and management personnel of A Maritime Bureau, and these issues also need to be reviewed and evaluated.

2.2.6. Issue audit report. Big data auditing reflects the business from a data perspective, and at the same time, doubts can only be discovered from the data. Whether there are problems in the actual business needs to be further collected based on the audit doubts. According to the process of big data auditing, the audit doubts obtained by auditors through big data auditing methods are classified according to audit indicators, and should ultimately be summarized to form a summary Table of audit doubts. Then, according to the actual audit business needs, tasks are assigned for audit evidence collection, and audit doubt point evidence collection reports are prepared to provide the final audit results [16]. When issuing audit results, summarize the audit opinions obtained through the implementation of big data audits and audit evidence collection, and combine the business background of the big data audit and the initially established audit objectives to obtain the final audit results. Then, based on the problems discovered during the audit process, propose management suggestions to the relevant leaders and management personnel of A Maritime Safety Bureau, and after communicating with them, obtain their relevant responses to the management suggestions, finally, issue an audit report.

3. Results and Analysis.

There are many types of business in A Maritime Bureau, and the author intends to analyze the internal control big data audit of A Maritime Bureau using data from some of the five types of business: Assets, budget, revenue and expenditure, infrastructure, and contracts as examples.

3.1. Internal Control Big Data Audit of Asset Business of A Maritime Administration.

3.1.1. Internal Control Audit Doubts Discovery Based on Random Forest Algorithm. After completing the mathematical expression of the internal control process, data samples were extracted from the audit knowledge base using Python language to form a training set, and a random forest algorithm was used to analyze some asset disposal data for 2016. The final summary results are shown in Table 3.1 [17]. The experimental results indicate that out of the 1421 selected data samples, 1419 are normal data samples, and 2 are doubtful data samples. The doubtful data samples reflect the asset disposal business of A Maritime Bureau, which is recorded in the 3rd accounting voucher in April and the 1st accounting voucher in November, respectively.

Table 3.2: Internal Control of Asset Business

Business number	Accounting Month	Accounting voucher number	amount of money	Asset nature
19	10	168	3250	Marine Department

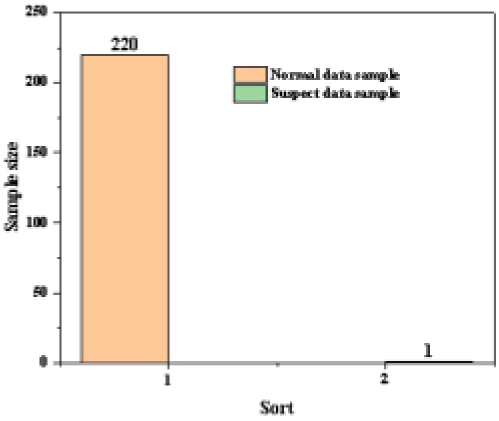


Fig. 3.1: Comparison of Findings of Doubtful Points in Internal Control Audit of Accounting Business

3.1.2. Verification of audit doubts . After using the random forest algorithm to find audit doubts, it is necessary to verify the audit doubts, that is, in order to search for raw data through SQL statements in the SQLSERVER2008R2 environment to confirm the effectiveness of the random forest algorithm in the discovery of internal control audit doubts [18]. Through verification, it was found that the asset recorded in the accounting voucher No.33 of A Maritime Bureau in April belongs to the bureau, but was approved by a subordinate maritime department leader in the actual disposal process; The asset recorded in the November 2nd accounting voucher of the A Maritime Safety Bureau belongs to technical assets, but during the disposal process, it was not approved by the Science and Technology Information Department, and instead was approved by the Infrastructure and Equipment Department.

3.2. A Maritime Bureau Budget Business Internal Control Big Data Audit.

3.2.1. Internal Control Audit Doubts Discovery Based on Random Forest Algorithm. After completing the mathematical expression of the internal control process, data samples were extracted from the audit knowledge base using Python language to form a training set. The random forest algorithm was used to analyze some additional budget adjustment data for 2016. The final summary results are shown in Table 3.2 and Figure 3.2. The experimental results indicate that out of the 221 selected data samples, 220 belong to normal data samples, and 1 is a doubtful data sample. The doubtful data sample reflects the additional budget adjustment business of A Maritime Bureau, which was recorded in the accounting voucher No. 168 in October.

3.2.2. Verification of audit doubts. After using the random forest algorithm to find audit doubts, it is necessary to verify the audit doubts, that is, in order to search for raw data through SQL statements in the SQLSERVER2008R2 environment to confirm the effectiveness of the random forest algorithm in the discovery of internal control audit doubts. Through verification, it was found that the application data submitted by the grassroots maritime department in the accounting voucher No.68 of A Maritime Bureau in October is missing, and the reason is unknown.

Table 3.3: Internal Control of Revenue and Expenditure Business

classification	Number of samples	-		
Normal Data Samples	6643			
Doubtful data samples	3			
Business number	Accounting Month	Accounting voucher number	Amount of money	Nature of expenditure
13	1	84	624	Daily expenses
142	3	66	1443	Daily expenses

Table 3.4: Internal Control of Infrastructure Business

Classification	Number of samples
Normal Data Samples	30
Doubtful Data Samples	0

3.3. Internal Control Big Data Audit of Revenue and Expenditure Business of A Maritime Bureau.

3.3.1. Internal Control Audit Doubts Discovery Based on Random Forest Algorithm. After completing the mathematical expression of the internal control process, data samples were extracted from the audit knowledge base using Python language to form a training set, and a random forest algorithm was used to analyze some expenditure approval data for 2016. The final summary results are shown in Table 3.3 [19]. The experimental results show that out of the 6644 selected data samples, 6642 belong to normal data samples, and 2 are doubtful data samples. The doubtful data samples reflect the expenditure approval business of A Maritime Bureau, which is recorded in the accounting voucher No.84 in January and the accounting voucher No.66 in March, respectively.

3.3.2. Verification of audit doubts. After using the random forest algorithm to find audit doubts, it is necessary to verify the audit doubts, that is, in order to search for raw data through SQL statements in the SQLSERVER2008R2 environment to confirm the effectiveness of the random forest algorithm in the discovery of internal control audit doubts. Through verification, it was found that the transactions recorded in the January 84th accounting voucher and March 66th accounting voucher of A Maritime Bureau should be daily expenses, but they were approved by the leaders of the responsible business units, which is inconsistent with the internal control system.

3.4. A Maritime Safety Bureau's Internal Control Big Data Audit of Infrastructure Business. After completing the mathematical expression of the internal control process, data samples were extracted from the audit knowledge base using Python language to form a training set. The random forest algorithm was used to analyze the bidding data of some infrastructure projects in 2016. The final summary results are shown in Table 3.4. The experimental results indicate that out of the 30 selected data samples, 30 belong to normal data samples, undoubtedly point data samples.

3.5. A Maritime Bureau Contract Business Internal Control Big Data Audit. After completing the mathematical expression of the internal control process, data samples were extracted from the audit knowledge base using Python language to form a training set, and a random forest algorithm was used to analyze some contract signing data for 2016. The final summary results are shown in Table 3.5 [20]. The experimental results indicate that out of the 1021 selected data samples, 1021 belong to normal data samples without any doubts.

Table 3.5: Internal Control of Contract Business

Classification	Number of samples
Normal Data Samples	1021
Doubtful Data Samples	0

4. Conclusion. Big data auditing is an audit method that applies big data technology to audit business, with the aim of reflecting the business from a data perspective. Applying big data technology to internal control auditing of administrative institutions can discover problems in the internal control system and execution of administrative institutions from a data perspective. The author constructed a big data audit framework for internal control of A Maritime Bureau and sorted out the audit process. In terms of specific audit methods, a combination of random forest algorithms was used to effectively utilize a large amount of data from the cloud, and an audit model was constructed to carry out the audit work. Finally, relevant data on internal control of A Maritime Bureau's assets, budget, revenue and expenditure, infrastructure projects, and contract business were analyzed as examples. Research has shown that big data technology can effectively assist auditors in discovering internal control doubts in administrative institutions from a data perspective, greatly reducing audit workload, improving audit efficiency, and solving problems in internal control audits of administrative institutions, achieving comprehensive and real-time audits, meeting new requirements in the context of new technologies. Big data auditing utilizes big data technology for auditing, which can effectively help auditors improve audit efficiency while having a data foundation. But at present, big data auditing has just started to rise and has not been widely applied. It is necessary to accumulate experience and summarize lessons in the practical process, and promote and practice on the basis of continuously improving the theoretical system.

REFERENCES

[1] Chen, Y., Guo, A., Chen, Q., Quan, B., & Hao, Z. . (2021). Intelligent classification of antepartum cardiotocography model based on deep forest. Biomedical Signal Processing and Control, 67(2), 102555.

[2] Azhar, Y., Mahesa, G. A., & Mustaqim, M. C. . (2021). Prediction of hotel bookings cancellation using hyperparameter optimization on random forest algorithm. Jurnal Teknologi dan Sistem Komputer, 9(1), 15-21.

[3] Abbassi, A., Mehrez, R. B., Abbassi, R., Saidi, S., Albdran, S., & Jemli, M. . (2022). Improved off-grid wind/photovoltaic/hybrid energy storage system based on new framework of moth-flame optimization algorithm. International Journal of Energy Research, 46(5), 6711-6729.

[4] Zhang, Z., & Bai, D. . (2022). Optimization of improved pid control strategy based on genetic algorithm. Journal of Physics: Conference Series, 2417(1), 012025-.

[5] Zheng, X., Yi, S., & Deng, X. . (2021). Evaluation model construction of automobile appearance design based on random forest algorithm. Journal of Physics: Conference Series, 1941(1), 012072 (9pp).

[6] Zhang, X., & Wang, M. . (2021). Weighted random forest algorithm based on bayesian algorithm. Journal of Physics: Conference Series, 1924(1), 012006 (6pp).

[7] Zvonareva, T. A., Kabanikhin, S. I., & Krivorotko, O. I. . (2023). Numerical algorithm for source determination in a diffusion–logistic model from integral data based on tensor optimization. Computational Mathematics and Mathematical Physics, 63(9), 1654-1663.

[8] Li, Y., Li, F., & Song, J. . (2021). The research of random forest intrusion detection model based on optimization in internet of vehicles. Journal of Physics Conference Series, 1757(1), 012149.

[9] Lin, H., & Tang, C. . (2021). Analysis and optimization of urban public transport lines based on multiobjective adaptive particle swarm optimization. IEEE Transactions on Intelligent Transportation Systems, PP(99), 1-13.

[10] Zhao, Y., Ren, X., & Zhang, X. . (2021). Optimization of a comprehensive sequence forecasting framework based on dae-lstm algorithm. Journal of Physics: Conference Series, 1746(1), 012087 (12pp).

[11] Yu, H. . (2021). Economic dispatching optimization of power grid based on igwo algorithm. Journal of Physics: Conference Series, 1748(3), 032009 (6pp).

[12] Yiyue, L., Yu, F., & Xianjun, C. . (2021). Research on optimization of deep learning algorithm based on convolutional neural network. Journal of Physics: Conference Series, 1848(1), 012038 (5pp).

[13] Sun, J., Guo, B., Hu, Y., & Zhang, Y. . (2021). Multi-objective optimization of spectrum sensing and power allocation based on improved slime mould algorithm. Journal of Physics: Conference Series, 1966(1), 012018-.

[14] Li, Y., Chen, J., Xu, X., Lin, Z., & Chen, X. . (2021). Optimization of garbage dumping mechanism of intelligent sanitation vehicle based on particle swarm algorithm. Journal of Physics: Conference Series, 1939(1), 012063 (8pp).

[15] He, Y. H., Luo, Y., Li, A. H., Wang, T. F., & Peng, Y. H. . (2021). Research on protection optimization of distribution network containing distributed power generation based on sparrow algorithm. Journal of Physics Conference Series,

1820(1), 012147.

[16] Lu, C., Xian, X., & Li, C. . (2021). Research on optimization of fuzzy network control system based on new smith predictive time delay compensation. Journal of Physics: Conference Series, 2132(1), 012019-.

[17] Hu, B., & Li, J. . (2021). An edge computing framework for powertrain control system optimization of intelligent and connected vehicles based on curiosity-driven deep reinforcement learning. IEEE Transactions on Industrial Electronics, 68(8), 7652-7661.

[18] Jiang, F., Sha, K., Lin, C., & Wu, Z. . (2023). Node layout optimization strategy based on aquaculture water quality monitoring system. Wireless Personal Communications, 132(4), 2839-2856.

[19] Gülah Gülba, & Gürcan etin. (2023). Lifetime optimization of the leach protocol in wsns with simulated annealing algorithm. Wireless Personal Communications, 132(4), 2857-2883.

[20] Dao, T. K., Nguyen, T. T., Ngo, T. G., & Nguyen, T. D. . (2023). An optimal wsn coverage based on adapted transit search algorithm. International Journal of Software Engineering and Knowledge Engineering, 33(10), 1489-1512.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Oct 31, 2023
Accepted: Nov 15, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1872–1881, DOI 10.12694/scpe.v25i3.2779

RESEARCH AND DESIGN OF AN AUTOMATED SECURITY EVENT ANALYSIS AND HANDLING FRAMEWORK BASED ON THREAT INTELLIGENCE

LINJIANG XIE, ZHOUYUAN LIAO[*] AND HANRUO LI

Abstract. In order to deeply explore and utilize the value of threat intelligence, strengthen research on attack organizations, and grasp the correlation between attack organizations, the author proposes the research and design of an automated security event analysis and handling framework based on threat intelligence. The author extracts the behavioral characteristics of the attack organization based on known APT attacks, and uses the machine learning framework Light GBM to establish a multi classification model to complete the analysis of unknown APT attack organizations. Through the study of multi-dimensional analysis of multi-source threat intelligence, attack organization correlation and judgment, an attack organization correlation and judgment system has been designed and implemented. The system includes six modules: threat intelligence collection module, threat intelligence multi-dimensional analysis module, attack organization fingerprint library module, attack organization correlation module, attack organization analysis module, and user module, providing attack organization correlation and judgment services for network security. The test results show that the intelligence reading and search query function can achieve the reading of various information of attack organizations, and achieve visual display of threat intelligence. The intelligence management function can achieve operations such as adding, deleting, and updating intelligence. The user management function of the system can achieve the management of administrator users and ordinary users. After testing, all functions of the system have been implemented and meet expectations.

Key words: Threat intelligence; Security incidents; Judgment and disposal; Design; APT attack

1. Introduction. With the rise of big data and the development of ”Internet plus”, the collection scope of threat intelligence has been greatly expanded. Threat intelligence describes existing or imminent threats or dangers to assets, and can notify the subject to make some response to the relevant threats or dangers [1]. The modern information technology represented by the Internet, especially mobile payments, cloud computing, social networks, and search engines, has had a fundamental impact on human financial models, accompanied by increasing external risks. How to strengthen the information security management of banks and form a scientific and effective information security management system to prevent financial information risks has become a major issue currently facing the financial industry. Organization and scale are increasingly becoming the most prominent characteristics of network attacks, and each attack is carried out with premeditation. Planned and often with political objectives, more and more national dreams and the construction of attack organizations and the implementation of attack behaviors have increased the difficulty of network defense due to the background of national cyber attacks.

At present, the implementation of relevant security technologies does not receive legal protection, such as extracting P message information. Monitoring device status and other issues are all related to personal privacy, and these issues cannot be solved solely through technological means. They rely more on the updating and improvement of network security laws and regulations. In order to better maintain cyberspace security and respond to APT attacks, threat intelligence has emerged. Threat intelligence is seen as a ”visible” tool for quickly and effectively resisting attacks and maintaining cyberspace security Ability. From a personal perspective, threat intelligence can serve as an important defense indicator, and attackers also need threat intelligence to update technical means and achieve the goal of hiding their true identity through attack information forgery and intelligence. From the perspective of enterprises, different security companies can exchange different threat intelligence, such as vulnerability information[2]. P reputation information, license list. White list, etc., to achieve an integrated network of internal and external resources, thereby achieving collaborative street prevention, and forming products such as vulnerability detection, threat detection, etc. to generate

[*]Information Security Operation and Maintenance Center of Information Center of Yunnan Power Grid Co., LTD, Kunming, Yunnan, China, 650011 (Corresponding author, bgzzzy@126.com)

commercial profits.

For a country, threat intelligence is crucial for maintaining national security and safeguarding national interests. Safeguarding the legitimate rights and interests of citizens is of great significance, and even includes attackers. Attacking organizational information plays a strategic role in diplomacy and national defense. At present, research on threat intelligence mainly focuses on the productization of security vendors, with a general direction of providing traditional security products such as PS (Intrusion Prevention System) and information exchange modes. However, the value of threat intelligence is not as simple as productization, as it contains a large amount of attack organization information such as attack methods and tools. Attack behavior patterns, attack intentions, etc. can be correlated and judged through in-depth analysis of threat intelligence and value mining. Network attack and defense is essentially a major game and competition between attackers and defenders using the network as the battlefield. The analysis, research, and tracing of attack organizations can help develop targeted defense strategies to achieve effective defense and precise strikes. Therefore, the importance of threat intelligence analysis and research for attacking organizations in network defense cannot be ignored. Attackers use various techniques such as forging IP addresses, springboards, and anonymous networks to hide their identities and evade tracing, which undoubtedly poses a huge challenge to the tracing of attack organizations. However, it is precisely due to the existence of various tracing and forensics technologies such as host IP tracing, malicious code analysis, and packet logging that tracing tracing becomes possible, it is precisely the analysis results of these technologies that provide content support and technical possibilities for the association and analysis of attack organizations. At present, research on attack organizations around the world still presents fragmented characteristics, and there is no real system that can complete the analysis functions of various attack organizations. Therefore, the author's research on the correlation and judgment of attack organizations based on multidimensional analysis of threat intelligence and its method implementation have important research and application value [3,4].

2. Methods. The attack tactics and objectives of different APT organizations are different. The attack time, attack tactics, attack objectives, IOC, etc. of the attack organization are saved as threat intelligence, which includes the characteristic information of the attack organization. By extracting these features, an attack organization fingerprint library can be formed. By using the attack organization features, an association of existing attack organizations can be established. At the same time, When the network is attacked again, the attack organization can be determined by comparing the attack traces with the attack organization fingerprint database [5]. As shown in Figure 2.1. Based on the above research ideas, the following steps are required:

Acquisition of multi-source threat intelligence: Selecting multiple sources to obtain a large amount of threat intelligence and establishing a threat request intelligence base is the foundation and prerequisite of this study.

Multidimensional analysis of threat intelligence: Based on basic theories and analysis models, conduct multidimensional analysis of threat intelligence, extract feature indicators from each dimension, and establish a threat intelligence feature library;

Establishment of attacker organization fingerprint database: Analyze and process the collected threat intelligence to establish an attack organization fingerprint database [6];

Attack organization homology determination: Attack organization homology refers to the similarity between two attack organizations. By using this similarity analysis to determine whether two attacking organizations are unified or have associations, it can be achieved through similarity analysis between different organizations;

Consistency determination of attacking organizations: Research and judgment of attacking organizations. Comparing recently captured attack events with known attack organizations to determine the category of their attack organizations can be achieved through machine learning.

2.1. Acquisition of multi-source threat intelligence. At present, there are roughly four sources of threat intelligence for security teams: Public sources, commercial sources, open source intelligence sources, and internal data. The credibility of threat intelligence sources is often measured by the authority of the intelligence source. As shown in Equation 2.1, Au represents the credibility of different sources, where S is the intelligence

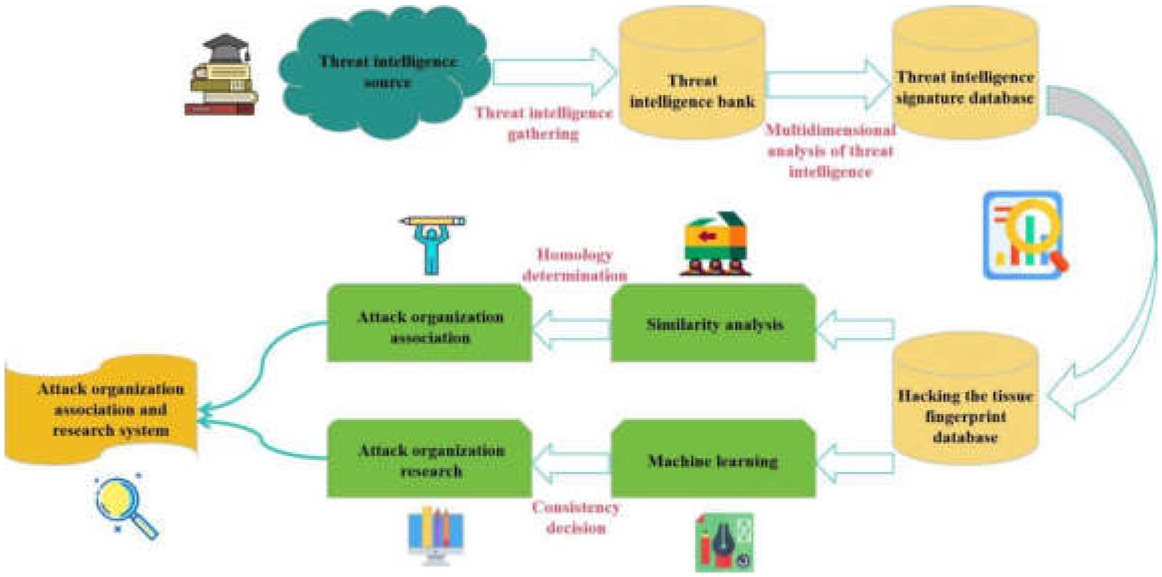

Threat intelligence source
Threat intelligence gathering
Threat intelligence bank
Multidimensional analysis of threat intelligence
Threat intelligence signature database
Hacking the tissue fingerprint database
Similarity analysis
Machine learning
Homology determination
Consistency decision
Attack organization association
Attack organization research
Attack organization association and research system

Fig. 2.1: Research ideas on attack organization association and judgment based on multi-dimensional analysis of threat intelligence

source and r is the Alexa ranking of the source site [7,8].

$$Au = \begin{cases} 0, & S\epsilon unknown \\ 0.2, & S\epsilon \text{Independent source station, blog} \\ 0.4, & S\epsilon site \quad \& \quad r < 10^6 \\ 0.6, & S\epsilon site \quad \& \quad r\epsilon[10^4, 10^6] \\ 0.8, & S\epsilon site \quad \& \quad r > 10^4 \\ 1, & S\epsilon \text{A well-known organization or institution} \end{cases} \tag{2.1}$$

The author's research is based on credible threat intelligence, and the credibility of threat intelligence from different sources also varies. Therefore, in order to ensure the credibility of threat intelligence, the author's sources of threat intelligence include public threat intelligence sources, open-source threat intelligence sources, and internal data, mainly including threat intelligence based on APT reports, shared intelligence on threat intelligence platforms, and accumulated data in practice (internal data) [9]. Research reports publicly released by national security departments, renowned security vendors or organizations, etc., on APT analysis intelligence. They rely on their own devices and platforms to mine various elements in APT attack events, then trace their sources, complete attack scenario reconstruction, and form analysis reports, which include specific analysis processes and results, as well as complete intelligence information such as attack tools, targets, and events. APT reports, as intelligence with rich knowledge content, have extremely high utilization value and are the highest level of threat intelligence. However, due to the wide distribution and large quantity of APT reports, it is difficult to collect them. At the same time, even with APT reports, accurately extracting threat intelligence from them also requires a lot of time. Therefore, this part of the workload is huge and difficult, which requires a lot of time and effort.

The publicly recognized machine readable intelligence provided by various threat intelligence sharing platforms is a clue to a network threat that has been analyzed and discovered at a certain time, and many platforms provide API interfaces, some of which require payment. At the same time, the data structure and information content of each shared platform are different, and not all threat intelligence is related to the attack organization, requiring screening and judgment.

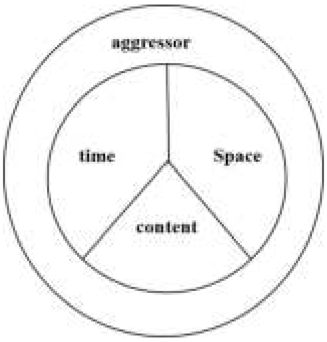


Fig. 2.2: Multidimensional Model of Threat Intelligence

2.2. Threat Intelligence Multidimensional Analysis Model. The research on attack organization involves a wide and complex range of technical means, including packet tracking, attack trace retention technology, malicious sample analysis, etc. These technologies obtain clue information of the attack organization through technical analysis of the attack process, which to some extent characterizes its characteristics [10]. The author focuses on the research of attack organizations, starting with threat intelligence, learning from the threat intelligence pyramid model, attack chain model, and diamond model, as well as different threat intelligence standards. With the ultimate requirement of attack organization association and judgment, the author integrates tactical intelligence, operational intelligence, and strategic intelligence. Through threat intelligence analysis and research, a multi-dimensional threat intelligence model guided by attack organization association and judgment is proposed. As shown in Figure 2.2, the model includes three dimensions: time dimension, spatial dimension, and content dimension, and proposes different feature indicators based on each dimension to establish a multidimensional analysis of threat intelligence. The proposal and establishment of this model mainly aims to deeply mine higher value threat intelligence information related to attack organizations through multidimensional analysis of threat intelligence, and use this model to achieve association and analysis of attack organizations.

(1) Characteristic indicators based on time dimension. Each attack is accompanied by a time characteristic, which is an inherent important characteristic of threat intelligence. Although IOC has timeliness and its role in the defense process may be reduced or even lost, for attacking organizations, each attack will last for a long time. Early threat intelligence can be used for early warning and defense in the later stage [11,12]. The time at which each attack organization initiates an attack varies, and the identity information of the attack organization can be inferred from the time. It can also be distinguished from other attack organizations. For example, APT28 (Fantasy Bear) is considered to have a Russian background, partly because during attack analysis, security engineers found that over 96% of the malware samples were compiled between Monday and Friday, and over 89% were in the UTC+4 time zone, Compiled between 8am and 6pm, similar to the working hours in Moscow and St. Petersburg, and judged to have a Russian background based on other information. At the same time, the attack time of different attackers varies over a long period of time. In order to clarify the relationship between attackers and attack time and support the time differentiation between different attack organizations, the author selected APT32 and BITTER as examples for analysis. Figure 2.3 shows the temporal distribution of all threat intelligence of APT32 and BITTER attack organizations from 2017 to 2022, with the horizontal axis representing time, the vertical axis represents the proportion of time distribution in all IOC within that time period.

From the above analysis, it can be seen that when an attacking organization initiates an attack, time clues are left behind. These time clues are also important content of threat intelligence. We can obtain the time dimension characteristics of the attacking organization by analyzing the time distribution of threat intelligence, which can be used as a classification basis for different attacking organizations. In summary, there is a certain correlation between time and attack events and attack organizations, which can be used as one of the specific indicators for classifying and judging attack organizations. Therefore, the characteristic indicator of the time

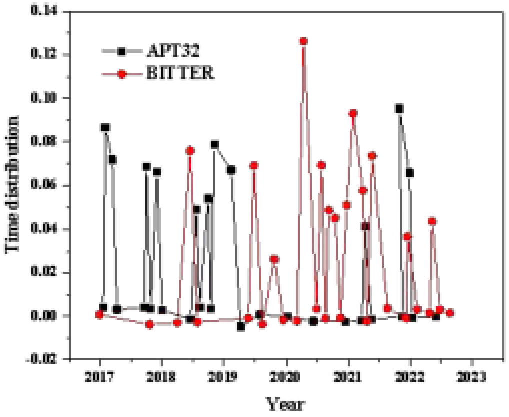


Fig. 2.3: Time distribution of APT32 and BITTER attacks

dimension of threat intelligence is the occurrence time of the attack event. Specifically, assuming that a threat intelligence is represented by ”T”, its time dimension is represented by ”Tt”, and the attack time is ”T”:

$$T_t = [t] \tag{2.2}$$

(2) Feature indicators based on content dimension. The collection of threat intelligence is a long and arduous process. Due to inadequate sharing mechanisms, the description of the same threat intelligence obtained through different sources and methods often does not have significant deviation, and there is a problem of one-sided content description [13]. Therefore, on the basis of analyzing the obtained threat intelligence, the author also comprehensively analyzed the intelligence content of multiple threat intelligence sharing platforms. In response to the main problem of the attack organization studied by the author, several characteristics with obvious directionality for the attack organization were abstracted from the multi-source threat intelligence. The main indicators include: attack target, attack purpose, attack event, code features IOC (URL, Hash, Domain, P, Emai1), Intelligence Description.

Specifically, assuming that ”T” represents a threat intelligence information with a content dimension of ”Tc”. ”C1” represents the ”attack target” of ”T”, ”c2” represents the ”attack purpose” of ”T”, ”c3” represents the ”attack event” of ”T”, and ”c4” represents the ”code feature” of ”T”, If ”c5” represents the ”URL” of ”T”, ”c6” represents the ”Hash” of ”T”, ”c7” represents the ”Domain” of ”T”, ”c8” represents the ”IP” of ”T”, ”c9” represents the ”Eamil” of ”T”, and ”c10” represents the ”intelligence description” of ”T”, then the content dimension Tc of threat intelligence ”T” can be expressed as:

$$T_c = [c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8, c_9, c_{10}] \tag{2.3}$$

(3) Feature indicators based on spatial dimensions. According to the diamond model of threat intelligence, infrastructure describes the physical or logical structures used by attackers to deliver capabilities, such as IP addresses, domain names, email addresses, etc. Due to the limitations of attack costs, the infrastructure chosen by attackers is often concentrated in a certain space, so security analysts often believe that some location information of the infrastructure is to some extent related to the attack organization or can be used to distinguish different attack organizations [14]. The ”geographic location” information contained in threat intelligence can analyze the infrastructure status of attacking organizations, determine their possible locations, and distinguish different attacking organizations, therefore it is considered as one of the indicators of the spatial dimension of threat intelligence. In addition, each attacking organization has many purposes for implementing network attacks, and is currently increasingly politicized, including the influence of ”geopolitical” factors. In

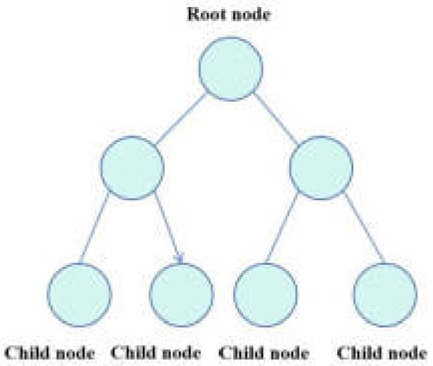


Fig. 2.4: Schematic diagram of tree structure

other words, the geographical location of the victims pointed to in these threat intelligence is also an important indicator for determining the attack organization. Therefore, the geographical location of the victim included in the threat intelligence is also considered as one of the characteristics of the spatial dimension.

In summary, the spatial dimension of threat intelligence includes two characteristic indicators, one is the distribution of address locations where the attack organization's infrastructure is located, and the other is the distribution of victim location information. Specifically, assuming that a threat intelligence information is represented by ”T”, its spatial dimension is represented by ”T1”, its home location is ”l1”, and the attacked country or region is ”l2”, then the spatial dimension T1 of threat intelligence ”T” can be expressed as:

$$T_l = [l_1, l_2] \tag{2.4}$$

2.3. Attack Organization Analysis Based on LightGBM Model. The analysis of attack organizations is the determination of attack organizations for APT attacks that have already occurred, which is the consistency determination of attack organizations. It is based on the characteristics of attack organizations that have already occurred to define the initiator of new attack events. Ultimately, it is essentially a multi classification problem. There are two methods for this, one is to use the idea of attack organization similarity. Every time a new APT attack occurs, the similarity between the event and each attack organization in the attack organization fingerprint database is calculated to determine the direction of analysis. It is obvious that the time complexity of this method is linearly related to the number of known APT organizations in the fingerprint database. When the number of known APT organizations is too large, it will lead to excessive computational complexity. A more efficient method is to train multiple classification models through machine learning to complete the analysis of attack organizations.

(1) Fundamentals of Classification Models. The structure composed of nodes and edges is called a tree, as shown in Figure 2.4. The root node splits into child nodes, and the byte point splits into new child nodes by re serving as the parent node. Decision Tree, as the name suggests, utilizes a tree structure for decision-making, using each non leaf node (i.e. the connecting line in the graph) as a judgment condition and each leaf node (i.e. the circle in the graph) as a conclusion. Starting from the root node, draw a conclusion through multiple judgments. LightGBM is a gradient boosting framework based on decision tree learning algorithms. LightGBM chooses a decision tree algorithm based on histogram, which first divides accurately continuous values into discrete ”bin” and then accumulates statistics in the histogram with ”bin” as the index, then, based on the discrete values of the histogram, traverse to find the optimal segmentation point. This can simplify data expression, reduce memory usage, and avoid overfitting.

Unlike GBDT, LightGB uses a Leaf wise leaf growth strategy with depth constraints. Compared to Level wise, Leaf wise is more efficient, can reduce errors, and has higher accuracy. Level-wise splits one layer of leaves at a time, resulting in unnecessary overhead due to its indistinguishable nature. However, the Leaf-wise strategy first searches for the leaf with the highest splitting gain from all the current leaves before splitting, which has the disadvantage of overfitting, therefore, LightGBM has added a maximum depth limit on this basis to avoid overfitting.

Category features are often used in machine learning processes, and most machine learning tools require the transformation of category features into multidimensional one hot encoded features. LightGBM directly supports category features without the need for additional one hot encoding, which reduces spatial and temporal efficiency. Compared to other frameworks, LightGBM has better performance, with advantages such as high training efficiency, low memory usage, and high accuracy.

(2) LightGBM model based on attack organization features. In this project, we will transform attack organization analysis into a multi classification problem solution. The specific classification algorithm process is as follows:

Step 1: Feature Engineering. A multi-dimensional model based on threat intelligence. By analyzing the attack organization, we establish a feature library for the attack organization. The characteristics of the attack organization include attack tactics, attack population, attack purpose, attack target, etc. Extract 51 feature dimensions from 99 attack organizations.

Step 2: Input. Encode the attack sample according to each feature, with 1 for the feature included and 0 for the feature not included, and map it to the vector space. And mark each vector category with the name of the attack organization, representing the 99 types of attack organizations as 99 dimensional vectors. Assuming that a sample x belongs to attack organization 2 (the attack organization has already been encoded in advance), then the sample is classified as category [0,1,0,0...]. By analogy, label all samples with attack organization categories;

Step 3: Training. In fact, a classification tree is trained for each possible category of the sample. There are a total of 99 attack organization categories in this round, which means that each sample includes 99 trees during each training round. For example, the first tree is for the first category of sample x, with the input being $(x, 0)$, and the second tree is for the second category of sample x, with the input being $(x, 1)$... Ultimately, the predicted values $f1(x)$, $f2(x)$ for each tree for sample x can be solved... Then, using softmax to generate probability, the probability of belonging to a certain attack organization n is:

$$p_n = \frac{exp(f_n(x))}{\sum\limits_{k=1}^{99} exp(f_k(x))} \tag{2.5}$$

Step 4: Iteration. Calculate the residuals $f11(x) = 0 - p1(x)$, $f22(x) = 1 - p2(x)$, $f33(x) = 0 - p3(x)$ for each attack organization categor. Then use $(x, f11(x))$, $(x, f22(x))$, $(x, f33(x))$... As input for each category, iterate through M rounds, constructing 99 trees for each sample in each round. After training, we obtain a decision model labeled n for each category:

$$F_{nM} = \sum_{m=1}^{M} \hat{C}_{nm} I(x \epsilon R_{nm}) \tag{2.6}$$

Step 5: Determine. When an unknown attack occurs, it can be used as sample input to determine the type of attack organization. Assuming the sample to be tested is y, the probability that y belongs to a certain category c can be obtained by inputting the trained decision model as follows:

$$p_c = \frac{exp(f_c(x))}{\sum\limits_{k=1}^{99} exp(f_k(x))} \tag{2.7}$$

We will use all attack samples from 99 attack organizations containing these features as training samples, and then use the Light GBM framework to train a model that can be used for attack organization classification [15]. In this process, our main focus is on parameter tuning. Light GBM through num leaves, learning rate, max depth, feature_fraction, objective, num_Several main parameters of class are used to control and optimize the algorithm, among which max_depth, feature_fraction, bagging, and fraction can all control or assist in controlling overfitting phenomena.

2.4. Overall Design of Attack Organization Association and Analysis System.

(1) Design concept of the system. The implementation of this system is based on the analysis of threat intelligence. The first problem to be solved is the acquisition of threat intelligence, which integrates multiple sources such as threat intelligence platform intelligence data, APT reports, and internal data to collect threat intelligence and form a threat intelligence library; Process the collected multi-source heterogeneous threat intelligence data to solve the problems of low credibility and high error rate, and form a threat intelligence feature library based on a multi-dimensional threat intelligence model, thereby forming an attack organization fingerprint library; Then, use similarity algorithms with different dimensions to calculate the association of attack organizations and determine their homology; Utilize the LightGBM framework to achieve consistency determination of attack organizations and output analysis results. Finally, the attack organization association and analysis system will be implemented.

(2) System architecture. Based on the above requirements analysis, the overall architecture of the prototype system includes six modules, namely threat intelligence collection module, threat intelligence multi-dimensional analysis module, attack organization fingerprint library module, attack organization association module, attack organization analysis module, and user module. Each module has different functional designs [16].

The acquisition of threat intelligence is the foundation and support of this study. This topic selected multiple threat intelligence sources to collect data, including obtaining APT reports through crawler technology and extracting threat intelligence. This includes collecting various intelligence data from multiple threat intelligence platforms, including intelligence data accumulated in practice, and enriching intelligence by comprehensively collecting relevant information from multiple threat intelligence sources. At the same time, the sources of threat intelligence are diverse and the structure is relatively complex. In order to facilitate subsequent analysis and processing, it is necessary to screen and process the collected intelligence data, eliminate threat intelligence with low credibility or errors, and correct errors that occur during the collection process. Based on the multi-dimensional analysis module of threat intelligence, deep analysis of threat intelligence is achieved, extracting feature indicators and attack organization information from time, space, and content dimensions. Organize threat intelligence with attack organizations as the core, and form a fingerprint database of attack organizations that can be used for association and analysis of attack organizations. Based on the multi-dimensional analysis module of threat intelligence and the fingerprint database module of attack organizations, the similarity calculation is used to determine the homology of attack organizations and establish the correlation relationship between attack organizations. Using the attack organization (event) to be detected as input, consistency is determined using the LightGBM model, and the analysis results are output. User module: Implement the management of user information and permissions.

(3) The hierarchical design of the system structure. Based on the analysis of market demand and the functional positioning of the system, as well as the usage scenarios of this association and analysis system, the author designed and implemented an attack organization association and analysis system using BS architecture, which combines a three-layer architecture and MVC using a hierarchical pattern for structural design. Through request and response data operations between layers, communication between layers is achieved [17]. Within each layer, modular development is carried out according to the overall system architecture of the system. Layering modules can achieve high cohesion and low coupling, and reduce development complexity by improving code reusability. When users use the system, they initiate requests through a browser and do not directly operate background programs, ensuring the security and stability of the system. The structural hierarchy of the system is designed from top to bottom as the presentation layer, application layer, and data layer. The content of each level is introduced as follows:

Presentation layer: Refers to the human-machine interaction interface provided by the system service, including visual display and system interface. It presents data in a user-friendly manner on the page with a user orientation, and is simple and clear, in line with user habits, and convenient for user operation. On the graphical system interface, users can perform operations such as querying.

Application layer: The application layer is the center of the system's business logic processing and calculation, mainly including threat intelligence collection module, threat intelligence multi-dimensional analysis module, attack organization fingerprint library module, attack organization association module, and attack organization analysis module. It is located in the middle layer of the entire structure and serves

Table 3.1: Display of Analysis Output Function

ID	Test file name	Output Results
1	test sample	78-Urpage-52.48%

Table 3.2: User Management Function

ID	Login Name	Joined on	Is it enabled
1	admin	2021-1-1-11:11	Enabled
2	yeshun	2021-1-1-11:11	deactivated

as a connection between the presentation layer and the data layer.

Data layer: The data layer is the bottom layer of the entire structure, playing the role of the "database" decision-making and serving as the database resources of the entire system, mainly including threat intelligence data, threat intelligence dimension features, attack organization data, and user data.

3. Results and Analysis.

3.1. Testing Environment. The attack organization association and analysis system developed in this project is a WEB system with a B/S architecture, and users need to request access through a web browser when using the system [18]. The testing environment for this system is different browsers for different operating systems, such as IE browser on Windows and Google browser on Mac. From the browser side testing verification, all the service functions provided by the system backend are normal.

3.2. System functional testing. The system has been tested for functionality in the Chrome browser on Windows 10 as follows. The intelligence reading and search query function can achieve the reading of various information about attacking organizations, and achieve intuitive display of threat intelligence. The intelligence management function can achieve operations such as adding, deleting, and updating intelligence. The association analysis function can clearly display the association relationship of attack organizations after calculating the similarity of each dimension. In this section, the similarity threshold setting and attack organization input are specially designed. By outputting the name of the attack organization to be queried and the lowest similarity, the association relationship graph related to it with a similarity greater than the threshold is analyzed and output. The results can serve as the basis for determining the homology of attack organizations. Table 3.1 shows the analysis output function, where the input is the attack samples to be detected and the output is the analysis results of the attack organization. Table 3.2 shows the user management function of the system, which enables the management of administrator users and ordinary users. After testing, all functions of the system have been implemented and meet expectations [19,20].

4. Conclusion. In recent years, an increasing number of APT attacks have threatened China's cyberspace security, bringing heavy pressure to cybersecurity defense. As the saying goes, "Knowing oneself and the enemy is invincible in a hundred battles." In order to occupy a favorable position in the game of network attack and defense and make breakthroughs in deep tracing, it is essential to study and master attack organizations. The author has conducted research on the association and analysis of attack organizations based on threat intelligence, and designed and implemented a system for the association and analysis of attack organizations. The attack organization association and analysis system developed in this project is a WEB system with B/S architecture, and users need to request access through a web browser when using the system. The testing environment for this system is different browsers for different operating systems, such as IE browser on Windows and Google browser on Mac. From the browser side testing verification, all the service functions provided by the system backend are normal. At present, the analysis technology of threat intelligence is rapidly developing, and research on APT is also receiving increasing attention from domestic and foreign researchers. The author proposes a multi-dimensional analysis based on threat intelligence to study the correlation and judgment methods of attack organizations, and designs and implements an attack organization correlation and judgment system, which to some extent solves the problem of determining the homology and consistency of attack organizations, and can provide research ideas and guidance for security analysts. However, there are still some issues worth further research in future learning. Although certain achievements have been made in multi-dimensional analysis based on threat intelligence, the characteristics of each dimension are still not detailed enough, and there are still some features that cannot be obtained. How to obtain more detailed content and detailed dimensional features of threat intelligence needs to be further deepened in the analysis of threat reports.

REFERENCES

[1] Hou, H. , Cao, G. , Ding, H. , Zhao, C. , & Wang, A. . (2021). Research on automatic detection system of encoder accuracy based on pid algorithm. Journal of Physics: Conference Series, 1754(1), 012233-.

[2] Onyema, E. M. , Dalal, S. , Romero, Carlos Andrés Tavera, Seth, B. , Young, P. , & Wajid, M. A. . (2022). Design of intrusion detection system based on cyborg intelligence for security of cloud network traffic of smart cities. Journal of Cloud Computing, 11(1), 1-20.

[3] Olukoya, O. . (2021). Distilling blockchain requirements for digital investigation platforms. Journal of Information Security and Applications, 62(1), 102969.

[4] Liu, Y. , Hou, Z. , Cui, J. , & You, K. . (2021). Design and research of automatic fishing machine based on acoustic adjustment. Journal of Physics: Conference Series, 1802(2), 022052 (6pp).

[5] Liu, X. , Zhu, S. , Yang, F. , & Liang, S. . (2022). Research on unsupervised anomaly data detection method based on improved automatic encoder and gaussian mixture model. Journal of Cloud Computing, 11(1), 1-16.

[6] Li, G. , Zhai, J. , Luo, C. , & Li, A. . (2021). Retraction note: research and application of automatic monitoring system for tunnel-surrounding rock measurement based on gis. Arabian Journal of Geosciences, 14(24), 1-1.

[7] Lin, S. . (2021). Research on automatic inspection system of printed circuit board based on computer vision. Journal of Physics Conference Series, 1861(1), 012093.

[8] Zhang, L. H. , Liang, Y. , Tang, Y. , Wang, S. , Tang, C. , & Liu, C. . (2021). Research on unknown threat detection method of information system based on deep learning. Journal of Physics: Conference Series, 1883(1), 012107 (6pp).

[9] Gao, W. , Tang, J. , & Wang, T. . (2021). An object detection research method based on carla simulation. Journal of Physics: Conference Series, 1948(1), 012163-.

[10] Dang, L. . (2021). Research on landscape design assistant system based on artificial intelligence and information technology. Journal of Physics Conference Series, 1744(2), 022103.

[11] Bagga, P. J. , Makhesana, M. A. , Bhavsar, D. L. , Joshi, J. , Jain, K. , & Patel, K. M. , et al. (2022). Experimental investigation of different nn approaches for tool wear prediction based on vision system in turning of aisi 1045 steel. International Journal on Interactive Design and Manufacturing (IJIDeM), 17(5), 2565-2582.

[12] Zheng, S. , Li, J. , Chen, S. , Liang, Y. , & Lin, J. . (2021). Research on breakpoint area detection of computer communication network transmission data based on cloud framework. Journal of Physics: Conference Series, 2083(4), 042045-.

[13] Pan, A. , & Wang, N. . (2021). Design and implementation of crop automatic diagnosis and treatment system based on internet of things. Journal of Physics: Conference Series, 1883(1), 012062 (6pp).

[14] Zong, Y. , Zhao, X. , & Ba, Z. . (2021). Design and research on the fatigue detection system of ship bridge duty based on image processing. Journal of Physics: Conference Series, 2131(3), 032119-.

[15] Liao, X. , & Xie, J. . (2021). Research on network intrusion detection method based on deep learning algorithm. Journal of Physics: Conference Series, 1982(1), 012121-.

[16] Belousov, K. I. , Bashirov, R. K. , Zelianskaia, N. L. , Labutin, I. A. , Ryabinin, K. V. , & Chumakov, R. V. . (2023). Profiling of conceptual systems based on a complex of methods of psychosemantics and machine learning. Automatic Documentation and Mathematical Linguistics, 57(4), 193-205.

[17] Li, Y. , Li, F. , & Song, J. . (2021). The research of random forest intrusion detection model based on optimization in internet of vehicles. Journal of Physics Conference Series, 1757(1), 012149.

[18] Luo, M. , Ke, Q. , & Li, J. . (2021). Research on automatic braking and traction control of high-speed train based on neural network. Journal of Physics: Conference Series, 1952(3), 032048-.

[19] Zhang, B. , Bai, L. , & Chen, X. . (2021). Research on the design of fire alarm and pre-treatment robot system. Journal of Physics: Conference Series, 1865(4), 042106-.

[20] Lin, T. , Zhao, Y. , Zhang, H. , Li, G. , & Zhang, J. . (2021). Research on information security system of ship platform based on cloud computing. Journal of Physics: Conference Series, 1802(4), 042032 (7pp).

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Nov 1, 2023
Accepted: Nov 8, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1882–1893, DOI 10.12694/scpe.v25i3.2782

APPLICATION OF INTELLIGENT ALGORITHMS AND BIG DATA ANALYSIS IN FILM AND TELEVISION CREATION

WEIWEI WU*

Abstract. With the rapid development of social media, people can access a large amount of data in a short period of time, and big data technology has emerged. With the vigorous development of cloud computing and big data, the method of mining audience interests through a large amount of data to guide film and television creation has attracted more and more attention from experts and scholars. In order to understand the current situation of film and television drama creation in China and provide suggestions for the shortcomings in the industry, this article mainly analyzes the application of intelligent algorithms for traffic prediction models and big data analysis in film and television creation. This intelligent algorithm can predict the potential audience of movies or TV dramas, helping producers and investors make decisions. This system utilizes artificial intelligence technology to select suitable actors for characters based on their matching degree and past work performance. This article applies intelligent algorithms to big data processing to improve the accuracy of data processing. This article explores the application of intelligent algorithms and big data analysis in film and television creation. Using machine learning algorithms to predict the potential audience of a movie or TV series based on historical data, providing decision-making basis for investors and producers. This system utilizes AI technology to select suitable actors for characters based on their matching degree with characters and past work performance, improving the scientific and accurate selection of roles. The application of these technologies helps to improve production efficiency and quality, reduce costs and risks, and inject new impetus into the sustainable development of the film and television industry.

Key words: Intelligent Algorithms; Big Data; Analysis; Film and TV Creation

1. Introduction. Big data, or huge amount of data, refers to the amount of data involved that is too large to be captured, managed, processed and sorted into information that can help enterprises make more positive business decisions in a reasonable time through traditional software tools [1]. It is a feature of The Times and a synonym for openness, integration and development [2]. The concept of ”big data” is widely recognized mainly because of the political TV series ”House of Cards” produced by Netflix in the United States in 2013. The popularity of the series has made people realize the value of massive user data analysis in film and television creation [3]. Film and television industry itself is the industry of producing and disseminating information. In the era of big data, film and television plays have great potential: more abundant information sources, better grasp the value and role of information through quantitative analysis of massive data [4]; more accurate understanding of the needs of audiences, and customization of the content and function of works based on the needs of audiences. Communication strategy to achieve accurate communication [5]. Big data is smart enough to make massive data worthwhile—content extraction, sharing, and interaction, so that they can better serve users and tap the value of commercial innovation [6]. In the media industry, big data analysis is deepening into the creative part of the film. This will have a profound impact on the choice of film and television from the choice of script, to the choice of directors and actors, to shooting and post-production and even marketing [7].Big data and intelligent algorithms have become the two driving forces of modern society. Especially in the film and television industry, the combination of these two technologies is revolutionizing the modes of creation, production, marketing, and playback. Big data has not only changed the way the film and television industry operates, but also brought unprecedented enormous value to this industry. Machine learning algorithms can be used to predict the potential audience of a movie or TV series based on historical data, providing decision-making basis for investors and producers. This system utilizes AI technology to select suitable actors for characters based on their matching degree with characters and past work performance. This largely solves the drawbacks of relying on perception and experience to select angles in the past, and improves the scientificity and accuracy of angle selection. Through intelligent algorithms, movies can be automatically

*Zibo Vocational Institute, Zibo, Shandong, 255300, China (Corresponding author, 10497@zbvc.edu.cn)

edited based on audience preferences and market expectations derived from big data analysis, optimizing their rhythm and structure, and enhancing their attractiveness.

Film and television big data refers to the mass data information generated by the network as the information platform in the creation, dissemination and acceptance of film and television works, and the general name of the system for storing, processing and displaying such information [8]. It mainly includes three aspects of user big data, content big data, and channel big data [9]. User big data refers to information about viewers who watch movies and television works online or offline [10]. By collecting and deeply analyzing the user data and mining its internal relationship, we can have a deeper insight into the personal and group information related to the user's experience of watching movies and TV, including the preferences of watching movies, etc. [11]. Content big data refers to the relevant information of film and television works stored in digital form [12]. Channel, as an important bridge connecting the creative works of film and television and the acceptance of users of film and television works, has always occupied an important position in the market [13]. In a certain sense, we should prevent and handle the network ideological risk of the youth. Ideological security is a necessary condition for China's overall national security and an important guarantee for the great rejuvenation of the Chinese nation. With the rise of the Internet, the integration of media has broken down the barriers between traditional media, network and Taiwan linkage, multi-screen interaction, greatly enriched the channels of film and television communication. The integration of channels brings about a great increase in data volume and continuous innovation in the form of film and television programs [14]. Big data may bring profound influence and great value to the film and television industry, which has been widely recognized by media workers.

Through the high-tech industry, big data can be comprehensively collected, mined, sorted out, summarized and refined, so as to achieve unprecedented creative results in the field of film and television creation [15]. As early as 2004, the United Kingdom had established a data analysis company called Epagogix. Through the semantic analysis of the screenplay, he built a model to evaluate the future box office. In 2013, some scholars in the film and television industry have recognized the importance of data analysis for a long time. But when the data grows to a certain extent, it is difficult for people to understand the mystery contained in the data by manual or traditional means of data analysis. In this case, new receipts must be provided. Analytical method [16]. In 2015, some scholars pointed out that data has penetrated into all business functions of the film industry at this stage. In the process of film production, information provided by big data is needed to guide all links, such as conception, production, marketing and so on. People's use of massive data will lead to a new wave of growth in the film industry [17]. In 2015, some scholars pointed out that big data analysis of film and television operation has three most important characteristics. Secondly, there are many types of data, which need to be sorted after data filtering, so that the context of data can be clearer. The third is that the speed of analysis must be fast. Now film investors all require to produce high-quality films in the fastest time. If the speed of data analysis is too slow, obviously better and more beneficial benefits cannot be obtained [18]. In 2017, some scholars realized the layered mixed interest model based on neural network and applied the model to the big data analysis of movies. The system input the original data of user behavior into the neural network and output the movies that users may be most interested in after continuous learning and training [19].

2. Materials and Methods. Big data is the core of prediction, data is not to solve the problem of ”why”, but to focus on and solve the ”what” and ”to do” problem, nowadays, the rapid development of the film industry has aroused the concern of the society from all walks of life people and big data using its own characteristics, has a profound internal [20] deep into the film and television industry. In this era of ”we media”, everyone has the right to speak, and everyone can change into the initiator of information, especially for Internet marketing. The speed of information transmission on the Internet is extremely fast, and the loss of the right to start is bound to lose a lot of attention and weaken attention. In the era of big data, we can get the focus of public attention by sorting out and analyzing big data. Big data is a data accumulation of network user behavior in the Internet era [21]. With the rapid development of new social media, online ticket sales and review websites, China's film and television industry has accumulated a lot of effective and useful data. For example, the user search index such as baidu index, micro blog index, micro index; Video websites are similar to iQIYI, youku and tencent video. Film review websites are similar to douban and mtime. New movie ticket purchasing and movie review websites such as maoyan and ticketing have accumulated a large number of users and provided a lot of usable data for the development of film and television industry. Visual data such as playback volume and geographical distribution can be obtained on video websites; daily box office can be obtained on Cat's Eye and Time Network; and a large number of film reviews can be obtained in comparison on Douban and other websites. At present, the planning of film and television content often relies on the subjective industry experience and professional sensitivity of practitioners. There are often problems such as inadequate grasp of current hot spots and inaccurate target audience. However, iQiyi Company makes full use of the big data analysis algorithm. It launched the "Green Mirror" video editing function, which is obtained by filtering, sorting and analyzing the massive user viewing information collected in the iQiyi PPS dual-brand back-end database. The user watches the video, especially the behavior habits when watching the video for a long time, and derives the user's viewing behavior data from it.

"Big data" shows its unique advantages in all walks of life. People are using it more and more. Big data also has a great impact on the film and television industry. Big data is analyzed rapidly with massive data scale, rapid data flow and various data types, which not only saves market costs, but also ensures the accuracy of data to the maximum extent, meets the different needs of different audiences, creates films and TV works with high commercial value, and achieves win-win situation [22]. Through big data analysis, it can also be concluded that the viewing habits of audiences have changed from big dramas and formal dramas to live dramas and idol dramas. As a result, the concept of screenwriters of films and TV plays has gradually changed and the theme of their works has also changed. In the 1990s, American blockbusters entered the Chinese market. With their gorgeous pictures and strong visual impact, people once again went to the cinema to watch films. Hollywood-style film and television forms also have a great impact on the literary creation of film and television creators. By using big data analysis, we can know which form of film and television is more acceptable to the audience [23]. In the era of big data in the Internet, big data has played a positive role in film and television creation, providing a way for the film and television creation industry to use information efficiently. It also created a batch of film and television works with rough content and single form, which seriously affected the art of film and television works. Value has had a huge negative impact on film and television creation. Therefore, it is more meaningful to use intelligent algorithms to screen TV dramas.

Structured data is relatively easy to draw analysis conclusions through mathematical tools, while a large number of unstructured data is the key and difficult point of big data analysis. Unstructured data refers to data that cannot be stored and read in a unified structure and format, typically including text, images, audio, video, and other types. Unstructured data has diversity and complexity, making it relatively difficult to process and analyze these data. Due to the diversity and complexity of unstructured data, its processing and analysis require the use of various technologies and tools, such as natural language processing, image processing, audio processing, video processing, as well as algorithms such as machine learning and deep learning. By analyzing and mining unstructured data, more information and knowledge can be obtained, providing more valuable support for decision-making. The emergence of data mining can solve the key problems of how to discover the important information hidden behind the data and analyze it at a higher level so as to make better use of the data. Data mining is a computational process to discover the laws of large data, and it is an interdisciplinary sub-field of computer science. The overall goal of data mining is to extract information from data sets and convert it into understandable structures for further use. Big data provides quantifiable indicators for film and television, and it is no longer a simple experience or feedback from traditional channels as in the past.

This article uses machine learning algorithms to predict the potential audience of a movie or TV series based on historical data, providing decision-making basis for investors and producers. This system utilizes AI technology to select suitable actors for characters based on their matching degree with characters and past work performance, improving the scientific and accurate selection of roles. Based on big data analysis of audience preferences and market expectations, automatically edit movies, optimize their rhythm and structure, and enhance their attractiveness. Suppose there are W search users and there are d big data retrieval resources. If the number of big data resources required by the t-th search user is dt, then the amount of big data resources that the user needs to retrieve is Wj, then the following formula can be established.

$$W_j = d_j / \sum_{j=1}^{m} d_j \tag{2.1}$$

Then, the film and television works are coded to obtain the distribution of retrieval purposes on large data resources. A sequence of retrieval purposes numbered by large data retrieval resources can be generated:

$$A^{\mathsf{T}} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \dots & \dots & \dots & \dots \\ 1 & 1 & \dots & 1 \\ y_1 & y_2 & \dots & y_p \\ y_1 & y_2 & \dots & y_p \\ \dots & \dots & \dots & \dots \\ y_1 & y_2 & \dots & y_p \\ x_1 & x_2 & \dots & x_p \\ x_1 & x_2 & \dots & x_p \\ \dots & \dots & \dots & \dots \\ x_1 & x_2 & \dots & x_p \end{bmatrix} \tag{2.2}$$

$$B^{\mathsf{T}} = \begin{bmatrix} d_1^{(1,j)} & d_2^{(1,j)} & \dots & d_p^{(1,j)} \\ d_1^{(2,j)} & d_2^{(2,j)} & \dots & d_p^{(2,j)} \\ \dots & \dots & \dots & \dots \\ d_1^{(n,j)} & d_2^{(n,j)} & \dots & d_p^{(n,j)} \end{bmatrix} \tag{2.3}$$

The complexity of decoding sequences can affect processing and storage requirements. This article needs to consider some indicators, such as the complexity of decoding algorithms, data compression rate, etc. In the actual process of big data retrieval, parallel processing and optimization techniques are usually used to improve performance. For example, using MapReduce or Spark's parallel processing framework can accelerate processing speed. According to the decoded sequence and the ETC matrix, the time consumed for large data retrieval on different large data resources is calculated. The ETC matrix represents the time required for the operation of the first retrieval purpose on the m large data retrieval resources. The time required to complete all the retrieval tasks can be calculated by the following formula:

$$P = \frac{\sigma_i^2}{\Sigma_{i=1}^{m} \sigma_i^2} \tag{2.4}$$

The decoded sequence and ETC matrix can be used to calculate the time required to complete the h-th retrieval purpose. The addition formula is shown as follows:

$$P_h = \frac{\Sigma_{i=1}^{h} \sigma_i^2}{\Sigma_{i=1}^{m} \sigma_i^2} \tag{2.5}$$

The average time taken to implement user retrieval can be calculated using the following formula:

$$t_{max} = \frac{1}{\lambda} In(\frac{1}{i_0} - 1) \tag{2.6}$$

The purpose of big data search is to retrieve useful information in a short time. At the same time, the average time required to complete the search task should also be considered. Therefore, two fitness functions are defined:

$$S_j = \Sigma_{i=1}^{N} W_{ij} X_i \tag{2.7}$$

$$P_\lambda = \frac{(M_p \bullet Q_{cal} + A_p)}{Sin(\theta_{SE})} \tag{2.8}$$

Table 2.1: Big Data Search Engine Results

Parameter	Traditional method search	Intelligent algorithm search
Search time/ms	94	35
Search accuracy%	61%	89%

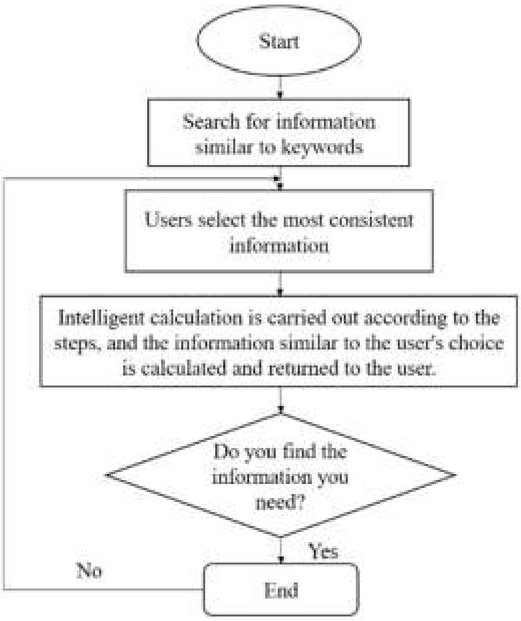

Start
Search for information similar to keywords
Users select the most consistent information
Intelligent calculation is carried out according to the steps, and the information similar to the user's choice is calculated and returned to the user.
Do you find the information you need?
Yes
No
End

Fig. 2.1: Search Engine Search Process

In the formula, Xi denotes the jth largest data resource in the ith individual, and wijxi denotes the time required for the ith individual to complete the search task. The results of valid data can be roughly shown in Table 2.1.

From Table 2.1, we can see that using intelligent algorithm search can improve the efficiency of large data search, less time-consuming than traditional methods of large data search, and higher accuracy. The search engine with intelligent algorithm can enable users to actively participate in the search process, and the search results are the closest to the user needs, so as to ensure the accuracy and satisfaction of the search results. Its working flow chart is shown in figure 2.1. Intelligent algorithm can effectively improve the search efficiency of users. The corresponding process is shown as follows:

Step 1: According to the user's search needs, the resources related to the keywords entered by the user are obtained from the large database and displayed to the user as an initialization population.

Step 2: Search users according to their own needs, from the information obtained to select the most useful.

Step 3: The big data retrieval system takes the information with the highest evaluation value as the best, and uses the intelligent algorithm to perform the calculation. Through the crossover and mutation operation, the information close to the user's demand can be obtained, and the information is fed back to the user.

Step 4: Terminate the algorithm when the information fed back to the user meets the requirements; otherwise, return to step 2.

3. Result Analysis and Discussion. A large amount of data is generated on various websites, social media platforms, forums, and other internet platforms, which can be used for big data analysis. For example,

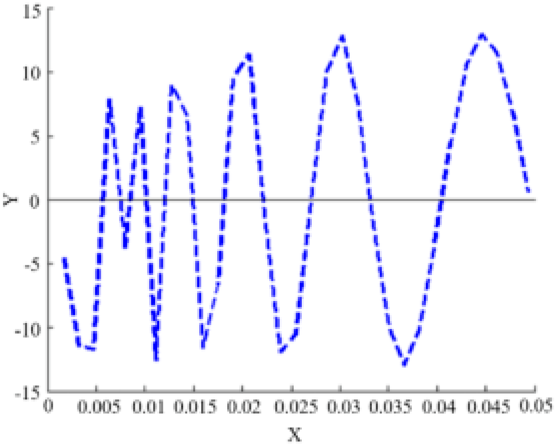


Fig. 3.1: Intelligent Algorithm Optimization Problem

search engines such as Google and Baidu can collect a large amount of user search data, while e-commerce platforms such as Amazon and Taobao can collect user shopping data. In the era of big data, the idea that "everything can be quantified" has begun to penetrate the hearts of the people. Big data provides the basis for quantitative analysis of film and television. Successful film and television works have some repetitive and referable features in terms of material selection, content, and form. Grasping these characteristics can provide guidance and support for the creation of film and television works [24]. We can also use the data to analyze the box office of the film and sort out the factors affecting the box office accurately.

Intelligent algorithm is a search algorithm. Search algorithm is essentially a process of finding the best advantage in a high dimensional space. In figure 3.1, for example, to find the maximum value of this random curve, we need to check the values of different abscissa as much as possible. However, there are countless points in the continuous curve. It is impossible to test each point, so it can only be tested with a certain precision. This is what intelligent algorithms mean: find a suitable path and let the algorithm converge to the target point as quickly as possible.

In films and TV plays, the whole group can be decomposed into a series of independent and interconnected action groups. In the planning stage, by using big data analysis technology, in-depth exploration and analysis of market trends, audience preferences, and topic popularity can be conducted, providing producers with more accurate decision-making basis. At the same time, intelligent algorithms can also provide inspiration and direction for script creation based on these data, making the story more in line with market demand and audience taste. Secondly, during the casting stage, through the application of intelligent algorithms such as the "Yihui" intelligent casting system, suitable actors can be selected for the character based on their matching degree with the character and past performance. This not only improves the scientific and accurate selection of roles, but also provides guarantees for the integration between actors and characters. In addition, intelligent algorithms and big data analysis can also play an important role in the filming and production stages. For example, using AI technology for automated editing and special effects production can improve production efficiency and quality. Meanwhile, through big data analysis technology, real-time monitoring and analysis of the environment, shooting equipment, shooting effects, etc. at the shooting site can be carried out, providing more accurate data support and guidance for shooting. The interaction between them is determined by the potential energy function which varies with the group. Assuming that each interacting group is independent and does not receive the influence of other groups, its expression is as follows:

$$I(X,Y) = \Sigma_{y\in Y}\Sigma_{x\in X}p(x,y)log(\frac{p(x,y)}{p_1(x)p_2(y)}) \tag{3.1}$$

where p is the interaction between the film and television groups. The form of interaction between the simplified

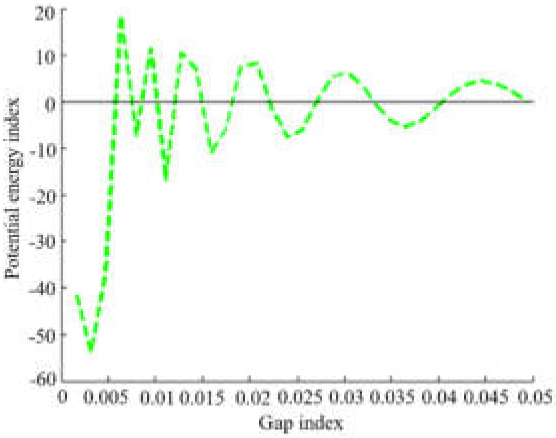


Fig. 3.2: Potential Energy Curve

film and television groups can be expressed by the following formula:

$$P_R = \frac{P(t+1) - P(t)}{P_N} \tag{3.2}$$

where P represents the potential energy between two groups, t is the difference between them, R and N are two empirical parameters. The following graph can be obtained from equation 9, as shown in figure 3.2.

As shown in the figure 3.2, when P is greater than 0, it means that the two are mutually exclusive. When they are less than 0, they attract each other. When the difference between the two is small, they are mutually exclusive. As the gap increases, the repulsive force gradually decreases. The above figure can better describe the overall characteristics of the interaction between film and television groups, but the detailed description of the interaction between group inspection is not enough. Therefore, the following formula is used to describe the interaction in detail:

$$P_i = \frac{f_i}{\Sigma_{i=1}^{N} f_i} \tag{3.3}$$

Among them, P represents the potential energy between two groups, f represents the gap between two groups, and N, I are two empirical parameters. The curve can be shown in Figure 3.3.

As shown in the figure 3.3, the force between the two can not be accurately described at a very small gap, but it is closer to the reality. In general, intelligent algorithms will adopt some forms of root mean square error for moderate evaluation. The fitness function used in this paper is shown as follows:

$$dF_r = \tau b dx \tag{3.4}$$

$$dF_r = 2b \int_0^L \tau dx \tag{3.5}$$

where b is the value to be formulated, x is the different output of each value to be fitted when the input values are different, and d is the number of values to be fitted. The fitness optimized by the intelligent algorithm varies with the number of iterations as shown in figure. The fitness value of the intelligent algorithm is compared with that of the traditional method as shown in Table 3.1, and the search range is compared with that of the traditional method as shown in figure 3.4. From the chart, we can see that in the film and television industry, the optimization effect of intelligent algorithm is better.

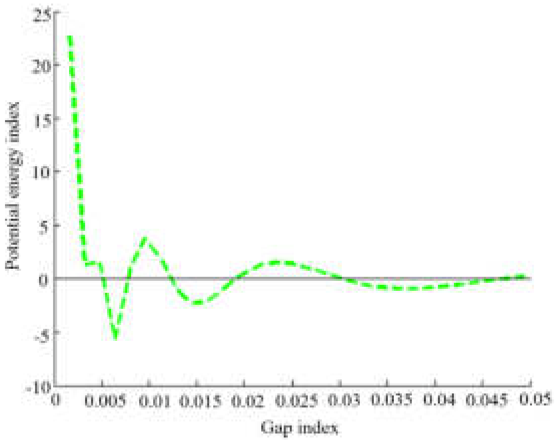


Fig. 3.3: Potential Energy Curve

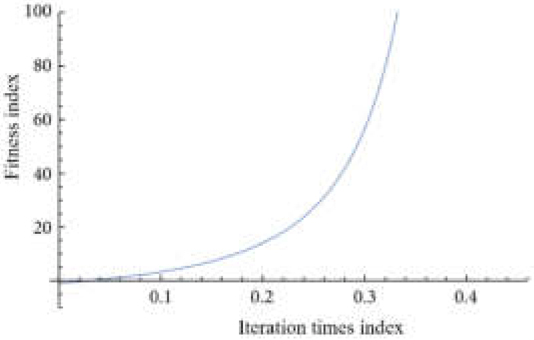


Fig. 3.4: Intelligent Algorithm Optimization Renderings

It can be seen from Table 3.2 that the potential energy function parameters optimized by the intelligent algorithm are closer to the actual one; and the search range is broader and more effective. This paper analyzes intelligent algorithm search and traditional method search through computer big data. The results can be represented by Tables 3.3 and 3.4, Figures 3.5 and Figure 3.6.

In Figure 3.3, the control group method may refer to a design of a controlled experiment used to compare the performance of intelligent algorithms and big data analysis methods with traditional methods. The control group method is an experimental design strategy that randomly divides subjects into two groups, one group accepting new methods such as intelligent algorithms and big data analysis, and the other group accepting traditional methods to evaluate the superiority of the new method. As can be seen from the chart, the search execution time of intelligent algorithm in large data environment is less than that of traditional method, and the effect is better.

In the actual creation process of films and TV plays, cloud storage and cloud computing are also needed to analyze the data to guide the selection of themes, scripts, actors, directors and later marketing. The data model is used to calculate multiple schemes, and then the optimization is selected to ensure the combination's efficiency and economy [25]. Before the creation of film and television dramas, relevant masters can use the big data technology to collect, deepen and analyze the data on the subject direction, delivery platform, target audience and other factors, and then draw whether the subject matter is attractive enough, whether the platform traffic is Reasonable and critical issues such as the main online behavior habits, and based on this, conduct correct

Table 3.1: Comparison of Film and Television Parameters

	Intelligent algorithm	Traditional method
First principles	3.7216	5.0231

Table 3.2: Scope of Film and Television Parameters Search

	Intelligent algorithm	Traditional method
Minimum value	0.1	1
Maximum value	100.1	85.9

market evaluation and provide decision support for the follow-up film and television drama creation practice. At the same time, before the official start of the shoot, the use of big data technology to explore the market positioning, distribution channels, post-production and other data mining and analysis of film and television drama, can effectively improve the market visibility and competitiveness of film and television drama under the premise of ensuring the quality of shooting. In the filming team building, using big data technology, the ability, experience, achievement and influence of team members such as scheduled directors, execution, late stage and actors are thoroughly evaluated, and then the team members are timely adjusted according to the objective evaluation results, so as to realize the optimization of the structure of creative personnel. In the preparatory period of film and television drama creation, the use of big data technology to accurately calculate the total number of shooting, different actors' parts, etc., so as to ensure the scientificity of the actors' shooting schedule, enable actors to better plan the shooting time, and ensure the high efficiency of film and television drama creation. At the same time, the application of big data technology in the field of preliminary preparation can comprehensively reduce the shooting cost, shorten the creation cycle in the scientific shooting plan, and maximize the investment and benefit ratio of film and television drama creation. To sum up, big data analysis plays a role as a platform for film and television production and investment advice. To some extent, it can provide a more rational expectation of the market and calculate the possible rate of return on investment with accurate quantitative figures.

By collecting and analyzing audience evaluations and feedback, the quality and popularity of works can be quantified. For example, indicators such as average rating and positive feedback can be used to measure the reputation of a work. For TV or online videos, ratings or views are commonly used indicators to measure the popularity of a work. High viewership or viewing frequency means that more viewers are interested in the work. The era of big data brings not only opportunities but also great challenges to film and television creation. We need to look at the problem from a scientific and rational point of view. In the creation of movies and TV plays, using big data technology to understand the data related to the audience, although it has the advantages of intellectualization and objectivity, it also has the drawbacks of inadequate optimization. China's Weibo, Weixin, Baidu, Taobao and other websites have developed rapidly, and some video websites have also developed rapidly. However, few of these websites can be used to analyze the big data of film and television creation. Although the big data contains a large amount of data, the actual usable data is far from meeting this standard. Not only do we need enough data storage space, but we also need sophisticated techniques to analyze, process, and reorganize data. At the same time, in the current situation of big data film and television creation, its data lacks authenticity and there is a lot of water. Therefore, we need to objectively analyze the content value of information. China still lacks a truly professional data analyst who specializes in big data. Data analysts not only need to fully understand the data content of related fields, but also need to understand the direction of future data needs. From the perspective of big data technology industry at the present stage, all enterprises engaged in big data work are non-standard and unprofessional, and current data analysts do not have enough data analysis ability, nor have they established rigorous thinking of data analysis. In the film and television creation industry, data analysts need to develop a complete data framework. First, they need to sort out the data sources, then establish a chart or table for daily monitoring, find out the core data, communicate with film and television creators, and understand the focus of today's film and television creation. Only by fully

Table 3.3: Intelligent Algorithms Search Running Time Changes

Kernel number	1	2	4	6	8	10	12
Running time	76153	56419	34672	10294	8026	6113	5716

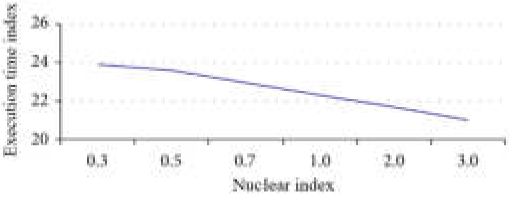


Fig. 3.5: Intelligent Algorithm Search Running Time Change

understanding the needs of the audience can scientific data analysis be carried out, so as to provide effective decision-making support.

In order to ensure the accuracy of intelligent algorithms and big data analysis, the author should establish a mechanism for evaluating data quality and accuracy. This includes evaluating the reliability of data sources, the effectiveness of data cleaning and preprocessing, and the accuracy of algorithm models. At the same time, it is also necessary to consider how to handle possible inaccurate results to avoid negative impacts on artistic creation. Faced with the opportunities and challenges of film and television drama creation in the context of big data, we must take effective measures to optimize and adjust, seize opportunities, meet challenges, and explore the path of change in film and television drama creation. First of all, we must pay attention to data sharing and optimize the data structure. Open data access interfaces to improve the sharing of data sources can effectively enhance the authenticity and accuracy of big data analysis. Secondly, we should cultivate strategic vision and increase talent reserve. In the context of the era of big data, decision-makers of film and television drama creation should have keen market insight and strong execution ability, so as to ensure the optimization and integration of resources and reach a high level of consensus at the strategic level.

4. Conclusions. With the advent of the era of big data, the research of big data processing has become a new social hot spot, and the role of big data is more precise and clear. It will play an important role in the early, middle and late stages of the creation of film and television dramas. It will also play an important role in the creation of scripts, the construction of team, the positioning of audiences and marketing. Aspects have an effective guiding function. In this paper, the problem of optimization in big data processing of film and television is studied by using intelligent algorithm. Compared with other forms of media art, film and television art in the blending zone of art and science has a natural dependence on new technology. It is in the era of digital media information that big data of film and television, as a core technology for the acquisition and application of film and television information, is bound to have a profound impact on the future creation and development of film and television, and become the fundamental of interactive experience and quantitative cognition of film and television. Through big data technology, strategic support and auxiliary operation can be provided for market insight, project incubation investment, marketing and distribution, cinema operation, so as to minimize the risk of film and television investment, improve the quality of film and television works, provide accurate positioning for film and television stars, and promote the prosperity of domestic cultural undertakings. The promotion of big data to promote the development of the film industry is also in line with the trend of the contemporary media industry. The film and television creation is more and more in line with modern advanced technology. The film and television creation gradually moves from single to plural, from closed to interactive, which is also the international trend of China's film industry. An important threshold for the development of China's film industry is an important ”milestone”.

Table 3.4: Traditional Methods to Search for Running Time Changes

Kernel number	1	2	4	6	8	10	12
Running time	91384	77261	53084	30021	10264	9546	7412

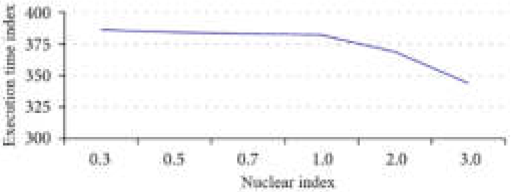


Fig. 3.6: Changes in Search Time Running of Traditional Method

REFERENCES

[1] Ebrahimi, P., Salamzadeh, A., Soleimani, M., et al. *Startups and consumer purchase behavior: Application of support vector machine algorithm. Big Data and Cognitive Computing, 6(2): 34, 2022.*

[2] Tang B, Chen Z, Hefferman G, et al. *Incorporating Intelligence in Fog Computing for Big Data Analysis in Smart Cities. IEEE Transactions on Industrial Informatics, 13(5): 2140-2150, 2017.*

[3] Somandepalli, K., Guha, T., Martinez, V. R., et al. *Computational media intelligence: Human-centered machine analysis of media. Proceedings of the IEEE, 109(5): 891-910, 2021.*

[4] Ajah, I. A., Nweke, H. F. *Big data and business analytics: Trends, platforms, success factors and applications. Big Data and Cognitive Computing, 3(2): 32, 2019.*

[5] Chow, P. S. *Ghost in the (Hollywood) machine: Emergent applications of artificial intelligence in the film industry. NECSUS_European Journal of Media Studies, 9(1): 193-214, 2019.*

[6] Wisetsri, W., *Systematic analysis and future research directions in artificial intelligence for marketing. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12(11): 43-55, 2021.*

[7] Kusal, S., Patil S, Kotecha K, et al. *AI based emotion detection for textual big data: techniques and contribution. Big Data and Cognitive Computing, 5(3): 43, 2021.*

[8] Awan, M. J, Khan R A, Nobanee H, et al. *A recommendation engine for predicting movie ratings using a big data approach. Electronics, 10(10): 1215, 2021.*

[9] Nti, I. K., Quarcoo, J. A., Aning, J., et al. *A mini-review of machine learning in big data analytics: Applications, challenges, and prospects. Big Data Mining and Analytics, 5(2): 81-97, 2022.*

[10] Hou, Q., Han, M., Cai, Z., *Survey on data analysis in social media: A practical application aspect. Big Data Mining and Analytics, 3(4): 259-279, 2020.*

[11] Kamrowska-Załuska D. *Impact of AI-based tools and urban big data analytics on the design and planning of cities. Land, 10(11): 1209, 2021.*

[12] Martin M E, Schuurman N. *Social media big data acquisition and analysis for qualitative GIScience: Challenges and opportunities. Annals of the American Association of Geographers, 110(5): 1335-1352, 2020.*

[13] Serrano, W., *Intelligent recommender system for big data applications based on the random neural network. Big Data and Cognitive Computing, 3(1): 15, 2019.*

[14] Shaddeli, A., Soleimanian Gharehchopogh F, Masdari M, et al. *An improved African vulture optimization algorithm for feature selection problems and its application of sentiment analysis on movie reviews. Big Data and Cognitive Computing, 6(4): 104, 2022.*

[15] Jagatheesaperumal, S. K., Rahouti, M., Ahmad, K., et al. *The duo of artificial intelligence and big data for industry 4.0: Applications, techniques, challenges, and future research directions. IEEE Internet of Things Journal, 9(15): 12861-12885, 2021.*

[16] Chen, K., Zu, Y., Wang, D., *Design and implementation of intelligent creation platform based on artificial intelligence technology. Journal of Computational Methods in Sciences and Engineering, 20(4): 1109-1126, 2020.*

[17] Lutz, C., *Digital inequalities in the age of artificial intelligence and big data. Human Behavior and Emerging Technologies, 2019, 1(2): 141-148, 2019.*

[18] Yanqiu, Tong, Song Y. *Large Data Based Research on the Editing of Film Trailer. Sociological Research, (1):23-28, 2015.*

[19] Bharadiya J P. *A Comparative Study of Business Intelligence and Artificial Intelligence with Big Data Analytics. American Journal of Artificial Intelligence, 7(1): 24, 2023*

[20] Chen, Y., Chen, H., Gorkhali A, et al. *Big data analytics and big data science: a survey. Journal of Management Analytics, 3(1):1-42, 2016.*

[21] Ahmed, A. A. A., Ganapathy A. *Creation of automated content with embedded artificial intelligence: a study on learning*

management system for educational entrepreneurship. Academy of Entrepreneurship Journal, 27(3): 1-10, 2021.

[22] Zhang, Y., Wilker, K., *Artificial intelligence and big data driven digital media design. Journal of Intelligent & Fuzzy Systems, 43(4): 4465-4475, 2022.*

[23] Agbehadji, I. E., Awuzie, B. O., Ngowi, A. B., et al. *Review of big data analytics, artificial intelligence and nature-inspired computing models towards accurate detection of COVID-19 pandemic cases and contact tracing. International journal of environmental research and public health, 17(15): 5330, 2020.*

[24] Hacking, K., *An Analysis of Film Granularity in Television Reproduction: Reprint of BBC Engineering Monograph. Journal of the Smpte, 73(12):1015-1029, 2015.*

[25] Li, S., Hao Z, Ding L, et al. *Research on the application of information technology of Big Data in Chinese digital library. Library Management, 40(8/9): 518-531, 2019.*

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Nov 1, 2023
Accepted: Nov 20, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1894–1903, DOI 10.12694/scpe.v25i3.2783

CONCEPT AND APPLICATION OF INDUSTRIAL HERITAGE PLANNING DECISION IN BIG DATA ERA

XIAOJUAN HAN[1,2], XUANGE ZHU[1], MIN XIE[1], HUI CHEN[1*] *

Abstract. With the rapid development of big data technology, more and more fields are realizing the importance of big data and applying it to various decision-making processes. Industrial heritage planning decisions are no exception. This article will explore the concept and application of industrial heritage planning and decision-making in the era of big data, aiming to emphasize the importance and application value of big data in industrial heritage planning and decision-making. This article analyzes the development relationship between the construction and transformation of urban industrial utilization and the theory and practice of industrial heritage. Through theoretical and practical technical analysis of specific regions, the value utilization of industrial simulation design has been improved. Through in-depth data mining, the government found that the industrial structure of the region is dominated by traditional manufacturing, lacking support from high-tech industries, and the pollution problem of the surrounding environment is also relatively serious. In response to these issues, corresponding planning plans have been formulated, including measures to introduce high-tech industries, optimize industrial structure, and improve environmental quality. By implementing these measures, the comprehensive strength of the region has been significantly improved and the economic transformation and upgrading have been successfully achieved.

Key words: Application; Industrial Heritage; Planning Decision; Big Data; Diagnosis and treatment services

1. Introduction. At the end of the nineteenth Century, Britain first appeared ”industrial archaeology”, which is different from the archaeology of common excavated relics. Although industrial archaeology includes earlier studies of the origins of pre-industrial and primitive industries, it is even more emphasized. Nearly 250 years of industrial revolution and industrial development period, Material Industrial Relics and relics are recorded and protected. The term ”industrial archaeology” means that people have begun to pay attention to the protection of industrial buildings [1,2].

Big data analysis is a powerful tool that can be used to improve decision-making, increase efficiency, optimize business processes, and drive innovation. In various fields of application, computational intelligence technology can help us better understand and solve real-world problems [3]. Post-industrial society is the further development of the industrial society of society [4]. The key variables of post industrial society are information and knowledge, and the main sectors of the economy are the third, even fourth, and fifth industries dominated by processing and service industries, such as transportation, public welfare, trade, finance, insurance, real estate, health, scientific research, and technological development [5]. Industrial heritage planning requires data support from various aspects, such as urban planning, environmental protection, tourism management, etc. Therefore, it is necessary to build a complete data application and management platform to achieve data sharing and application. At the same time, corresponding management measures need to be developed to ensure the accuracy, standardization, and security of data [6]. Typically, in the United States, Britain, Japan and other developed countries have 6 5% 75% of labour in the service industry, 30% 40% is engaged in the information service industry. 2 this represents a fundamental change in the production mode of industrialization. One is the characteristics of post-industrial society is the most intuitive, most labour no longer engaged in agriculture or manufacturing, but in the service industry, such as trade finance, insurance, transportation, entertainment, education, research, and management. It also marks the Europe and other developed countries have entered the post-industrial society [7]. In twenty-first Century, the human society on a large scale from the industrial age into the information age, from industrial society to post-industrial society, from city to city in the 3 century

*1. School of Architecture and Planning, Hunan University, Changsha, 410082, China 2. College of Architecture and Urban Planning, Hunan City University, Yiyang, 413000, China; (hanhan1380@163.com, xuangez@hnu.edu.cn, archxiemin@hnu.edu.cn, chenhui2012@hnu.edu.cn) Corresponding author: Hui Chen

". Based on the development of civilization for thousands of years, most people in the world (over 50%) will enter the city to Chinese for community life [8]. On the other hand, city is becoming the key node of the global economy on the network in the process of global integration. In respect of city development update the world is accelerating [9,10]. This also led to the rapid rise of the post-industrial society, industrial society gradually withdraw from the stage of history. This is due to the need to adjust the industrial layout of city, construction and infrastructure conditions of traditional industry lags behind the aging caused by the city industrial structure adjustment and other factors. The development of the information society and the global economic integration process directly affects the change of city the industrial structure of the third industry has become the leading industry is gradually replacing second. In the post-industrial era, information and service workers become the body of work; the city industry also began to migrate to the suburbs. The city of the old industrial base of lost economic activity and employment opportunities are resulting in industrial building renovation and reuse requirements.

The content of industrial heritage includes multiple aspects, and its core is the protection and utilization of industrial buildings, machinery, equipment, factories, mines, workshops, production lines, raw materials, ancillary facilities, surrounding sites, and technological processes with historical, technological, cultural, artistic, and economic value. Industrial heritage also includes intangible cultural heritage related to it, such as craft skills, traditional techniques, etc. These heritages, together with material cultural heritage such as industrial architectural heritage, constitute the complete content of industrial heritage. The industrial heritage consists of the legacy of industrial culture that has historical, technological, social, and cultural heritage, Architectural or scientific value. These remnants are made by buildings and machinery, workshops, and buildings Factories and factories, mines and disposal of refined sites, warehouses and storage rooms, energy production, transportation, use and transportation. The site of all underground structures, a place of social activity associated with industry, such as a house, Places of worship or educational institutions are included in the domain of industrial heritage industrial heritage is divided into broad and narrow sense of two kinds, in time, narrow sense of industrial heritage refers to Eighth Century The industrial remains of the industrial revolution that began in England [11]. The broad industrial heritage can include prehistoric processing the site of stone tools, ancient resources, smelting and smelting sites, and ancient works, including water conservancy projects Relics and remains of human technological creation reflected in various historical periods such as large-scale engineering sites [12].

Niu et al. proposed a framework for optimizing data management using big data analysis (ODM-BDA) to improve the intelligent effectiveness and decision analysis of organizations. Introducing backtracking methods in business intelligence and decision-making environments to enhance plan failure and risk-taking capabilities [13]. Novak et al. aim to explore product decision-making information systems, real-time sensor networks, and artificial intelligence driven big data analysis in Sustainable Industry 4.0. It analyzed and estimated the performance of the production network [14]. The willingness to invest in the Internet of Things (IoT) and Big Data Analysis (BDA) does not seem to depend on the supply and demand of technological innovation. The required sensing and communication technologies are mature and affordable for most organizations. On the other hand, enterprises need more operational data to address the dynamics and randomness of the supply chain [15]. Decision-making for manufacturing and maintenance operations is benefiting from the advanced sensor infrastructure of Industry 4.0, enabling the use of algorithms that analyze data, predict emerging situations, and recommend mitigating actions [16]. Operational Risk Management (ORM) is crucial for any organization, and in the era of big data, analytical tools for operational risk management are developing faster than ever before. Araz et al. examined the latest developments in academic ORM literature from the perspective of data analysis [17]. Tantalaki et al. reviewed the use of big data analysis practices in agriculture to solve various problems, revealing opportunities and promising areas of use. The large and complex amount of data generated poses a challenge to the successful implementation of precision agriculture [18]. Machine learning seems to have potential to address agricultural big data, but it needs to reshape itself to meet existing challenges. Ö zemre and Kabadurus use a large amount of open trade data to predict export volume [19]. The predicted values are included in the Boston Consulting Group (BCG) matrix for strategic market analysis. Big data analysis tools play a crucial role in establishing the knowledge required for decision-making and preventive measures [20]. Big data analysis and its application have great potential to enhance urban operations, functions, services, design, strategies, and policies in this direction. This is because big data computing enables informed decision-making and enhanced insight in the form of application intelligence [21]. In the era of digital transformation, big data has played a crucial role in changing the global tourism industry, providing significant challenges and opportunities for established companies and new entrants to the tourism industry [22]. With the significant increase in the availability of a large amount of data analysis methods, organizations are beginning to use talent analysis to manage their workforce. Nocker et al. discussed the benefits and costs of using talent analysis within organizations and emphasized the differences between talent analysis and other sub areas of business analysis [23]. Ogbuke et al. explored the application of big data in supply chain management and its benefits for organizations and society [24]. The paper also examines the ethical, security, privacy, and operational challenges of big data technology, as well as the potential damage to corporate reputation. Clinical decision-making is more promising and evidence-based, therefore, big data analysis to assist clinical decision-making has been expressed in various clinical fields. Rehman et al. presented the different architectures, advantages, and repositories of each discipline, comprehensively describing how different healthcare activities are completed in pipelines from multiple perspectives to facilitate individual patients [25].Through the big data platform, this article can collect more comprehensive and detailed data, including but not limited to medical behavior data, resource allocation data, medical service quality data, etc. These data provide more comprehensive data support for industrial heritage planning decisions, helping decision-makers better understand and grasp the situation of hospitals. Big data platforms can provide data analysis and mining functions, helping decision-makers extract valuable information from massive amounts of data, thereby achieving data-driven decision-making. For example, by analyzing medical behavior data, doctors' medical behavior characteristics and work patterns can be discovered, providing reference for hospital operation and management.

2. Overall Design of Energy Storage Braking Energy Recovery System. The energy storage braking energy recovery system is a technical solution that involves the storage and recycling of energy. This technology has important application value in the industrial field, which can effectively improve energy utilization efficiency, reduce energy consumption, and provide technical support for the planning and protection of industrial heritage. The era of big data provides more data support and information foundation for the planning of industrial heritage. Through big data platforms, a massive amount of data information can be collected, including data related to industrial heritage. These data can provide basis and support for the design and application of energy storage braking energy recovery systems, thereby better realizing the planning and protection of industrial heritage. With the change from industrial society to post-industrial society, more and more industrial buildings withdrew from the stage of history. However, each item as the industrial heritage of the industrial building is a record of the specific industrial history information, the information for the understanding of the values of industrial civilization, industrial technology, industrial level, industrial organization, industrial and cultural industry. Process, it is impossible to substitute. Based on the industrial heritage reservation or protection, can prove the historical events and historical information. Through the transfer of industrial heritage, can understand the mode of industrial society when production, relations of production development and changes. For example, we can understand the production environment and the production situation at that time from the study of industrial equipment, from the structure of plant and workshop to deduce workers working condition, Judge the social production capacity and consumption level from industrial products. Therefore, industrial heritage is a historical witness of industrial civilization and experience, to break the shackles of time, history and become the carrier of history; we convey the profound industrial civilization. As a historical basis in reality, industrial heritage in the condensation of the industrial historical value has certain universality.

2.1. The Overall Design of the System. The industrial architectural heritage has important scientific value. Site selection and planning of industrial building heritage in the production base, The construction and construction of buildings and structures, the commissioning and installation of mechanical equipment, the design of process flows and products And so on, it has scientific research value. As an important carrier of industrial civilization, the industrial architectural heritage has become a very important stage in the process of Ming Dynasty For the important part of the cultural heritage, it is important for a complete understanding, historical evolution and cultural heritage It is an important part of the cultural value of industrial architecture heritage, features and regional characteristics and humanistic spirit as shown in figure 2.1.

China's industrial land in some central area of the city is like as some city in the surrounding area. In

Fig. 2.1: Underground industrial site

Fig. 2.2: Industrial sites after renovation

recent years, the rapid development of city construction, start planning strategy, so that the original will in the central area of the city's industrial land value soared, as the Beijing Jeep factory, Beijing electronic meter factory economic. Value is an important characteristic of the industrial architectural heritage is different from other historical and cultural heritage. The city's Industrial Architectural Heritage Reuse, not only injects new vigour and power for the city, stimulating economic development, but also has a certain effect on the city heritage context. And reuse of industrial heritage can be save the cost of demolition large, but also avoid a lot of construction waste due to the demolition caused by the destruction of the natural environment, which is consistent with the sustainable development strategy Slightly. The life of industrial buildings than its use life is long, so the possibility of the transformation is also larger, can effectively avoid the waste of resources and economy. According to the different characteristics of industrial heritage, the heritage building for the transformation of the industrial museum, commercial office, leisure and entertainment facilities, creative industry Park industrial heritage tourism, industrial landscape, etc., through the adjustment and replacement of function layout makes the industrial heritage glow second spring, play the maximum value, and promote the third industry but also solve some employment problems, based on the industrial heritage protection, at the same time can be transformed into enormous economic benefits as shown in figure 2.2.

The artistic features of the industrial architectural heritage not only have the artistic characteristics and

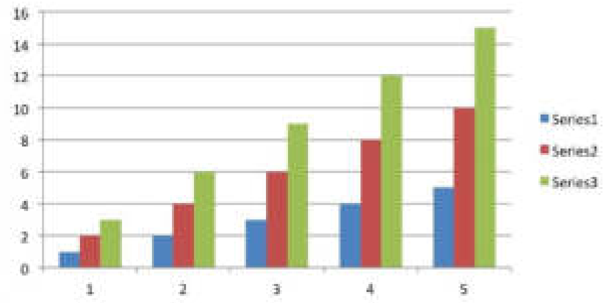


Fig. 2.3: Three series

general works of art is consistent, and it also has the characteristics of art and works of art has not. First, for the purpose of art appreciation and the industrial heritage is practical. Second, art has unique art forms such as sculpture, art etc.; and the industrial architectural heritage in industrial entities as the artistic carrier of industrial heritage. There are a lot of architectural art value of ”Bauhaus” style of modernist style buildings retain industrial architectural heritage.The artistic value of industrial architecture is reflected in three aspects: the architectural style and genre characteristics of a period, the expressive and infectious power of industrial architectural works, and the impact on urban space. As shown in Figure 3.

Industrial heritage is not only a huge material wealth, but also have enormous spiritual values. They witnessed the brilliant period of industrial development, but also a record of work life a large number of industrial workers, can be said that the industrial heritage is to industry oriented development of modern society. Because of its typical example of the era the spirit of enterprise culture and the excellent quality of workers has become an important symbol of that era, so it has important educational significance and incentives, is the historical contribution of another batch of old industrial workers of the respect and affirmation. At the same time, the industrial architectural heritage has more emotional value and a special carrier for long-term working memory the number of technical staff and workers and their families, to be properly protected will give the industrial community residents with psychological stability and return to them Sense of it.

2.2. Basic Structure of the System.. This article is very convincing from an archaeological perspective. The threat of extinction and the preservation of industrial space and the value of industrial construction have attracted the attention of British academia and civil society. These discussions prompted the British government to investigate and document. The development of industrial archaeology has promoted a deeper understanding of industrial heritage and museums through plans and related protection policies. Many industrial relics are protected in the form of people interested in traveling and sightseeing. This has also enabled industrial heritage tourism to embark on its journey and become a new form of tourism, as shown in Figure 2.4..

The industrial sample plot drawing software in Figure 2.4 may refer to AutoCAD. The relationship between the industrial template field drawing software in Figure 2.4 and the context may refer to the use of drawing software such as AutoCAD to create and display detailed design drawings and models of the industrial template field. These drawings and models can be used to guide actual construction, plan industrial heritage tourism routes, and showcase the historical and cultural value of industrial heritage. Study on this field in our country generally appear in the middle and late 1990s, mainly including the study of city waterfront renovation development city government concern directly, such projects take the ”top-down” overall mode of operation, such as many traditional city waterfront area, industrial area and warehouse land reform and especially a person with breadth of vision. The research carried out two aspects of the arts community and professional use of the transformation of traditional industrial construction concern as shown in figure 2.5.

At the same time the case and reuse of industrial heritage protection outstanding in Europe and other places like bamboo shoots after a spring rain up, such as the British Cardiff Docklands development, industrial development Glanville island in Vancouver, Switzerland, Zurich power plant renovation and reuse, reconstruc-

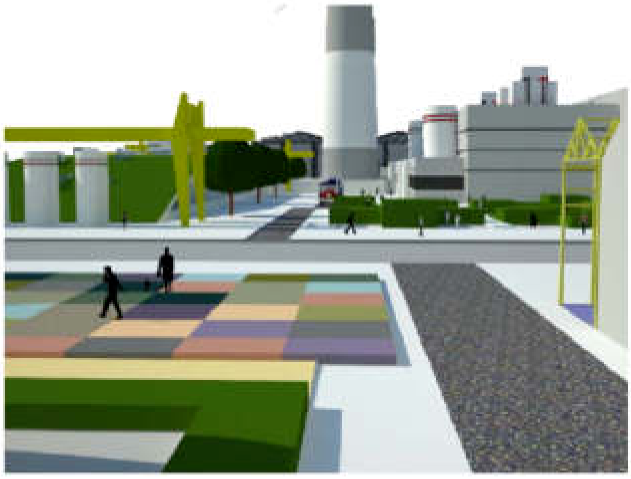


Fig. 2.4: Industrial model site I

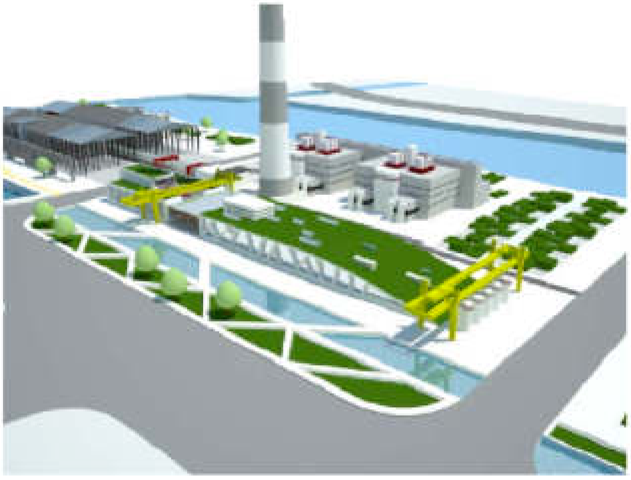


Fig. 2.5: Industrial model site II

tion and development of Sydney power plant renovation of Seattle gas, Vienna Gas Tank Museum renovation, renovation and reuse in Zurich, Switzerland Stephen Bruner mills reformation. At present, the world heritage of Europe is not only the church and other ancient buildings, including the ruins and industrial civilization, which only has three mining areas, located in Belgium, Germany and Sweden. Chinese are included in the world heritage the list of heritage projects are mostly archaeological sites, religious temples, imperial tombs and the royal garden. In 90s, due to the city the change and development of exhibition mode to speed up the construction speed, resulting in a century of industrial buildings and lots of decline. The transformation of city economy and lead to "update mode, and the transformation of the city will" retreat "is the decline of second industry industrial buildings and sites. An important object of the transformation of the old city also includes industrial buildings and the objective of the overall planning area. Shanghai and Nanjing have been developed; Shanghai and Nanjing have been developed; Shanghai will put the function replacement 66.2 square kilometres of industrial land. By 2010, 66.2 square kilometres of industrial land reserved for 1/3, 1/3 to third of industrial land, the last 1/3 transfer to the suburbs and outer suburbs.

Although our country pays more and more attention to the protection of industrial heritage, but theoretical research and practice of reform of industrial heritage at present in our country has just started, the existing research results are not enough to face a large number of industrial heritage issues put forward the best solution, still need to pass the theory and method of experience and practice to further enrich. Approved in the Sixth Batch of national key cultural relics protection units, and will continue to be a group of ancient iron, copper and other sites also included in the protection unit, and the Arsenal site, the early construction of Qingdao brewery, hydropower station, River bridge and a number of modern industrial heritage protection into list. Protection and reconstruction of industrial heritage reuse is imperative, deepen the research of industrial heritage protection and renovation has important significance in our country.

2.3. The Main Working Process of the System. The city's cultural landscape and the spirit of place are composed of different periods and different types of city buildings and sites. Industrial buildings and other construction, a witness history and reproduce the historical role of the industrial building is the industrialization of our human society experience, the same is also a witness. They are a cultural city even the national and historical material, have an indelible effect on the industrial heritage and record the history of civilization. Industrial heritage is not only because the building is "historical stone", city culture and city image as constituted by construction area is formed, it will arouse people's sense of identity and sense of existence so, the protection and renovation of industrial architectural heritage is the inheritance and protection of culture of the city. But the industrial heritage is different from other ancient architectural heritage, is the use of it. Because the long life of industrial buildings and construction level is higher, resulting in a lot of abandoned industrial buildings has very strong function. So for the protection of industrial heritage not as the legacy of ancient building protection, renovation and reuse of industrial heritage to meet the new functional requirements, is the most effective of the industrial architectural heritage protection means.

3. Overall Design of Energy Storage Braking Energy Recovery System. Sustainable development has put forward new requirements for the construction sector. The industrial heritage area of the original building and infrastructure have the possibility to continue using, compared to construction, renovation and reuse development method can reduce the cost of demolition, and reduce the environmental pollution, which is consistent with the strategy of sustainable development of mass. The development and use of it by way of discarded resources has been questioned and warned the Rome club, said such a method will lead to resource depletion, mankind will face a dilemma. China's current construction waste has accounted for the total amount of city garbage is 3 0%-40%, there is a great possibility for the use of the industry architectural heritage, demolition and reconstruction is obviously not a wise choice, increase the cost of demolition waste also increases and a large number of new waste. So, Reasonable reconstruction and reuse of industrial building heritage is of great significance to environmental protection and sustainable development.

3.1. The Main Working Process of the System. The energy shortage has become an important problem now facing the world, and as building energy consumers should bear responsibility. Usually the service life of the building than its design life is long, and many of the buildings in did not reach actually with life when they have been demolished for new buildings, is a great waste of resources and industry. The building function and require the use of the space, the building at the time of using the advanced construction technology and building materials, and it mostly has the characteristics of stable large space structure, provides a precondition and possibility for its good function replacement. And some industrial buildings because of its huge volume, the cost of demolition it will cost more than pay reform. However, the historical building renovation not all cases are cost savings, "1 970/1980's, Cheng Bending was building re-use Higher than the new cost, but this change to the end of 1980s, the building again started using competitive. 5 this is because the development of economy and technology, now the cost of a building structure is about the total cost of 1 /3, 6 transformations than the new structure can save most of the money spent. And the construction period is short, so is cost-effective for investors and owners.

The beginning of modern architecture is generally considered an industrial building Peter Behrens shoe factory. Faust industrial standardization construction, function oriented concept reflects the essence of modern architecture. Its "form follows function" and geometric aesthetics design rule until now also has a guiding role. The research on modern building concept and the method has important practical significance, has become a modern building in the fresh material. Some industrial workshop column spacing and span has made great breakthrough such as, the span and column a British ship hull workshop reached 75 meters; industrial workshop area has been greatly improved, such as the former Soviet Union coin Schiff V Its main car manufacturing plant construction area reached 740000 square meters, the Luis Weil truck factory assembly workshop area of 244000 square meters, industrial construction volume reached a considerable degree, a Soviet designed atomic power station up to 164 meters, the building number reached the peak.

3.2. The Main Working Process of the System. The arrival of the information society and the global economic integration process directly affects the industrial structure of the city. The development of information technology has made traditional manufacturing supply chain management more efficient and precise. Through technologies such as e-commerce and the Internet of Things, enterprises can grasp market demand and supply in real-time, achieving the goals of inventory management and logistics optimization. This not only reduces costs, but also enhances the competitiveness of the enterprise. With the advent of the information society, the traditional manufacturing industry gradually decline, the third industry gradually replaced the second industry to become the leading force in the industrial structure. The second industry is an industry with high resource consumption and high pollution, while the third industry is an industry with less resource consumption and less environmental pollution. With the increasing emphasis on environmental protection and sustainable resource utilization, the tertiary industry has gradually become the dominant force. Such as manufacturing, transportation and warehousing industry gradually decline, financial, trade, culture, information as the industry has become the city's main functions. In the past in the manufacturing industry developed on the basis of the industrial city of different structural decline. Also leads to many of the old industrial zone has not adapt to the industrial structure of city planning, new face relocation, demolition of fate, resulting in transformation of industrial heritage and need to use.

Due to the development of production technology, production conditions and production scale requirements increases, some of the traditional industrial area and building area has been unable to meet the modern needs. Or because of the development of the mode of transport, the original old wharf, the station conditions and transport equipment cannot meet the requirements. Such as the London Docklands, Shanghai along the Suzhou River Glanville island of Vancouver Industrial Zone, industrial zone, and Beijing industrial zone. There is due to the expansion of the city, the capacity of the city was a breakthrough, so that some areas of city overload, infrastructure and environmental conditions are relatively backward and aging, so that it cannot meet the new requirements of the city, the traditional industrial zone also there are serious problems in this issue.

At the same time, the value of the land itself and the economic benefits created is no longer equal; the land produced a new demand. The higher value of the city centre of the land is industrial land occupied, became a waste of land resources. Due to the commercial, office, finance and other three industry profits should be higher than the industry, so some of the city's land value high land becomes their location, but there are a lot of land is still industrial land occupied, so I in the industrial building large demolition and construction, and transformation of the industrial architecture reuse can take into account the rational use of industrial heritage protection and land value.

For some of the traditional city industrial zone has not adapted to the modern city life, so to reconstruction and recycle. It is a pity in China since the beginning of 1950, with the production of the transformation of the old city "by the end of 1970s after the reform and opening up the city of the old residential renovation, renovation of the city centre area and city structure requirements are the reconstruction of traditional industrial area most used, led to another batch of valuable industrial heritage was destroyed. No matter what the reasons mentioned above is ultimately proposed modification of Industrial Heritage Reuse, transformation of industrial heritage and reuse of more and more people the attention and concern, the contradiction between these with values in many aspects of the architectural heritage protection and regional new city functions, transformation and reuse is undoubtedly A balanced approach.

This reuse mode is the base of the historical and cultural industry based on the industrial historical value and ecological value of landscape architectural decisions. As is the historical and cultural places retain the base context, a protection of the ecological landscape and the traditional way of life, but also the protection of industrial heritage. When use of such the site, to conduct a thorough investigation on the whole area and construction, rigorous evaluation, the protection of places of cultural value and industrial building reuse.

Conservation and reuse of industrial heritage is reserved for industrial history appearance and its historical characteristic of the construction or expansion in the periphery, or reasonable the reform of the internal space, without affecting the industrial historical value under the most possible out of the possibility, through reuse not only make it become a witness of history, Can the new vitality to bring new benefits. Protective reuse is very popular in Europe and the United States, in China still belongs to the starting stage in the first, also like the United States s oho District, Beijing 798 factory districts by keen artist found and re-use of its protection, make this a pack of Moor house style industrial architectural heritage to be retained down.

3.3. The Main Working Process of the System. Whether foreign or domestic, the heritage renovation reuse of industrial building relatively sharp smell is artists. Like 798. Of New York's s oho and Beijing compared to the original plant users and residents, the artist's cultural acumen make them aware of the industrial heritage value. Through years of hard work in 798 rental the staff and the United artists seven Hadrian bold attempt of new management and operation mode, positive for all aspects of social support and help of the government, the joint efforts of the industrial architectural heritage building protection industry to survive. This way can be temporary or even longer preservation of industrial heritage; there is the guiding significance in use.

For the party, due to the limitation of plant uncertainties and their own economic conditions, the construction industry can only rent. So the plant transformation will not invest a lot of money, the transformation is at a relatively early stage, and the transformation of the project quality is not very high. But for the lessor to speak, rental industry the construction of the income is very low, if it is not of direct rental plant because of its poor infrastructure and supporting facilities, the rent is lower, the economic return is not high due to the lessor quality more willing to invest capital to improve the infrastructure and industrial construction, such a vicious spiral will only make the situation worse, more and more protection the industrial heritage more disadvantageous.

The timing of the first is the spontaneous reuse, flexibility, try the possibility re-use of industrial heritage will have more, according to the base of the surrounding environment, the human factor is the most suitable for reuse, or commercial or residential or exhibition, to explore the possibilities for trying to get the most reasonable results. Next, the government intervention, to guide developers to invest, before not to damage the interests of developers under the premise of spontaneous to maximize the economic benefits of the site, to optimize the environmental quality, fill the infrastructure of the field full of new vitality. This re use due to early low investment can be discussed a wide range of possibilities, the functional replacement site to achieve the most reasonable economic state.

4. Conclusion. In this paper, based on large data through reasonable means of protection and transformation of the industrial civilization heritage and respect for the industrial history, based on the realistic attitude, analysis of the domestic and foreign industrial heritage protection and excellent case reuse, reuse of industrial architectural heritage protection and the theory of learning and the analysis and summary of previous experience. Then the industrial building heritage protection and reconstruction were to conduct research and discussion, in order to make some useful supplement. Through the case design of the protection and reuse of industrial buildings of the method and mode of practice, strengthen the ability to use the theory in practice. And try new techniques and models in practice, application of rich industrial architectural heritage protection and transformation in our country in order. But because the case is not the actual design project the new approach proposed and model remains to be verified in the actual situation. Make a detailed and comprehensive description of the protection of industrial heritage and transformation, in the case of proposed method of innovation but, due to my limited knowledge and training, still need further study.

5. Acknowledgments. The study was supported by the Hunan International Innovation Cooperation Base on Science and Technology of Local Architecture, Changsha 410082, China; Hunan Provincial Key Area RD Program - Key RD (2021WK2004).

REFERENCES

[1] Dong, M., Jin, G., *Analysis on the protection and reuse of urban industrial architecture heritage, IOP Conference Series: Earth and Environmental Science. IOP Publishing, 787(1), pp. 012175, 2021.*

[2] Zhang, J., Cenci, J., Becue, V., et al. *Recent evolution of research on industrial heritage in Western Europe and China based on bibliometric analysis. Sustainability, 12(13), pp. 5348, 2020.*

[3] Ibrar, Y., Victor, C., Abdullah, G., Salimah, M., Ibrahim, A. T. H., Ejaz, A., Nor Badrul, A., Samee, U. K., *Information Fusion in Social Big Data: Foundations, State-of-the-art, Applications, Challenges, and Future Research Directions. International Journal of Information Management, 12(4), pp. 56-59, 2016.*

[4] Rahat, I., Faiyaz, D., Brian, M., Shahid, M., Usman, Y., *Big Data Analytics: computational intelligence techniques and application areas. International Journal of Information Management, 23(6), pp. 125-129, 2016.*

[5] Bukodi, E., Goldthorpe, J. H., *Intergenerational class mobility in industrial and post-industrial societies: Towards a general theory. Rationality and Society, 34(3), pp. 271-301, 2022.*

[6] Mariani, M., Borghi, M., *Industry 4.0: A bibliometric review of its managerial intellectual structure and potential evolution in the service industries. Technological Forecasting and Social Change, 149, pp. 119752, 2019.*

[7] Victor, C., M, Ramachandran., Gary, W., Robert, J. W., Chung, S. L., Paul, W., *Editorial for FGCS special issue: Big Data in the cloud. Future Generation Computer Systems, 65, pp. 123-125, 2016.*

[8] Zoidov, K., Medkov, A., *Transit Economy in Global Post-Industrial Eurasia. Post-Industrial Society: The Choice Between Innovation and Tradition, 193-209, 2020.*

[9] Marijn, J., Haiko, van der V., Agung, W., *Factors influencing big data decision-making quality. Journal of Business Research, 70(9), pp. 56-58, 2017.*

[10] Gries, T., Grundmann, R., Palnau, I., et al. *Technology diffusion, international integration and participation in developing economies - a review of major concepts and findings. International Economics and Economic Policy, 15(1), pp. 215-253, 2018.*

[11] Ebrahim, A., Brown, L. D., Batliwala, S., *Governance for global integration: Designing structure and authority in international advocacy NGOs. World Development, 160, pp. 106063, 2022.*

[12] Alessandro, M., *Regulating big data. The guidelines of the Council of Europe in the context of the European data protection framework. Computer Law &amp; Security Review: The International Journal of Technology Law and Practice, 12(3), pp. 45-46, 2017.*

[13] Ahmed, O., Fatima, Z. B., Ayoub, A.-L., Samir, B., *Big Data technologies: A survey. Journal of King Saud University - Computer and Information Sciences, 10(4), pp.99-102, 2017.*

[14] Niu, Y., Ying, L., Yang, J., Bao, M., & Sivaparthipan, C. B. *Organizational business intelligence and decision making using big data analytics. Information Processing & Management, 58(6): 102725, 2021.*

[15] Novak, A., Bennett, D., Kliestik, T., *Product decision-making information systems, real-time sensor networks, and artificial intelligence-driven big data analytics in sustainable Industry 4.0. Economics, Management and Financial Markets, 16(2): 62-72, 2021.*

[16] Koot, M., Mes, M. R. K., Iacob, M. E., *A systematic literature review of supply chain decision making supported by the Internet of Things and Big Data Analytics. Computers & Industrial Engineering, 154: 107076, 2021.*

[17] Bousdekis, A., Lepenioti, K., Apostolou, D., et al. *A review of data-driven decision-making methods for industry 4.0 maintenance applications. Electronics, 10(7): 828, 2021.*

[18] Araz, O. M., Choi, T. M., Olson, D. L., & Salman, F. S. *Role of analytics for operational risk management in the era of big data. Decision Sciences, 51(6): 1320-1346, 2020.*

[19] Tantalaki, N., Souravlas, S., Roumeliotis, M., *Data-driven decision making in precision agriculture: The rise of big data in agricultural systems. Journal of agricultural & food information, 20(4): 344-380, 2019.*

[20] Özemre, M., Kabadurmus, O., *A big data analytics based methodology for strategic decision making. Journal of Enterprise Information Management, 33(6): 1467-1490, 2020.*

[21] Alsunaidi, S. J., Almuhaideb, A. M., Ibrahim, N. M., Shaikh, F. S., Alqudaihi, K. S., Alhaidari, F. A., ... & Alshahrani, M. S. *Applications of big data analytics to control COVID-19 pandemic. Sensors, 21(7): 2282, 2021.*

[22] Bibri, S. E., *On the sustainability of smart and smarter cities in the era of big data: an interdisciplinary and transdisciplinary literature review. Journal of Big Data, 6(1): 1-64, 2019.*

[23] Ardito, L., Cerchione, R., Del, Vecchio, P., et al. *Big data in smart tourism: challenges, issues and opportunities. Current Issues in Tourism, 22(15): 1805-1809,*

[24] Nocker, M., Sena, V., *Big data and human resources management: The rise of talent analytics. Social Sciences, 8(10): 273, 2019.*

[25] Ogbuke, N. J., Yusuf, Y. Y., Dharma, K., & Mercangoz, B. A. *Big data supply chain analytics: ethical, privacy and security challenges posed to business, industries and society. Production Planning & Control, 33(2-3): 123-137, 2022.*

[26] Rehman, A., Naz, S., Razzak, I., *Leveraging big data analytics in healthcare enhancement: trends, challenges and opportunities. Multimedia Systems, 28(4): 1339-1371, 2022.*

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Nov 1, 2023
Accepted: Nov 20, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1904–1913, DOI 10.12694/scpe.v25i3.2789

ELECTRIC ENERGY METERING ERROR EVALUATION METHOD BASED ON DEEP LEARNING

TIANFU HUANG[*], ZHIWU WU[†], WEN ZHAN[‡], CHUNGUANG WANG[§] AND TONGYAO LIN[¶]

Abstract. The measuring accuracy of the electric energy meter, voltage transformer and current transformer shows a dynamic state under the influence of its factors and external factors. The error of the voltage transformer and current transformer cannot be measured by traditional method. This paper establishes a multidimensional error analysis and fault diagnosis system for power metering based on Hadoop architecture and Spark memory calculation. The platform extracted the error signal from the measurement data and calculated the characteristic value of the error signal. Then, dependent cloud and dynamic time rules are used to estimate the transformer's and voltage transformer's continuity. Then, a half-step membership degree cloud generation algorithm is constructed to overcome the error bias randomness and fuzzy characteristics under the influence factors. Finally, the system uses the dynamic correction method to estimate the similarity of error timing and quantitative factors to realize the error calculation of the current transformer and voltage transformer. The power metering error processing system was built with the support of Hadoop and Spark. The timing increment is introduced in the process of data collection. Dependent cloud and dynamic time-repair methods can improve the accuracy of diagnosing errors in electric energy metering. The parallel optimization of big data platforms by belonging to the cloud and dynamic time-warping algorithm is verified.

Key words: Deep learning; Energy metering error; Affiliated cloud; Dynamic time warping; System Estimation

1. Introduction.

Power is essential to our country's economy and People's Daily lives. The key to national construction and development is modernizing information construction and management. It is necessary to study all kinds of errors in electric energy measurement systems and reduce their influence as much as possible. Compared with the previous calibration equipment, the multi-purpose electronic calibration equipment has more functions, better performance and higher intelligence. This method has high advantages in the accuracy and speed of inspection. The multiple meter calibration equipment developed and put into operation by the State Grid in 2015 can verify multiple units of measurement at the same time. The measurement accuracy can reach 0.01 magnitude. Literature [1] elaborated on the failure causes of electronic watt-hour meters from internal circuit structure, external humidity sensitivity and chip packaging technology and proposed corresponding countermeasures. Literature [2] uses OOK dynamic test signal modeling combined with the Monte Carlo method to study the error characteristics of digital watt-hour meters. The functional relationship between the factors and the resultant results is examined. The measuring accuracy of the energy meter, voltage transformer and current transformer shows a dynamic state under the influence of its factors and external factors. Reference [3] calibration of frequency multiplier resonance equipment with calibration electrode.

At the same time, the corrected multiplier is used to calculate the correction factor together with the measured output voltage and equivalent resistance. A correction factor is introduced to optimize the uncertainty caused by the resonance effect. In this way, a contactless electrostatic voltmeter calibration scheme is realized. At the same time, the power and quantity errors of different voltage and current are simulated respectively under sinusoidal and non-sinusoidal conditions. The real-time monitoring of the energy meter and the secondary loop is realized [4]. Traditional methods cannot measure the voltage transformer and current transformer error. The measurement error is estimated by the extrapolation method. But this calculation only involves secondary loads, primary currents and primary voltages. Ambient temperature, applied electric and magnetic fields, and

[*] Marketing Service Center of State Grid Fujian Electric Power Co., Ltd., Fuzhou, Fujian, 350013, China (Corresponding author, m18650869579@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

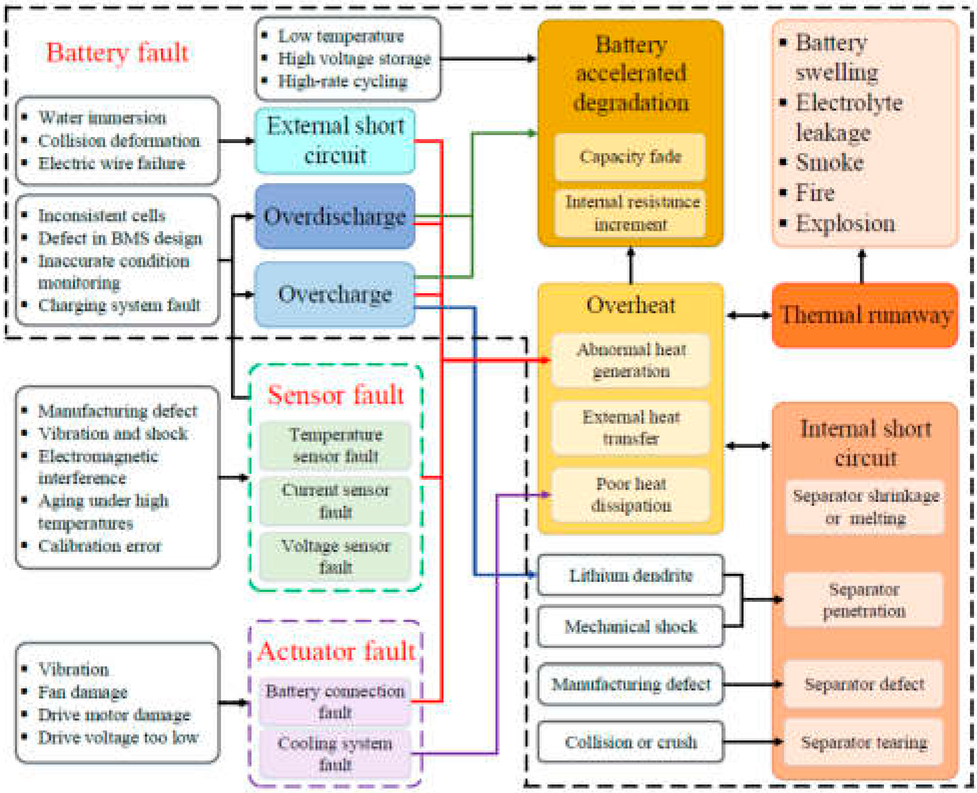


Fig. 2.1: Fault diagnosis method of electric energy metering error information.

leakage current affect the measurement results. The influence of each factor on the measurement result is random and fuzzy. Therefore, in this paper, dependent cloud and dynamic timing algorithms are used to continuously estimate the current and voltage transformer's ratio and phase errors. A half-step membership cloud generation algorithm is constructed [5]. The purpose is to overcome the error bias, randomness, and fuzzy characteristics under the influence of various factors.

2. Digital energy metering analysis based on big data. An error detection method of electric energy metering based on big data is proposed. Hive is selected as the data warehouse management method in this project. Spark computing architecture's efficient characteristics are used to research fast storage of energy metering errors, efficient calculation and distributed parallel optimization for big data [6]. The organic integration of Hadoop and Spark can improve the storage capacity, parallel optimization processing and computing speed of large-scale measurement data. The electric energy measurement error measurement method based on big data is proposed. The system mainly includes four parts: data acquisition, storage and calculation, analysis and diagnosis, and engineering application.

As shown in Figure 2.1 (Picture quoted from Advanced Fault Diagnosis for Lithium-Ion Battery Systems), the system's primary function is to collect, clean, transform and encapsulate the collected information in real time. Among them, Hadoop and Spark are used as carriers to achieve adequate storage of massive data. The characteristic values of electric energy measurement errors are extracted in the analysis and diagnosis, and the type diagnosis is realized. In the engineering application, real-time query and fault analysis of power measurement data is realized. The units showed a strong correlation in the test. Each stage from data collection to actual project implementation is the support and guarantee of subsequent work. The collection and storage of data is the prerequisite for accurate analysis of measurement results [7]. The centre of measurement error is analytic diagnosis. The measurement error calculation method is to show the measurement results to the user.

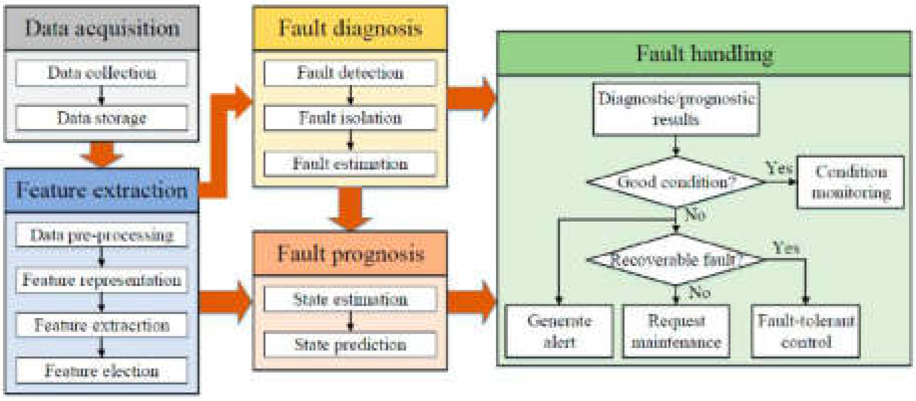


Fig. 2.2: Data processing flow of electric energy metering error.

The error analysis process of electric energy measurement is shown in Figure 2.2.

The Kettle cluster mode is selected during data entry. The work of multiple data collection devices is synchronized and distributed to each PC in the cluster. When the device status is upgraded, not only must the device status data file table be entered into the database, but also the upgraded device file data information must be upgraded. In this way, it is possible to complete the periodic incremental input of data information for various energy metering equipment [8]. As the underlying platform, Hadoop can store a large amount of energy-metering data at high speed. The distributed Spark algorithm can quickly and efficiently process intra-cluster resources and ensure the algorithm's accuracy. The parallel optimal algorithm based on Spark can diagnose and analyze the cause of specific faults in power metering devices. The steps to analyze the problems existing in power measurement using big data are as follows:

1. In periodic increments, The Kettle cluster model inputs power-related data from the customer's power information collection system into the big data infrastructure platform.
2. The analysis of error characteristics in power measurement is divided into three categories: transformer error, analog input merging error and digital error. Secondly, the eigenvalues of electrical energy errors are extracted and calculated.
3. The method of slice and rotation is used to analyze and diagnose the multidimensional error characteristic quantity.
4. Using dependent cloud and dynamic time correction methods to diagnose and type judge measurement errors.

3. Measurement error estimation based on affiliated cloud.

3.1. Membership Cloud Theory. The membership degree cloud method is a method that transforms qualitative and quantitative indexes into each other. Suppose σ is a set of ordinary values. S is the qualitative idea, which relates to V. Assume a random embodiment B of type $x \in \sigma$. The correspondence between x and S, σ, is determined by the following formula:

$$\begin{cases} \sigma : v \to [0, 1] \\ \forall x \in v, x \to v(x) \end{cases} \tag{3.1}$$

$v(x)$ is the degree of membership of x relative to S. Merging and distributing between $(x, v(x))$ is called a slave cloud. The member cloud describes the qualitative concept with three parameters:

1. The expected value W_x determines the mean value of the member cloud.
2. The entropy value E determines the variation amplitude of the cloud cluster;
3. Super entropy F_e is an important factor affecting the dispersion degree of cloud water droplets. The cloud droplet data is replaced by the reversely owned cloud to obtain the parameters [9]. Backward

dependent cloud uses statistical methods to translate accurate data into two qualitative concepts, W_x, W_n and F_e. The three parameters of$N_s x_{ai}$ samples can be calculated in the following way:

$$\begin{cases} W_x = \dfrac{1}{N_a} \displaystyle\sum_{i=1}^{N_a} x_{ai} \\ W_n = \dfrac{\sqrt{\pi/2}}{N_a} i = \displaystyle\sum_{i=1}^{N_a} |x_{ai} - W_x| \\ F_e = \sqrt{\dfrac{1}{N_a - 1} \displaystyle\sum_{i=1}^{N_a} (x_{ai} - W_x)^2 - W_n^2} \end{cases} \tag{3.2}$$

After judging W_x, W_n and F_e of x_{ai}, the owning cloud distribution of $(x, v(x))$ can be generated from the positively owning cloud cluster. The membership degree $\sigma(x_{ai})$ of x_{ai} to S can be obtained using the parameters W_x and W'_n in fuzzy mathematics. Where W'_n is an arbitrary number, it conforms to the normal distribution [10]. Its expected value is W_n. The standard deviation is F_e. $\sigma(x_{ai})$ has many different functions for x_{ai}, W_x and W'_n, and can generate sets of many membership degrees. The semi-trapezoid and semi-normal membership clouds are selected to examine the influence of various factors on the measurement results. A half-step fuzzy mathematical model is established to describe the influence of various parameters on the measurement results.

3.2. Temperature and frequency belong to the cloud. The temperature of the operating environment of the voltage transformer and current transformer is -25℃ to 55℃. The measurement deviation is temperature-independent in the temperature range close to the Calibration. However, the measurement errors in the high and low-temperature regions vary significantly with the increase and decrease of temperature. In this paper, a half-step membership function is established to characterize the influence of atmospheric temperature on the measurement results. Figure 3.1 is a subgroup of measurement errors as a function of temperature [11]. It is A semi-trapezoidal cloud layer, denoted by $S(W_{xR1}, W_{nR1}, F_{eR1})andS(W_{xR2}, W_{nR2}, F_{eR2})$. If the temperature is between W_{xR1} and W_{xR2}, the degree of membership of this error deviation value is 0. Their corresponding membership functions can quantify the two members outside this interval. Using entropy weightW_{nR1}, W_{nR2}, superentropy F_{eR1} and F_{eR2}, a semi-stepped cloud model is constructed to describe the variation amplitude and the dispersion degree of cloud droplets. The following procedure is used to process the hybrid half-ladder dependent cloud algorithm where error deviation produces the surrounding temperature clouds:

1. Generate random values W'_{nR1} and W'_{nR2}, which are normally distributed. $W'_{nR1} \sim N(W_{nR1}\text{and}F^2_{eR1}), W'_{nR2} \sim N(W_{nR2}\text{and}F^2_{eR2})$.
2. Generate random numbers W'_{nR1} and W'_{nR2}, which are normally distributed. $x_{R1} \sim N(W_{xR1}\text{and}W'^2_{nR1}), x_{R2} \sim N(W_{xR2}\text{and}W'^2_{nR2})$.
3. Repeat steps 1 and 2 until $D \times 1$ binding vectors x_R of x_{R1} and x_{R2} and $D \times 1$ binding vectors W'_{nR} of W'_{nR1} and W'_{nR2} are generated.
4. By substituting the values of the surrounding temperature x_R and W'_{nR} into formula (3), the degree of membership of the surrounding temperature x_R for the deviation of the measurement error can be found:

$$(x_R, W'_{nR}) \begin{cases} 1 - e^{\frac{(x_R - W_{xR2})^2}{2{W'_{nR}}^2}}, & x_R < W_{xR2} \\ 0, & W_{xR2} \leq x_R \leq W_{xR1} \\ 1 - e^{\frac{(x_R - W_{xR1})^2}{2{W'_{nR}}^2}}, & x_R > W_{xR2} \end{cases} \tag{3.3}$$

Using the forward-owning cloud cluster algorithm from step (1) to step (4), D cloud droplets of (x_R, σ_R) can be generated. The distribution of temperature clusters in each region under each error deviation is given. Figure 3 shows the distribution of the owning group when $\begin{matrix} W_{xR1} = 30°C, W_{nR1} = 10°C, F_{eR1} = 4°C, \\ W_{xR2} = -5°C, W_{nR2} = 5°C, F_{eR2} = 1°C \end{matrix}$ occurs. Formula (3) is a function with three components. The midpoint here is 0. The left half is in a downward

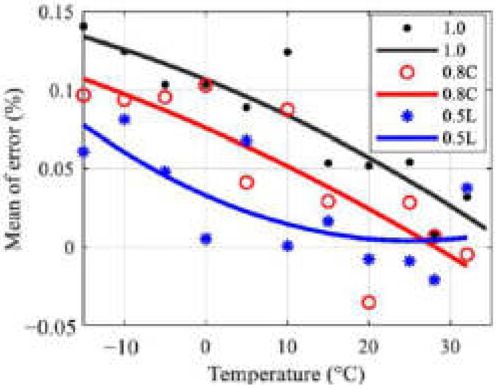


Fig. 3.1: Subordinate cloud distribution of ambient temperature causing error deviation.

trend, and the right half is up until each reach zero. The dependent cloud in Figure 3.1 is a trapezoidal distribution with a broad upper side and a narrow bottom side [12]. The semi-ladder-shaped subordinate cloud in the left region $S(W_{xR1}, W_{nR1}, F_{eR1})$ and the semi-ladder-shaped subordinate cloud rising in the right region $S(W_{xR2}, W_{nR2}, F_{eR2})$ combine to form the hybrid semi-ladder-shaped subordinate cloud in the figure.

It is difficult to obtain the parameters of hybrid semi-trapezoidal dependent cloud in the manufacturers and brands of current transformers and voltage transformers [13]. The above parameters can be obtained from the measured temperature sample data. The measurement result is divided into two parts: (1) the measurement result of the left part is lower than the calibration result; (2) The group on the right contains all other temperature data. The values of $W_{xR2}, W_{nR2}, F_{eR2}$ and $W_{xR2}, W_{nR2}, F_{eR2}$ can be calculated by substituting the temperature sampling data of the two groups on the left and right into equation (2). A half-step fuzzy model is proposed to estimate the effect of atmospheric temperature x_{Rr} on the power measurement results. In the central region $B_R = [x_{Rr} - (F_{eR1} + F_{eR2})/3, x_{Rr} + (F_{eR1} + F_{eR2})/3]$, the deviation of measurement error caused by atmospheric temperature to the number of cloud droplets Z can be expressed as

$$G_R(x_{Rr}) = \frac{\lambda_R G_{\lim}}{Z} \sum_{x_R \in B_R} \sigma_R \tag{3.4}$$

$G = g, \varphi$ is the commutation and phase deviation of the transformer. $G_{\lim} = g_{\lim}, \varphi_{\lim}$ is the corresponding limit value. $\lambda_R$is the allowable error range of the measurement result and the ambient temperature range. G is used to replace the R of the subscript in formula (3) and thus $W_{nG1} = W_{nG2} = W_{nG}$ and $W_{nG1} = W_{nG2} = W_{nG}$ to obtain a stepwise dependent cloud, which has a frequency shift relative to the measurement error. The model contains only four parameters: W_{xG1}, W_{xG2}, W_{nG} and F_{eG}. A cloud model based on a positive membership degree is proposed. The membership function should adopt the above symmetric ladder function [14]. The influence of the number of observations on the measurement accuracy is estimated using a symmetrical ladder statistical model.

3.3. Affiliated clouds of other influencing factors. The relationship between the measurement error of the voltage transformer and the external electric field shown in Figure 3.2 can be represented by a rising semi-trapezoidal dependent cloud. Its expression is $S(W_{xW}, W_{nW}, F_{eW})$. The calculation method of the influence of the external electric field on the measurement accuracy is similar to the previous part. The central rain-type cloud system based on gradient is proposed. The member function of $\mu_W(x_W, W'_{nW})$ in the semi-trapezoidal appendage cloud can be expressed as:

$$\mu_W(x_W, W'_{nW}) = \begin{cases} \overset{field}{e^{\frac{(x_W - W_{xW})^2}{2{W'_{nW}}^2}}}, & x_W < W_{xW} \\ 1, & x_W \geq W_{xW} \end{cases} \tag{3.5}$$

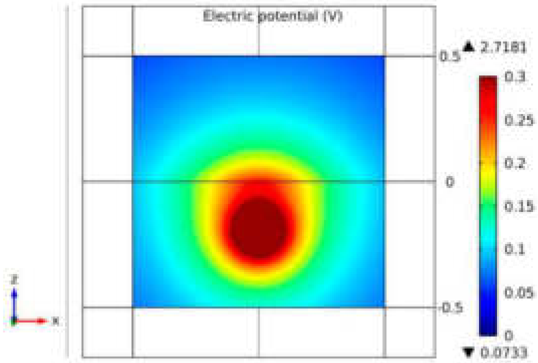


Fig. 3.2: Membership cloud of external electric field measurement error.

x_W and W'_{nW} represent the random number and standard deviation of the electric field cloud drop number, respectively. Previous formula is a fragment function with two parts. The right-hand part is 1. The left paragraph goes from 0 to 1. This causes the degree of membership to change from a discrete growth trend to a continuous saturation state in a specific area of the chart.

The causes of residual magnetism in the transformer core are the breaking of the secondary winding and the sudden drop of current. Therefore, the permeability of the core will be reduced, and the accuracy of the transformer will be affected. The DC component of the residual magnetic field tends to zero with increasing time. Thus, the measurement error of the current transformer is reduced. In this way, the time t_U lost from the most recent current can represent the effect of the remaining magnetic field [15]. The effect of the residual magnetic field on the measurement error deviation is shown in Figure 3.3 with the descending semi-normal dependent cloud cluster. Its expression is $S(W_{xU}, W_{nU}, F_{eU})$. The error is most significant at point t_U. A cloud model based on descending semi-normality is proposed. The membership function $\mu_U(x_U, W'_{nU})$ of the lower semi-normal dependent cloud is shown as follows:

$$\mu_U(x_U, W'_{nU}) = \begin{cases} 1, & x_U \leq W_{xU} \\ e^{\frac{(x_U - W_{xU})^2}{2{W'_{nU}}^2}}, & x_U > W_{xU} \end{cases} \tag{3.6}$$

x_U is the cloud droplet in the remaining magnetic field. W'_{nU} is a normally distributed random value of the standard deviation. The belonging cloud system with a semi-normal residual magnetic field differs from the one with a semi-stepped magnetic field. As seen from Figure 5, the standard deviation W'_{nU} is expressed in terms of normal distribution. Its expected value and standard deviation are W_{nU}, and the standard deviation is F_{eU}. This allows the dependent cloud to incorporate cloud droplet distribution.

3.4. Similarity between measurement results and factors. Assume that the time series for the measurement error deviation is $x = \{x_1, x_2, \cdots, x_m\}$. The influence factors of voltage transformer are $y = \{y_{X1}, y_{X2}, \cdots, y_{Xm}\}$. $X = R, W, G$ and M stand for temperature, electric field, frequency, magnetic field$y_Y = \{y_{Y1}, y_{Y2}, \cdots, y_{Yn}\}$. The influencing factors of current transformer are. $Y = U, R, M$ and S represent remanence, temperature, magnetic field, and leakage current, respectively. The similarity measurement based on dynamic time warping is also valid for various influencing factors. Figure 6 shows the principle of dynamic time adjustment. The regularization route $k = \{k_1, k_2, \cdots, k_D\}$ is searched from the start point (x_1, y_{Y1}) to the end point (x_M, y_{Y_n}). Wherek_D represents the optimal structured path of distance $B(x_i, y_{Yj}) = |x_i - y_{Yj}|, k_{opt}$ can be chosen to minimize the cumulative distance of dynamic time warping along this path.

$$B_{DTW} = \sum_{k_{opt}} B(x_i, y_{Yj}) = \min(\sum_{k=1}^{D} k_k) \tag{3.7}$$

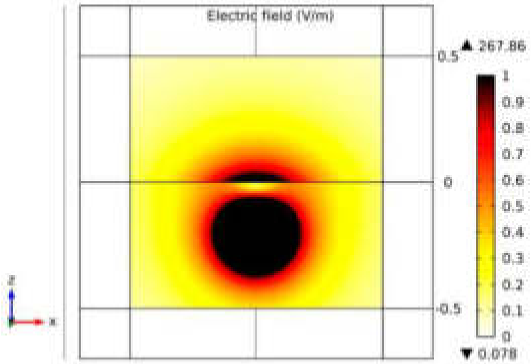


Fig. 3.3: Membership cloud of remanent magnetic measurement error.

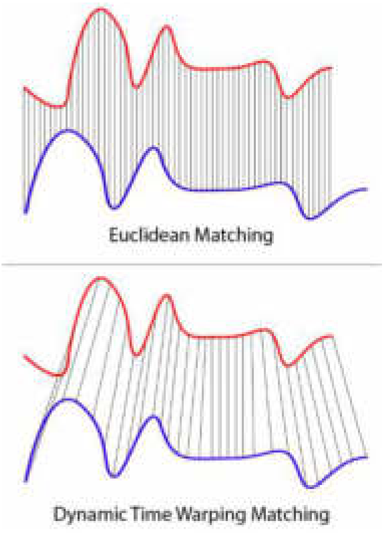


Fig. 3.4: Dynamic time warping of time series.

A self-repeating algorithm based on time series data is proposed. However, if you repeat it too much, either horizontally or vertically, the fragment shown in Figure 3.4 will not match the other longer fragments. The traditional dynamic time correction methods add gradient restriction to compensate for this shortcoming. A modified rule path by constraining self-repeating on any node. There are the following constraints on the improvement of dynamic time adjustment method:

1. Boundary conditions of regularized paths: $k = \{k_1, k_2, \cdots, k_k\}, k_1 = B(x_1, y_{Y1})$, $k_k = B(x_m, y_{Y_n})$;
2. The monotonic condition is as follows: when$k_{k-1} = B(x_{i'}, y_{Yj'})$ and $k_k = B(x_i, y_{Yj})$, $i - i' \geq 0, j - j' \geq 0$ and $i - i' + j - j' \neq 0$.
3. Continuity condition: $i - i' \leq 1, j - j' \leq 1$ when $k_{k-1} = B(x_{i'}, y_{Yj'})$ and $k_k = B(x_i, y_{Yj})$.
4. The slope restriction conditions are:

$$0 \leq A = \frac{\max(S_{xm}, S_{yn})}{S_{\lim}} < 1 \tag{3.8}$$

S_{xm} and S_{yn} represent the automatic copy time in vertical $(i - i' = 0)$ and horizontal $(j - j' = 0)$, respectively. $S_{\lim}$ is for persistent limit. A is the tilt factor of the critical value. Points within the restricted area may not be considered. They are on an unwanted, twisted trajectory. In the closed region, the slope constraint can write the set of points in B_ε as:

$$B_\varepsilon = \begin{Bmatrix} B(x_i, y_{Yj})|S_{\text{lim}} \leq i \leq m, 0 \leq j \leq \left\lceil \dfrac{i - S_{\text{lim}}}{S_{\text{lim}}} \right\rceil \\ or\ S_{\text{lim}} \leq j \leq n, 0 \leq i \leq \left\lceil \dfrac{j - S_{\text{lim}}}{S_{\text{lim}}} \right\rceil \end{Bmatrix} \tag{3.9}$$

$\lceil \cdot \rceil$ is the upper bound function. The distance $B(x_i, y_{Yj})$ in the B_ε region is replaced by a larger constant $B_{\max}$ to minimize the cumulative distance of the optimal regularization route so that the points in the restricted region are found on the optimal regularization route. The cumulative distance from point (x_1, y_{Y1}) to (x_i, y_{Yj}) on the optimal regular route is defined as $S(i, j)$. The recursive formula for the cumulative distance $S(i, j)$ is as follows:

$$S(i, j) = B(x_i, y_{Yj}) + \Delta B \tag{3.10}$$

ΔB is the distance accumulated in front of the point (x_i, y_{Yj}). It depends on the number of self-repeats S_{xm} and S_{yn} make in succession along both vertical and horizontal lines. The recursive formula can improve the dynamic time adjustment and find the optimal regular route to achieve the shortest cumulative distance. At this point, an improved dynamic time adjustment scheme is obtained:

$$B_{MDTW} = S(m, n) \tag{3.11}$$

4. Test and result analysis.

4.1. Establishment of test environment and collection of sampling data. Two hosts are used as cluster hosts. The other eight are normal. The two central nodes are ”one active, one standby” cases. The data storage system mainly completes the storage, management and scheduling of test data. The latter is to ensure the reliability of the cluster system. Select CentOS6.4 as the operating system. Maximize your use of the /home directory. The experimental case data is obtained from the primary platform file of the power measurement equipment. The Kettle software is used to preprocess data and then input all data into the Hive data warehouse. After analysing the fault intelligently, the Spark SQL model queries and displays the fault in real-time.

4.2. Research on Multidimensional Testing Methods. The fault diagnosis system of electric energy measurement error analysis is constructed. Since the operation rate and accuracy of the algorithm are directly related to the overall effect of the entire power measurement error analysis, the test results of the speed and accuracy of the proposed scheme's power measurement error feature extraction are listed in Table 4.1.

By analysing the data of different fault categories, the fault diagnosis effect of the proposed method based on cloud ownership and dynamic time correction is further verified. At the same time, the accuracy of the power measurement error extensive data analysis system is judged. The diagnostic conclusions of the analysis are shown in Table 4.2.

Table 4.2 lists five typical errors in electrical energy measurement. There are 100 experimental data. The experiment was carried out in single-machine mode and cluster mode. When the number of test cases is larger, the accuracy of the test can reflect the true degree of test error. True diagnostic results were obtained from 100 test data (Figure 4.1)). The solid lines on the left represent each of the five error types from R1 to R5. The dotted line on the right clearly represents the error type's diagnostic accuracy. The fault diagnosis accuracy of post-cluster methods using dependent cloud and dynamic time-repair methods differs significantly from that of a single method. The identification accuracy of R2 error in cluster mode is significantly higher than that in single-machine mode. The experimental results show that the dependent cloud and dynamic time repair method can improve the accuracy of diagnosing the types of energy metering errors. The parallel optimization of big data platforms by belonging to the cloud and dynamic time-warping algorithm is verified.

Table 4.1: Test and verification results of calculation of error eigenvalues.

Error information evaluation index	**The electric energy meter reversed**	**Electric energy meter stalls**	**Electric energy meter spinning**
Record time	2022.08.01	2022.09.01	2022.108.01
Output records (strips)	8	356	9
Time (s)	5.994	3.731	1.410
verify	Table number is 31139	The table code is 28313	The table code is 30382
	The table code of 2022.07.03 is 10.03	PAP_R=0.02 on 2022.09.01. PRP_R=1.10	RUN_CAP=23958
	The table code of 2022.08.01 is 0.0	PRP_R =0.00 on 2022.09.02. PRP_R=0.00	PRP_E=415333
Verify conclusion	It is verified that the electric energy meter is running backward, and the conclusion is correct.	It is verified that the electric energy meter has stopped, and the conclusion is correct.	It is verified that the electric energy meter is flying, and the conclusion is correct.

Table 4.2: Test and verification results of calculation of error eigenvalues.

Error type	**ID**
The transformer is improperly connected	R1
The collection device is disconnected	R2
The energy meter is incorrectly connected	R3
Analog input combined cell quantization error	R4
Normal state	R5

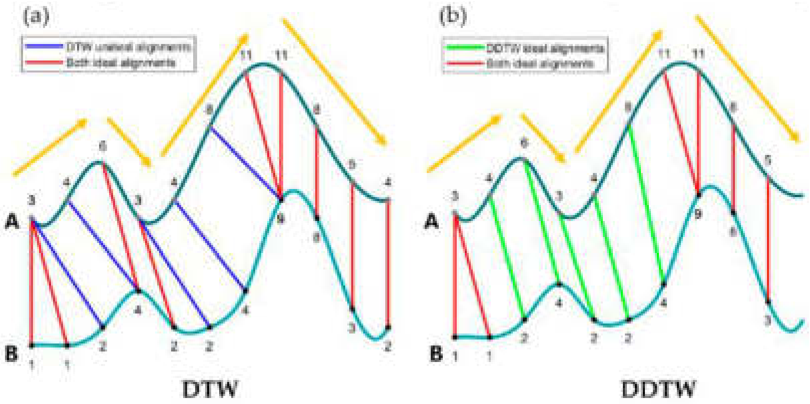


Fig. 4.1: Comparison of accuracy rates of membership cloud and dynamic time warping diagnosis under two modes.

5. Conclusion. In this paper, the deviation problem of digital electric energy measurement is discussed in detail from the transfer of sample value, the input and combination of analogy quantity, the measurement error of digital electric energy meter and their effect on electric energy measurement. The power metering error processing system is constructed with the support of Hadoop and Spark. The timing increment is introduced in the process of data collection. The flexible distributed database based on Spark architecture is used to solve the problem quickly, and the error characteristic of the system is obtained. The power metering extensive data analysis system based on the owning cloud and dynamic timing restoration is constructed to optimize real-time data. At the same time, it can also be used to diagnose all kinds of faults in electric energy measurement. The experiment proves that an effective distributed data processing method is realized on the big data platform.

REFERENCES

[1] Yu, X., Zheng, D., & Zhou, L. Credit risk analysis of electricity retailers based on cloud model and intuitionistic fuzzy analytic hierarchy process. International Journal of Energy Research, 2021;45(3): 4285-4302.

[2] Asres, M. W., Ardito, L., & Patti, E. Computational cost analysis and data-driven predictive modeling of cloud-based online-NILM algorithm. IEEE Transactions on Cloud Computing, 2021; 10(4): 2409-2423.

[3] Tajalli, S. Z., Kavousi-Fard, A., Mardaneh, M., Khosravi, A., & Razavi-Far, R. Uncertainty-aware management of smart grids using cloud-based lstm-prediction interval. IEEE Transactions on Cybernetics, 2021;52(10): 9964-9977.

[4] Geng, Y., Pan, F., Jia, L., Wang, Z., Qin, Y., Tong, L., & Li, S. UAV-LiDAR-based measuring framework for height and stagger of high-speed railway contact wire. IEEE Transactions on Intelligent Transportation Systems, 2021; 23(7): 7587-7600.

[5] Koivumäki, P., Steinböck, G., & Haneda, K. Impacts of point cloud modeling on the accuracy of ray-based multipath propagation simulations. IEEE Transactions on Antennas and Propagation, 2021; 69(8): 4737-4747.

[6] Tran, M. K., Panchal, S., Chauhan, V., Brahmbhatt, N., Mevawalla, A., Fraser, R., & Fowler, M. Python-based scikit-learn machine learning models for thermal and electrical performance prediction of high-capacity lithium-ion battery. International Journal of Energy Research, 2022; 46(2): 786-794.

[7] Yang, L., Yu, K., Yang, S. X., Chakraborty, C., Lu, Y., & Guo, T. An intelligent trust cloud management method for secure clustering in 5G enabled internet of medical things. IEEE Transactions on Industrial Informatics, 2021; 18(12): 8864-8875.

[8] Miller, C., Picchetti, B., Fu, C., & Pantelic, J. Limitations of machine learning for building energy prediction: ASHRAE Great Energy Predictor III Kaggle competition error analysis. Science and Technology for the Built Environment, 2022; 28(5): 610-627.

[9] Ghafari, R., Kabutarkhani, F. H., & Mansouri, N. Task scheduling algorithms for energy optimization in cloud environment: a comprehensive review. Cluster Computing, 2022; 25(2): 1035-1093.

[10] Pretto, S., Ogliari, E., Niccolai, A., & Nespoli, A. A new probabilistic ensemble method for an enhanced day-ahead pv power forecast. IEEE Journal of Photovoltaics, 2022; 12(2): 581-588.

[11] Azimi Nasab, M., Zand, M., Eskandari, M., Sanjeevikumar, P., & Siano, P. Optimal planning of electrical appliance of residential units in a smart home network using cloud services. Smart Cities, 2021; 4(3): 1173-1195.

[12] Saxena, D., Gupta, I., Singh, A. K., & Lee, C. N. A fault tolerant elastic resource management framework toward high availability of cloud services. IEEE Transactions on Network and Service Management, 2022; 19(3): 3048-3061.

[13] Alam, T. Cloud-based IoT applications and their roles in smart cities. Smart Cities, 2021; 4(3): 1196-1219.

[14] Mansour, R. F., Alhumyani, H., Khalek, S. A., Saeed, R. A., & Gupta, D. Design of cultural emperor penguin optimizer for energy-efficient resource scheduling in green cloud computing environment. Cluster Computing, 2023; 26(1): 575-586.

[15] Belgacem, A., & Beghdad-Bey, K. Multi-objective workflow scheduling in cloud computing: trade-off between makespan and cost. Cluster Computing, 2022;25(1):579-595.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Nov 2, 2023
Accepted: Nov 20, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1914–1922, DOI 10.12694/scpe.v25i3.2790

RESEARCH ON DATA MINING AND REINFORCEMENT LEARNING IN RECOMMENDATION SYSTEMS

YUERAN ZHAO*AND HUIYAN ZHAO†

Abstract. This paper aims to help students better grasp the required professional knowledge and core concepts. This paper presents a design method for a multi-layer knowledge base based on XML. According to learners' identity characteristics and learning behaviour, using the mathematical statistics method, the feature expression for the learning system is constructed. Multivariable linear regression theory establishes convergence constraints for accurate and deep mining. The average detection results of the collected samples are used for high-quality deep mining of user portraits in the learning system. This project intends to study the method of solving accurate confidence intervals for user portrait data in the education system. Excel and Access are used to complete the data collection of the teaching object and the construction of the database. A multi-mode interactive editing and processing method of user portrait information for education systems is studied in cloud computing. Finally, a learning system based on mathematical loading mode is proposed, and an object-oriented learning recommendation system is designed. The developed teaching software can enable students to get more teaching guidance when they acquire the required knowledge to improve students' learning effect effectively.

Key words: Knowledge recommendation; Learning needs; Personalization; Learning guidance; Learning system user profile data; Deep excavation; System Design

1. Introduction. Internet autonomous teaching has developed rapidly in the field of computers and the Internet because of its characteristics of ”individuality,” ”autonomy,” ”initiative,” and ”non-timeliness.” Scholars have built a teaching resource-sharing and management system based on network technology. They set up large-scale open courses for teachers and students so that teachers and students can use the online teaching platforms to carry out interactive learning. Currently, online self-study faces the following problems:

1. learner-oriented online learning environment can not integrate many learning resources well. Students often have difficulties in finding the materials they need for their studies.
2. Students' hidden learning needs cannot be discovered from their behavioral characteristics.
3. Lack of individualized knowledge recommendation and dynamic learning trajectory generation mechanism.
4. There are various forms of research data.

The lack of semantic information makes it difficult for computers to understand and automate. Reference [1] uses OpenCL to implement the KNN parallel computation method. The fine-grained parallelism method and the improvement of multiple thread sets are adopted in the ranging process. In the classification process, the memory model is doubly classified, and the depth of classification is increased. However, this algorithm has a high amount of computation and a substantial real-time. Literature [2] provides an effective NB-MAFIA (Maximum Logistic Organization System) method. Using the compressional coefficient of the N-List and the effective cross-algorithm, the support of the item collection can be solved quickly. Pruning the search space and discovering supersets are used to improve the performance of this method. The performance of this algorithm is not ideal for deep data mining. This paper uses mathematical statistics to build a learning system feature model based on the learner's identity characteristics and learning behavior characteristics [3]. Multivariable linear regression theory establishes convergence constraints for accurate and deep mining. The average detection results of the collected samples are used for high-quality, in-depth mining of the user portraits of the learning system. The behavioral characteristics of learners are intensely studied. Excel and Access are used to complete

*Academic Affairs Office, Zhengzhou Shengda University, Zhengzhou 451191, China (Corresponding author, rachel_yr2023@126.com)

†School of Information Engineering, Zhengzhou Shengda University, Zhengzhou 451191, China

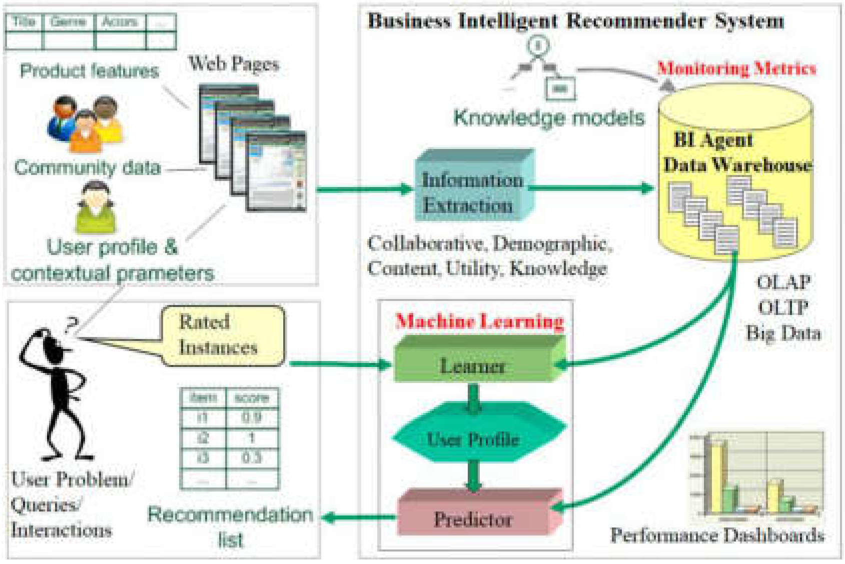


Fig. 2.1: Framework of learning recommendation system based on data mining.

the data collection of the teaching object and the construction of the database. A multi-mode interactive editing and processing method of user portrait information for education systems is studied in cloud computing.

2. Learn the design of the recommendation system. In essence, the learning recommendation system is realized through machine learning. Through the collection of students' basic information and analysis of students' learning habits, learning behaviors and test results, the external and internal learning needs of students are identified [4]. Students actively seek knowledge in the knowledge base to meet their needs—a dynamic generation of learning paths to facilitate learners to complete learning better.

2.1. Overall system framework. The general framework of the learning recommendation system is shown in Figure 2.1 (image cited in Informatics 2017, 4(4), 40). The basic steps of this method are as follows:

1. For the students who adopt this method for the first time, a questionnaire survey is conducted, and their basic information is registered. Through the self-introduction of the students and the analysis of the questionnaire, the students' subject is classified. It includes the relevant body of knowledge, cutting-edge information, and essential citations. In this way, the basic knowledge can be deduced.
2. Students can evaluate the knowledge points presented. All information is stored in the student's database and the student's learning behavior database.
3. Identify students' possible learning interests through users' basic information, learning behavior, and exam results.
4. Make knowledge recommendations based on user feedback, combined with the user's learning requirements, learning performance and the interrelation between knowledge points.
5. Students will learn the next course or topic according to the information the system pushes.

2.2. Building a Knowledge Base. Through the establishment of students' personal information and learning behaviour database, students' learning needs can be found. These include students' personal information, evaluation of learning performance and so on. All the data are stored in the database. The knowledge base includes a central course, knowledge points and related teaching and research resources [5]. Among them, knowledge processing mainly classifies, organizes and transforms the data to form a multi-level knowledge base based on XML. The hierarchy of the knowledge base is shown in Figure 2.2.

Expertise is the highest level in the hierarchy of the knowledge base. The XML file uses majorList.xml to describe the specific content in each discipline. The course information is described in the XML file [6]. Each

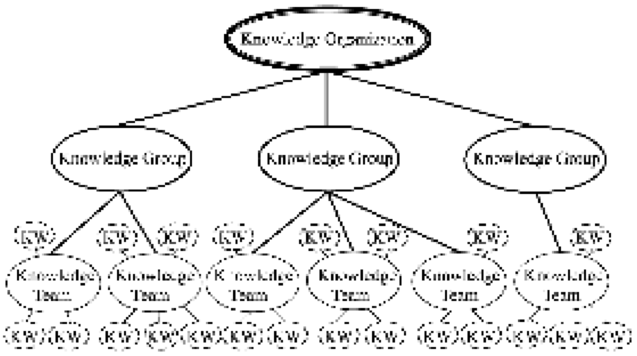


Fig. 2.2: Knowledge base hierarchy diagram.

lesson has a knowledge tree of chapters. XML is represented as separate parts. Knowledge in all sections of the XML file is represented by "partial nodes." Some nodes support hierarchical nested structures. The knowledge points in each chapter include more than one meta-knowledge point. Use a separate knowledge section ID.xml representation. The meta-knowledge point is the smallest unit of knowledge point. It can no longer be divided. Each meta-knowledge node includes related teaching resources, research resources, background resources, expansion resources and so on. This information comes in a variety of ways. These files can be Word, PDF, text, videos, etc. When describing the characteristics of chapter and meta-knowledge, it should consider not only the number, name, keywords, difficulty, and importance but also the interrelation of knowledge.

3. Data deep mining algorithm.

3.1. Mathematical modeling of multiple linear regression. The user portrait of the learning-oriented system is deeply analyzed using the sample mean detection method, and its stability is analyzed [7]. Suppose that the deep mining of the user profile data in the learning system makes the statistical function $g(u), g(u, e)$ continuously bounded on the range of U. It is expressed by $g(U), g(U, e)$. Since $G(U)$ is the unique minimum modular eigenvalue of $g(u)$. Therefore, the initial value of the Gaussian function, which trains the user portrait data for in-depth mining, is expressed as follows:

$$\bar{g}_e^{(0)} = \bar{g}_e = g(n(U^{(0)}), e^{(0)}) \tag{3.1}$$

The principle of mathematical statistics is used to construct a state characteristic equation for the user portrait data in the learning system:

$$\inf G(U, e) - \inf g(U, e) \leq \frac{T}{2} ||C(U)||_{\infty}^{2} \tag{3.2}$$

The results show that all the constraints $\ln x, e^u$are monotonically growing functions in the regressive distribution space. In the homogeneous Sobolev space, the order of the user portrait data used in the learning system is intensely mined, and its weight is:

$$C(G(U, e)) - C(g(U, e)) \leq T ||C(U)||_{\infty}^{2} \tag{3.3}$$

The robust optimal solution of user portrait data for the learning system is obtained using the sample mean detection method [8]. This project intends to use deep neural networks for research. Deep learning processing of data mining is expressed through the Jacobian matrix $H(u)$:

$$H(u) = \begin{pmatrix} \frac{\partial z_1(u)}{\partial u_1} & \frac{\partial z_1(u)}{\partial u_2} & \cdots & \frac{\partial z_1(u)}{\partial u_n} \\ \frac{\partial z_2(u)}{\partial u_1} & \frac{\partial z_2(u)}{\partial u_2} & \cdots & \frac{\partial z_2(u)}{\partial u_n} \\ \vdots & \cdots & \ddots & \cdots \\ \frac{\partial z_N(u)}{\partial u_1} & \frac{\partial z_N(u)}{\partial u_2} & \cdots & \frac{\partial z_N(u)}{\partial u_n} \end{pmatrix} \tag{3.4}$$

This project intends to study an accurate method for solving user portrait data in the education system. The vector function of edge solution for deep data mining is:

$$c_{ji}(t+1) = c_{ji}(t) - \lambda \frac{\partial G}{\partial c_{ji}} \tag{3.5}$$

$$y_{tj}(t+1) = y_{tj}(t) - \lambda \frac{\partial G}{\partial y_{tj}} \tag{3.6}$$

At the balance point $Q_0(u_1^0, u_2^0)$ of the delay discontinuity, the spectral features of the user portrait in the learning system are extracted. The output spectral characteristic quantity is obtained:

$$\dot{u}(t) = \alpha u(t) + \theta u(t - s_1(t) - s_2(t)) \tag{3.7}$$

where $u(t) = \gamma(t), t \in [-h, 0]$. By using deep learning, the data mining process is adaptive and optimized. This paper obtains the training vector:

$$u(t) = (u_0(t), u_1(t), \cdots, u_{t-1}(t))^T \tag{3.8}$$

3.2. Learning system user portrait data mining output.

Optimization solution of stable features for deep mining of user portrait data of the learning system. The existing SVM algorithm and BP neural network algorithm have significant differences in the recognition accuracy of different samples due to the interference of sampling data. Text recognition based on a deep confidence network can be divided into two stages: pre-training artificial neural network and network adjustment [9]. Most existing classification methods use dimensionality reduction to avoid dimensionality disaster, while deep belief networks (DBN) can extract low-dimensional features with strong discrimination ability from massive original features. In this way, the classification model can be built directly without dimensionality reduction. Meanwhile, it fully uses the rich information in the text. The weights of each BP neural network level are initialized using DBN network weights [10]. This method does not need to initialize any initial value of DBN, nor does it need to extend the BP neural network. BP neural network is used for global optimization to solve the local extreme value problem caused by DBN's randomness of weight parameters.

The robust optimal solution of user portrait data for the learning system is obtained using the sample mean detection method [9]. In the probability distribution interval, the user portrait data of the learning system is intensely mined. Divide the initial value $U^{(0)}$ of the initial cluster center into N parts $U^{(1)}, U^{(2)}, \cdots, U^{(N)}$,$U^{(0)} = \bigcup_{i=1}^{N} U^{(i)}$. This project intends to study an accurate method for solving user portrait data in the education system. A boundary value for convergence can be obtained:

$$C_N = \max_{1 \leq i \leq N} ||C(U^{(i)})||_\infty \tag{3.9}$$

The first-order inertial output vector of deep learning is $\omega_{tt} - \Delta\omega + |\omega|^p \omega = 0, (p > 4)$. And when$N \to \infty$ satisfies $e_N \to \infty$. An algorithm based on mathematical statistics is proposed and used to optimize the user characteristics in the learning system. Deep learning convergence can be expressed as follows:

$$s_j = \sum_{i=0}^{t-1} (u_i(t) - \eta_{ij}(t))^2, j = 0, 1, \cdots, N-1 \tag{3.10}$$

Where $\eta_j = (\eta_{0j}, \eta_{1j}, \cdots, \eta_{t-1,j})^T$,$\forall \varepsilon > 0, \exists \hat{N} > 0$. When $N > \hat{N}$ is $|\min_{u \in U^{(0)}} e = (g(u) - \lambda_N)| < \varepsilon$. The following formula describes the convergence and optimality of data mining.

$$\min_{0 \leq \lambda_i \leq c} C = \frac{1}{2} \sum_{i,j=1}^{l} v_i v_j \lambda_i \lambda_j T(u_i, u_j) - \sum_{i=1}^{l} \lambda_i + b \left(\sum_{i=1}^{l} v_j \lambda \right) \tag{3.11}$$

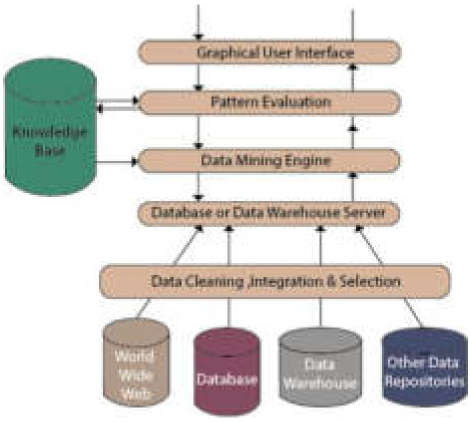


Fig. 3.1: Data mining design architecture flow.

Where (u_i, u_j) represents a controller with linear convexity. It satisfies $c \in (a_1, a_N]$ and satisfies the convergence range:

$$\bar{\delta}_j \leq \frac{f_i(\lambda) - f_i(\beta)}{\lambda - \beta} \leq \delta_j^+ \tag{3.12}$$

The average index $s_c = \frac{s-1}{2}$ is determined for the state feedback controller. The upper bound is taken according to the conditional variance$R_z^i(t)$ of boundary convergence for deep mining of user portrait data in the learning system.

$$\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_n) \neq 0 \tag{3.13}$$

$$\lambda^T L \lambda = \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j L_{ij} \geq 0 \tag{3.14}$$

In the learning system, user portrait data collection point is limited.

1) $\lim\limits_{n\to\infty} \sup |g^n(u) - g^n(v)| > 0, \quad \forall u, v \in R, u \neq v;$

2) $\lim\limits_{n\to\infty} \inf |g^n(u) - g^n(v)| = 0, \quad \forall u, v \in R;$

3) $\lim\limits_{n\to\infty} \sup |g^n(u) - g^n(v)| > 0, \quad \forall u \in R, \forall v \in Q(g)$

The deep confidence interval of user portrait in the education system is solved accurately. At the same time, the behaviour characteristics of learners are deeply explored.

Implementation of data collection flow chart. The process of data mining is designed. The deep mining system of learner portrait data based on embedded processor is studied. The system uses Lab Window/CVI as the software development tool. The deep mining technology of learner portrait data based on embedded Linux is studied. The wireless communication of the network is realized by using ZigBee protocol. The establishment of basic database is based on IEEE802.15.4 technical specification. Adaptive learning method is used in data acquisition to realize the optimal control of data acquisition [10]. Set the sample clock for the A/D module. Adds the required data sources to the user portrait dataset. The software loading function is realized by using the idea of embedded network service. This system can collect and process the learner's portrait data. Based on the above analysis, the design framework for deep mining of user portrait data in the learning system is obtained (Figure 3.1).

Table 4.1: Experimental data set.

Serial number	Data set	Number of transaction items	Number of transaction records
S1	T25I10D10K	1031	5104
S2	Retail	17156	91836
S3	Musroom	124	8463
S4	Kosarak	42990	1031252

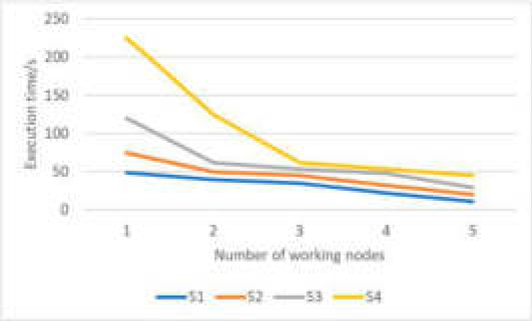


Fig. 4.1: Algorithm execution time.

4. Experiment and analysis. This project adopts the experimental clustering platform built by the Tongfang TR730 series server of Tsinghua University. One controller node and four working nodes are virtualized. Each node has 24 cores and 64 GB of storage [11]. All computing nodes are performed on Ubuntu16.10 OS. JDK1.8 and Eclipse compilation methods are used. Spark2.1.0 and Hadoop2.7.3 construct the cloud computing platform based on Hadoop2.7.3.

4.1. Dataset. The paper will collect and validate four distinctive big data sets on the mechanical data analysis and data mining platforms of FIMI and SPMF. The characteristics of this data collection are shown in Table 4.1.

S1 datasets are collections of artificial data generated by a random transaction database generator. S2 data refers to the retail data used in the supermarket basket mode. It keeps detailed records of customers' business in the mall. S3 is an open fungus data collection. The S4 data set provides a click action for a specific web news entry.

4.2. Experiment and analysis.

Scalability analysis of the algorithm. The experiment evaluated the algorithm's scalability by increasing the number of working nodes and replicating the original data set. The algorithm is dynamically tracked and improved to keep the data quantity constant [12]. Figure 4.1 shows that as the number of nodes changes from 1 to 2, the execution time of the method decreases almost linearly with the increase of the number of nodes. In the process from 3 to 5 nodes, the effect of this method is not significant. This shows that the performance of this method in parallelization has reached a relatively high level.

Under the condition that the system cluster size is five working nodes constant, the corresponding data sets are copied respectively [13]. The speed of the method changes as the data set size increases. Figure 4.2 shows that the running time of this method approximates a straight-line increase as the data set increases. At the same time, this project also proposes a parallel architecture suitable for mass data at various scales.

Algorithm performance analysis. The paper will use this paper's support vector machine, neural network, and nonlinear data mining algorithms to analyse four data types. The experiment was repeated five times. The average execution time of the algorithm is used as the final result to evaluate the performance of the algorithm. Figure 4.3 shows data set S1 with the minimum support set at 0.10%. The results show that

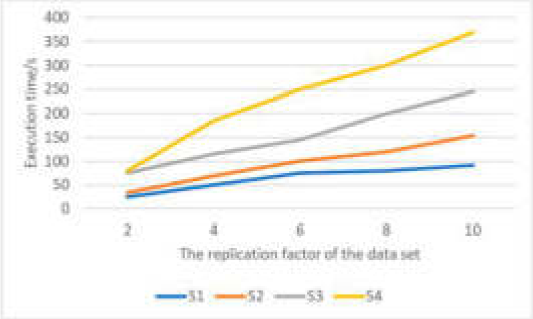


Fig. 4.2: Algorithm execution time.

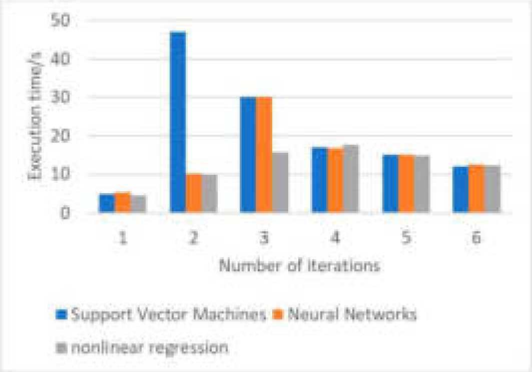


Fig. 4.3: Performance analysis of different algorithms on the S1 dataset.

the second iteration performs better than the SVM algorithm. In the third iteration, the performance of this algorithm is better than that of the two algorithms. After three iterations, the problems to be solved become small and stable [14]. When the number of frequent items in the iteration is small, the algorithm platform adaptively selects the traditional strategy to process the candidate set. In this case, the performance of the three algorithms is roughly the same.

Figure 4.4 shows the data set S2 with a minimum support pre-set of 0.15%. The results show that the proposed method has more advantages than the SVM method in the second iteration. Each iteration's performance is similar to that of the neural network algorithm [15]. This is because the set of pending objects gets smaller and becomes stable after two iterations. The number of occurrences in the last iteration process is small so that the algorithm can choose the standard solution according to the actual situation. This method is similar to the support vector machine and neural network algorithms.

Figure 4.5 shows the data set S3 with the minimum support set to 30% in advance. The results show that the proposed method has more advantages than the traditional neural network method in the second iteration. It is consistent with the iterative results of the support vector machine. This is because the set of pending objects becomes smaller and tends to be stable after two iterations. If the number of items often appearing in the iteration process is minimal, the method can choose the appropriate method to deal with the candidate set according to the need. In this case, the performance of the three algorithms is roughly the same.

Figure 4.6 shows data set S4 with the minimum support set at 0.60 percent. The results show that the nonlinear data mining method is much better than the SVM method in the second iteration. Because of the large scale of D4 samples, the project will select corresponding optimization strategies in 3,4,5 and 6 iterations. In the process of frequent monomial storage and transaction pruning, the execution time of the algorithm is

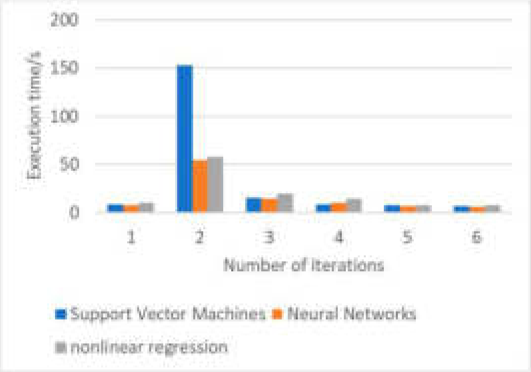


Fig. 4.4: Performance analysis of different algorithms on the S2 data set.

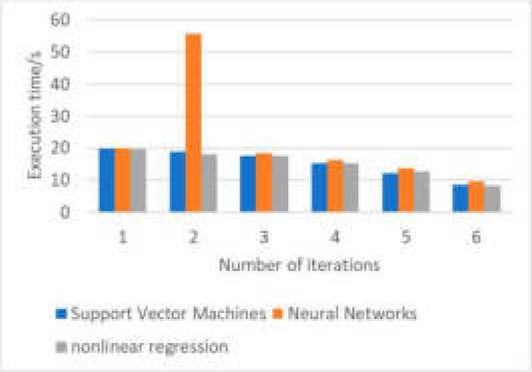


Fig. 4.5: Performance analysis of different algorithms on the S3 dataset.

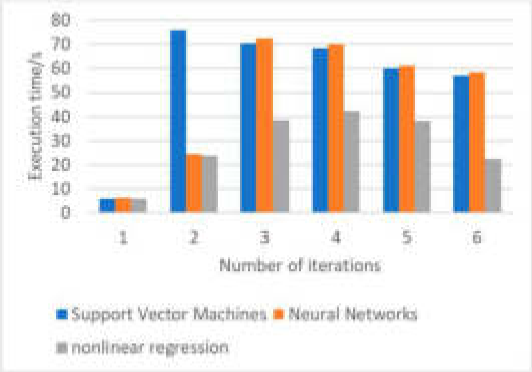


Fig. 4.6: Performance analysis of different algorithms on the S4 dataset.

shortened by using the Split Bloom Filter. This makes the performance of nonlinear data mining algorithms better than support vector machines and neural network algorithms.

5. Conclusion. This paper studies the efficient mining method of education-oriented user portrait data and integrates it with deep learning and adaptive learning to improve the information acquisition ability of education-oriented user portrait data. The user portrait data in the learning system is analysed in depth. The experiment proves the excellent applicability of this method. It is a scalable distribution algorithm. For large data sets of different sizes, the nonlinear data mining algorithm performs similarly or better than the support vector machine and neural network algorithms on the Spark platform.

REFERENCES

[1] Nitu, P., Coelho, J., & Madiraju, P. Improvising personalized travel recommendation system with recency effects. Big Data Mining and Analytics, 2021; 4(3): 139-154.

[2] Al Fararni, K., Nafis, F., Aghoutane, B., Yahyaouy, A., Riffi, J., & Sabri, A. Hybrid recommender system for tourism based on big data and AI: A conceptual framework. Big Data Mining and Analytics, 2021; 4(1): 47-55.

[3] Javed, U., Shaukat, K., Hameed, I. A., Iqbal, F., Alam, T. M., & Luo, S. A review of content-based and context-based recommendation systems. International Journal of Emerging Technologies in Learning (iJET): 2021; 16(3): 274-306.

[4] Singh, P. K., Pramanik, P. K. D., Dey, A. K., & Choudhury, P. Recommender systems: an overview, research trends, and future directions. International Journal of Business and Systems Research, 2021; 15(1): 14-52.

[5] Steck, H., Baltrunas, L., Elahi, E., Liang, D., Raimond, Y., & Basilico, J. Deep learning for recommender systems: A Netflix case study. AI Magazine, 2021;42(3): 7-18.

[6] Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., & He, X. Bias and debias in recommender system: A survey and future directions. ACM Transactions on Information Systems, 2023; 41(3): 1-39.

[7] Baczkiewicz, A., Kizielewicz, B., Shekhovtsov, A., Watróbski, J., & Sałabun, W. Methodical aspects of MCDM based E-commerce recommender system. Journal of Theoretical and Applied Electronic Commerce Research, 2021;16(6): 2192-2229.

[8] Deldjoo, Y., Noia, T. D., & Merra, F. A. A survey on adversarial recommender systems: from attack/defense strategies to generative adversarial networks. ACM Computing Surveys (CSUR): 2021; 54(2): 1-38.

[9] Joe, M. C. V., & Raj, D. J. S. Location-based orientation context dependent recommender system for users. Journal of Trends in Computer Science and Smart Technology, 2021; 3(1): 14-23.

[10] Wu, D., Shang, M., Luo, X., & Wang, Z. An L 1-and-L 2-norm-oriented latent factor model for recommender systems. IEEE Transactions on Neural Networks and Learning Systems, 2021; 33(10): 5775-5788.

[11] Ferrari Dacrema, M., Boglio, S., Cremonesi, P., & Jannach, D. A troubling analysis of reproducibility and progress in recommender systems research. ACM Transactions on Information Systems (TOIS): 2021; 39(2): 1-49.

[12] Huang, Z., Liu, Y., Zhan, C., Lin, C., Cai, W., & Chen, Y. A novel group recommendation model with two-stage deep learning. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2021; 52(9): 5853-5864.

[13] Yue, W., Wang, Z., Zhang, J., & Liu, X. An overview of recommendation techniques and their applications in healthcare. IEEE/CAA Journal of Automatica Sinica, 2021; 8(4): 701-717.

[14] Elbadawi, M., Gaisford, S., & Basit, A. W. Advanced machine-learning techniques in drug discovery. Drug Discovery Today, 2021;26(3): 769-777.

[15] Ageed, Z. S., Zeebaree, S. R., Sadeeq, M. M., Kak, S. F., Yahia, H. S., Mahmood, M. R., & Ibrahim, I. M. Comprehensive survey of big data mining approaches in cloud systems. Qubahan Academic Journal, 2021; 1(2): 29-38.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Nov 2, 2023
Accepted: Nov 20, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1923–1931, DOI 10.12694/scpe.v25i3.2792

CONTEXT-AWARE MUSIC RECOMMENDATION ALGORITHM COMBINING CLASSIFICATION AND COLLABORATIVE FILTERING

XIAOLING WU[*]AND GUODONG SUN[†]

Abstract. As an effective solution to the problem of information overload, personalized recommendations have received widespread attention in the music field. A context-aware music recommendation algorithm combining classification and collaborative filtering is proposed based on user context information. Firstly, the similarity analysis of the user situation is carried out. A preliminary list of recommended songs is obtained by collaborative filtering. The machine learning method is used to classify music in different scenes to get the preferences of music types in different situations. Finally, the recommendation list obtained by collaborative filtering is combined with the music type preference obtained by the classification model and personalized music recommendations for users in different situations. This algorithm not only effectively reduces the complexity of the recommendation process Experiments show that the proposed algorithm can effectively improve the accuracy of users' music recommendations.

Key words: Music recommendation; Personalized recommendation; Situational awareness; Improved Collaborative Filtering

1. Introduction. Music is a good pastime in human life. With the progress of science and technology, the development of music resources has reached an unprecedented height. A personalized recommendation is an effective way to alleviate the problem of "information overload." It has received more and more attention and application in the music industry. Major music platforms can now carry out personalized music recommendations, such as Spotify, Pandora, Douban Music, NetEase Music, etc. Many platforms have gained a reputation for recommending songs more accurately. Currently, most music recommendations are based on the user's usage habits to tap into the user's long-term preferences. However, the environment often affects users' short-term song preferences in song recommendations. Scenarios include people's psychological state, behavior, external environment and many other aspects [1]. Focusing only on users and recommended products will impact recommendation results. Take the user's environment into account in personalized music recommendations. It can make individual recommendations to users based on their situation.

By effectively utilizing a large amount of information in the context, personalized recommendations more in line with users can be achieved to improve its accuracy and user experience. This is the focus of scholars at home and abroad [2]. This paper uses multiple class models based on the classical collaborative screening method. Complete a scene-oriented music recommendation method. The traditional collaborative filtering algorithm can create situational awareness by integrating the similarity calculation method of user situational information. The fusion classification model improves the performance of the recommendation system, effectively reduces the complexity of the recommendation process and improves the cold start problem. After implementing this rule, users can be provided with a personalized song recommendation according to the user's actual situation.

2. Recommendation method based on collaborative screening. In essence, the learning recommendation system is realized through machine learning. Through the collection of students' basic information and analysis of students' learning habits, learning behaviors and test results, the external and internal learning needs of students are identified [4]. Students actively seek knowledge in the knowledge base to meet their needs—a dynamic generation of learning paths to facilitate learners to complete learning better.

2.1. Overall system framework. Collaborative screening is one of the most widely used methods at present. The algorithm can be divided into user and item-collaborative filtering [3]. The "nearest neighbor" is

[*] College of Music, Handan University, Handan, Hebei, 056000, China

[†] Students' Affairs Division, Hebei University of Engineering, Handan, Hebei, 056000, China (Corresponding author, sunguodong@hebeu.edu.cn)

closest to the target object and recommends the goods with "nearest neighbor preference" to it. The latter is the recommendation of items with similar historical interests to the user to the target user.

2.2. User recommendation methods in collaborative screening. Firstly, a new method based on the user-entry scoring matrix is proposed. Predict the evaluation of an item. Make practical information recommendations to specific users. Standard methods of user similarity calculation include the Pearson correlation coefficient method, cosine similarity method and improved cosine similarity method. Pearson correlation analysis is one of the most widely used:

$$sim(u,u\prime) = \frac{\sqrt{\sum\limits_{i\in I(u,u\prime)} \left|(r_{(u,i)} - \bar{r}_{(u)})(r_{(u\prime,i)} - \bar{r}_{(u\prime)})(r_{(u,i)}+\bar{r}_{(u)})(r_{(u\prime,i)}+\bar{r}_{(u\prime)})\right|}}{\sqrt{\left|\sum\limits_{i\in I(u,u\prime)} (r_{(u,i)} - \bar{r}_{(u)})^2\right|}\sqrt{\left|\sum\limits_{i\in I(u,u\prime)} (r_{(u\prime,i)} - \bar{r}_{(u\prime)})^2\right|}} \tag{2.1}$$

$r_{(u,i)}, r_{(u\prime,i)}$ indicates the evaluation of users u and $u\prime$ on item i. $\bar{r}_{(u)}$ and $\bar{r}_{(u\prime)}$ represent the average of the scores of users u and $u\prime$ for all entries. The nearest neighbor set D is generated to the target user based on similarity. Then, based on D, the user's score of the item is converted into the K-nearest neighbor score prediction formula (2.2) to predict the user's preference. The result is used as the basis for recommending the target users.

$$p(\varepsilon,i) = \bar{r}_{(\varepsilon)} + \frac{\sum\limits_{\varepsilon_t\in S} [sim(\varepsilon,\varepsilon_t)(r_{(\varepsilon t,j)} - \bar{r}_{(\varepsilon,t)})]}{\sum\limits_{\varepsilon_t\in S} sim(\varepsilon,\varepsilon_t)} \tag{2.2}$$

2.3. Recommended methods for item classification in collaborative screening. First, a similar group of items is found based on the score and solution process in formula (2.3). Record as project set I. This method predicts the target items and the information suitable for its characteristics is recommended.

$$sim(i,j) = \frac{\sum\limits_{u\in S_{(i,j)}} \sqrt{(r_{(u,i)} - \bar{r}_{(i)})(r_{(u,j)} - \bar{r}_{(j)})(r_{(u,i)}+\bar{r}_{(i)})(r_{(u,j)}+\bar{r}_{(j)})}}{\sqrt{\sum\limits_{s\in S_{(i,j)}} (r_{(u,i)} - \bar{r}_{(i)})(r_{(u,i)}+\bar{r}_{(i)})(r_{(u,j)} - \bar{r}_{(j)})(r_{(u,j)}+\bar{r}_{(j)})}} \tag{2.3}$$

$\bar{r}_{(i)}, \bar{r}_{(j)}$ is the average of all users' scores for items i and j. $S_{(i,j)}$ is the set of users. Equation (4) is obtained from the predicted user score ε in the nearest neighbor group I for entry i. Then give the corresponding product recommendation according to the predicted results.

$$p(\varepsilon,i) = \bar{r}_{(i)} + \frac{\sum\limits_{j\in I_{(i,j)}} [sim(i,j)(r_{(\varepsilon,j)} - \bar{r}_{(j)})]}{\sum\limits_{j\in I_{(i,j)}} sim(i,j)} \tag{2.4}$$

$I_{(i,j)}$ represents the set most closely related to item i.

2.4. Some critical issues about collaborative filtering.

Data sparsity problem. The recommended method of collaborative filtering is based on the user-entry score matrix. The higher the score density, the more accurate the similarity calculation and the higher the accuracy of its recommendation. However, due to the extreme sparsity of a large number of user-item evaluation samples, the efficiency of the recommendation algorithm is greatly affected. Literature [4] proposes an algorithm based on matrix and SVD to solve the problem of performance degradation caused by sparse data. The aim is to solve the sparse problem of the score matrix. Literature [5] adopted the clustering method to reduce the dimensionality of the score matrix to solve the problem of sparse data and improve the accuracy of recommendations.

Research on the recommended cold start problem. The cold start problem is when a new user or project is added to the recommendation system; because there is not enough object scoring data, the similarity of the target object will be difficult, failing to recommend the project to the new user. This problem will cause the user to lose confidence in the recommendation mechanism and thus be rejected by the user. At this point, the recommendation mechanism fails. The literature on the cold start phenomenon [6] proposes an algorithm based on n-sequence access analysis logic and most frequent item extraction based on the cold start problem of cooperative filtering. This eliminates the shortcomings of the user in the cold startup process. Literature [7] proposes a collaborative filtering algorithm based on a K-nearest neighbor-based attribute-feature graph. Literature [8] proposes a recommendation algorithm based on association rules, demonstrating its advantages in dealing with sparse data and cold start problems. Literature [9] designed the CUTA Time recommendation algorithm to solve the cold start problem of new projects based on user (project) attributes and rating date.

Real-time research of recommendation. In a recommendation system, the amount of recommendation computation increases with the increase of users and projects. At present, rapid and effective information recommendation is a serious challenge. At present, collaborative screening technology faces an urgent problem: improving its real-time performance while ensuring the accuracy of recommendations. Literature [10] proposes a collaborative screening algorithm that can enhance the real-time performance of recommendations and users' real-time feedback updates. Literature [11] designed a real-time recommendation system for collaborative filtering based on Spark distribution and demonstrated the system's reliability.

3. Discussion on personalized music situational cognition.

3.1. Research on personal style characteristics. Personalized music shows immediacy, situational dependence, mobility, and randomness of user needs. In the personalized music environment, the changes in users' time, space, social relations and other related situations make users' interests and needs change. The lives of music users are constantly changing. When the user's information needs to be timed, the situational factors will directly impact the user's interest tendency [12]. Environmental factors can be weather, seasons, time, space, etc. Mobile phone music's mobility, scene dependence and other characteristics make mobile phone music users will have more personal needs.

3.2. Overview of Situational Perception Knowledge.

Concept 1. $Z, Z = \{E_1, E_2, E_3, \cdots, E_n\}, Z \neq \delta$ scenario can be expressed as various types of non-empty sets a about users. E_j stands for item j in this text. n is the number of features in the scene. Case Z_i can be expressed in terms of $Z_i = \{E_{i1}, E_{i2}, E_{i3}, \cdots, E_{in}\}$. In a specific scenario Z_i, E_{ij} is the value of an attribute E_j. Situations can be divided into material situations and social situations.

Concept 2. Action scenario. Context attribute $Z_j = \{x|x = gZ_i(D)\}, D$ represents the variable of the user's location in the action context.

Concept 3. Situational awareness. Context awareness refers to the operation process in which the applied device can sense the context information, process the device's context information, and use the context data information. In fact, in a generalized computing environment, various applications of situation processing can be called situational cognition. Early scene cognition technology has been widely used in pervasive computing, data mining, information retrieval and other fields [13]. As a universal, flexible and personalized mobile scene, it has a broad application prospect in music, traffic navigation and tourism. Therefore, scene cognition service has become a profit point of music festivals.

3.3. Clustering algorithm. Clustering refers to dividing physical or virtual objects into groups of similar objects. So that objects in the same cluster have a high degree of similarity. The difference in goals is not the same in clustering various categories. An efficient clustering method based on K-means is proposed. Clustering entries and users in different scenarios obtain top-N suggestions.

Basic Model of Mobile Phone Situational Cognition Service. The platform collects many situational information, including current situational information, interest preferences, personal characteristics, etc. A customer-oriented personalized recommendation algorithm with high real-time performance and accuracy is proposed. Figure 3.1 shows the architecture of its standard architecture model [14]. The first layer is the mobile intelligent terminal interaction layer. The second layer is the mobile recommendation service layer. The third level is the data preprocessing layer. The fourth level is the data collection layer.

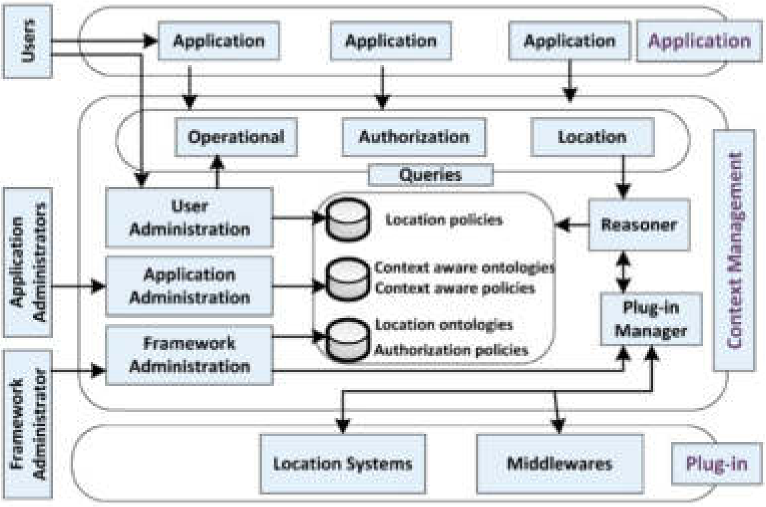


Fig. 3.1: Mobile situational awareness service model.

Context-aware user-item cluster collaborative screening. The traditional collaborative filtering recommendation algorithm mainly considers two dimensions: user and project. However, in the individual music behavior, the user's situation constantly changes, and the user's needs also change. Moreover, the calculation scale of the nearest neighbor also increases linearly with the expansion of the personalized music scale, resulting in a continuous decline in the recommendation accuracy [15]. This project proposes a collaborative screening strategy based on mobile context awareness and clustering. The collaborative filtering method of user-commodity clustering under scene perception is studied. Thus, it can better meet users' personalized requirements for personal music.

Build personalized music, user-project-context model. We adopt personalized music 3D $S-I-Z$ (User, Account, Events) mode. This study aims to more clearly present users' preferences for music items in various contexts (Figure 3.2). The vertical axis is the user dimension [16]. The horizontal axis represents the dimensional dimension of the music entry. Z represents a specific characteristic component of the situation (season, location, weather, occupation, gender, interests, etc.). Coordinate refers to the user S_i score value of I_j in the specific scenario Z_t is $g(S_i, I_j, Z_t)$. A context-based user-entry score matrix $H : S \times I \times Z$ is established.

Scenario-based User-Project clustering. Under the premise of thoroughly combining the scene elements, the user-project clustering operation is realized—offline clustering groups users and items with highly similar situations into the same category.

Determination of scenario similarity. Contextual information is diverse, such as the user's name, gender, online time, hobbies, etc. Specific scene information includes season, weather, time, temperature, location, etc. All the assumed data are quantified as numerical to facilitate the calculation. Due to the complexity and diversity of scenes, the conventional method of scene similarity has been unable to adapt, so the calculation of scene similarity can be carried out by formula (3.1).

$$sim(Z_i, Z_j) = \frac{\sum\limits_{v=1}^{m} \lambda_{ij}^{v} \times \kappa_{ij}^{v}}{\sum\limits_{v=1}^{m} \lambda_{ij}^{v}} \tag{3.1}$$

$\lambda_{ij}^{v}, \kappa_{ij}^{v}$ is the indicator function. If the $v$variable in Z_i, Z_j does not appear, it is $\lambda_{ij}^{v} = 0$; otherwise, it is $\lambda_{ij}^{v} = 1$. If v variables in Z_i, Z_j are $\kappa_{ij}^{v} = 1$, then $\kappa_{ij}^{v} = 1$. The closer the $sim(Z_i, Z_j)$ value is to 1, the higher the similarity is, and the lower the similarity is.

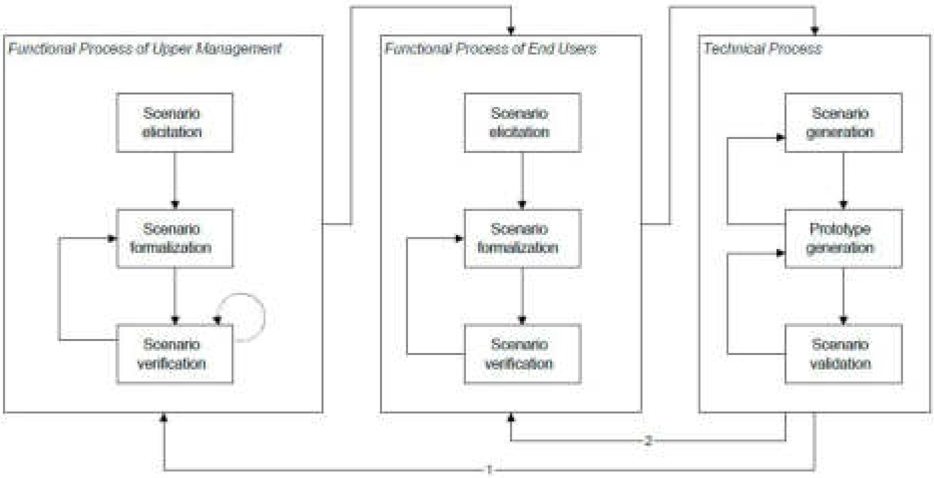
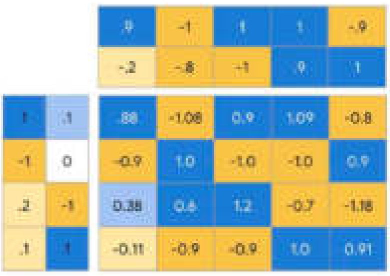


Fig. 3.2: User-project-scenario model.



		.9	-1	1	1	-.9
		-.2	-.8	-1	.9	1
1	.1	.88	-1.08	0.9	1.09	-0.8
-1	0	-0.9	1.0	-1.0	-1.0	0.9
.2	-1	0.38	0.6	1.2	-0.7	-1.18
.1	1	-0.11	-0.9	-0.9	1.0	0.91

Fig. 3.3: Steps to create a user preferences table.

Constructing User Item Type Preference Scoring Matrix (UPM). The score of an item can directly reflect the user's preference for the same category item. Here, the initial score matrix $H : S \times I \times Z$ can be reconstructed by formula (6). The evaluation matrix of various items by users containing contextual information is established, which is called the "user item category preference scoring matrix".

$$F(S_i)_{Ox} = \sum_{i \in IS_i, x, Z_t} r_{ui}, Z_t / |IS_i, Ox, Z_t| \tag{3.2}$$

I_u, x, Z_t represents the collection of items with characteristic x that are evaluated by the user in ScenarioZ_t. r_{ui}, Z_t is user u real evaluation of item i in the specific situation Z_t. $|I_u, x, Z_t|$ represents how many elements there are in the group I_u, x, Z_t. Start by reading the original score matrix and the relevant information about the categories of entries. Then, the UPM matrix is formed by extracting the user's interest tendency toward the commodity category from (3.2). Figure 3.3 illustrates a simple UPM generation process. For items that are not scored, the data is marked as 0 in the scoring matrix H_1. In the entry type matrix H_2, when the entry has a particular property $O\chi$, its corresponding entry represents 1, and vice versa. The value of the User Presence Matrix (UPM) H_3 is calculated by (3.2).

K-class clusters are obtained by the IC-KM method. K The appropriate value should be chosen so that the collection of items can be better classified rationally. Ensure all items entering the same cluster have similar scores [17]. Filling within the same cluster can effectively eliminate the influence of external indicators on the clustering results and minimize the error rate of the clustering results. This facilitates the filling of the sparse matrix for subsequent scoring.

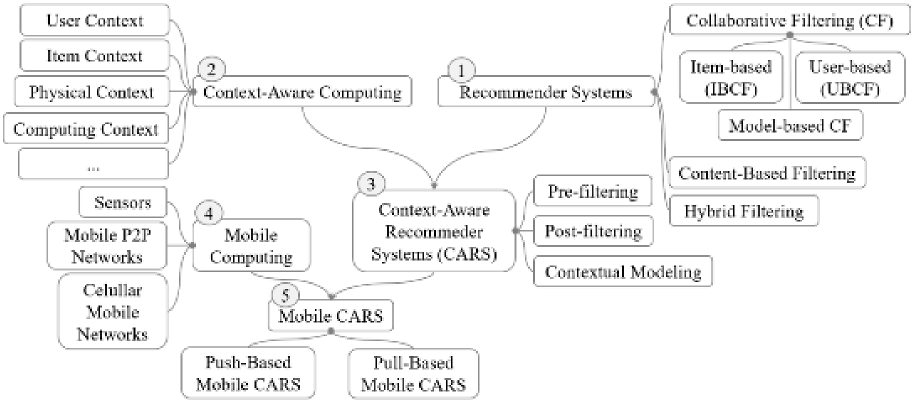


Fig. 3.4: Implementation flow of UIC-collaborative screening algorithm.

Complete user-entry score matrix. A weighted Slope1 algorithm is developed to fill the score matrix according to the similarity between entries. First, the mean of each item is calculated according to the mean in formula (3.3), and the similarity between the items is calculated according to the mean in formula (2.3). The value is then calculated by the weight of the formula (3.4). All the score values missing from the score matrix are obtained in order. Produces A new scoring matrix $H\prime : S \times I \times Z$ with no vacancies. Its scene composition has not changed.

$$dev_{ij} = \sum_{u \in S_{ij}} \frac{r_{ui} - r_{uj}}{sum(S_{ij})} \tag{3.3}$$

$$F(u)_i = \frac{\sum\limits_{j \in I_i} [(dev_{ij} + r_{ui}) \times sum(S_{ij}) \times sim(i,j)]}{\sum\limits_{j \in I_i} [sum(S_{ij}) \times sim(i,j)]} \tag{3.4}$$

S_{ij} is the set of users that evaluate i and j. r_{uj} is the score for item j given by user u. dev_{ij} is the mean of each index i, j. $sum(S_{ij})$ is used to evaluate the number of users for items i and j.

3.4. Context-based user-item cluster recommendation method. This paper presents a user-commodity cluster recommendation method in the context. The volume flow is shown in Figure 3.4 (image cited in Knowledge-Based Systems, 2021, 215:106740).

4. Experimental results and analysis.

4.1. Experimental data. In this paper, 80 volunteers collected music background information daily to test. Experiments and theoretical analysis verify the scenario-based personalized recommendation method. These subjects were wearing special sensors. Collect the user's heart rate, exercise status, and other information, and record the user's situation information in various scenarios. Try to ensure that all categories of background information are covered. The goal is to take full advantage of the value of these features. There are 40,000 records in the original data collected. The complete, clean, preprocessed dataset contains 38,600 records. Finally, the effectiveness of the proposed algorithm is verified by ten cycles of experiments.

4.2. Evaluation Indicators. The evaluation index of a personalized recommendation algorithm cannot be completely equivalent to that of the classification algorithm. This is because similar criteria, such as accuracy, can evaluate the method. In the evaluation system, its evaluation indicators are more diverse. The central performance is correct rate, recall rate, average absolute difference, diversity, surprise degree, etc. Too much accuracy will result in a smaller number of selected items. Increasing the diversity and surprise of the

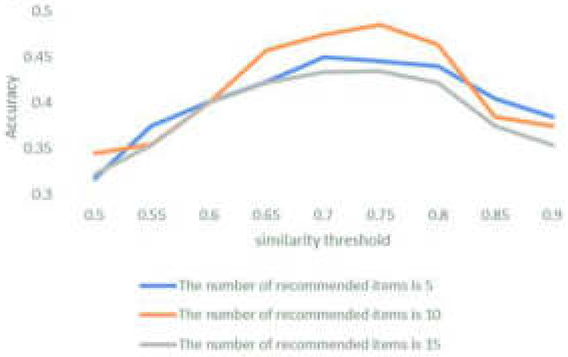


Fig. 4.1: Accuracy of the cooperative screening model.

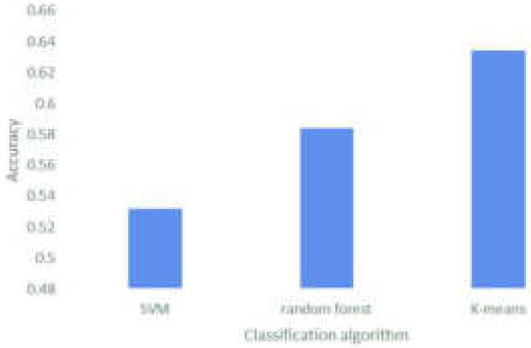


Fig. 4.2: Accuracy of classification model.

recommendation system will inevitably reduce its accuracy. Through the analysis of user satisfaction, a more comprehensive conclusion is drawn.

Accuracy refers to the percentage of recommended lists marked as favorites by the user. $T(u)$ represents the list of associations marked as preferences by the user. $R(u)$ represents the list of suggestions given, then the method of calculating accuracy is

$$\Pr ecision = \frac{\sum\limits_{u \in U} (R(u) \bigcap T(u))}{\sum\limits_{u \in U} |R(u)|} \tag{4.1}$$

In addition, 50 volunteers were tested under different recommendation algorithms to get the evaluation of the system. The user comments on the list of suggestions. The average of these ratings indicates the user's satisfaction with the suggestion system. In addition, based on the feedback information from users with clear goals, the whole recommendation system can be further improved.

4.3. Experimental results and analysis. In the process of collaborative screening, it is necessary to understand the influence of the threshold of user story similarity on the screening results. When the accuracy of the collaborative filtering model is tested according to the size of the threshold, the accuracy results are obtained (Figure 4.1).

Experiments verify the correctness of the algorithm. The average accuracy of the forecast for the classification model is shown in Figure 4.2. Through the analysis of experimental data, it is concluded that both random forest and K-means methods have greatly improved the prediction accuracy.

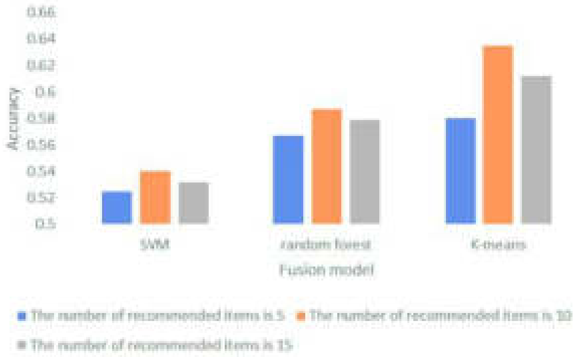


Fig. 4.3: Accuracy of the fusion model.

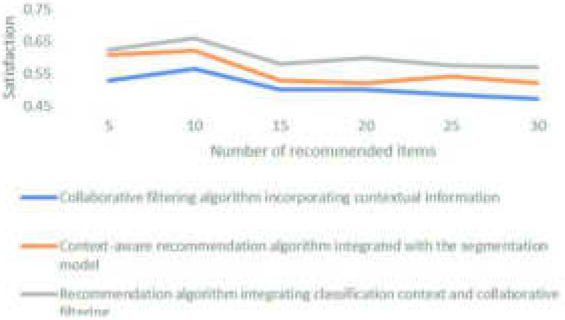


Fig. 4.4: Comparison of user satisfaction.

The precision of this fusion pattern is shown in Figure 4.3. Figure 4.4 shows a comparison of user satisfaction levels. The results show that the accuracy of the method has improved significantly. When K-mean is selected, and the number of recommended items is 10, the accuracy of this method is the best. Compared with the support vector machine method, the accuracy of this method is significantly improved. At the same time, the result obtained by this method is the most satisfactory. The feasibility and effectiveness of this method are demonstrated.

5. Conclusion. Context information is an essential factor affecting the effect of personalized music recommendations. A context-aware music recommendation method is designed by combining situational information and collaborative filtering. Experiments show that combining contextual information and collaborative filtering can obtain higher accuracy and user satisfaction. However, the research on this subject needs more expansion and deepening due to the limitation of the data used and the background data included. Only the music type label is used in the classification model, but no more profound analysis of its speech characteristics is carried out, which has a particular impact on the performance of the recommendation algorithm.

Acknowledgments. This work was supported by Research project of Humanities and Social Sciences in Colleges and universities of Hebei Provincial Department of Education: Research on the inheritance of local music in local universities (NO.SY2022099).

REFERENCES

[1] Zhou, K., Yang, C., Li, L., Miao, C., Song, L., Jiang, P., & Su, J. A folksonomy-based collaborative filtering method for crowdsourcing knowledge-sharing communities. Kybernetes, 2023;52(1): 328-343.

[2] Lv, Z., & Song, H. Trust mechanism of feedback trust weight in multimedia network. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2021; 17(4): 1-26.

[3] Liu, K., Xue, F., He, X., Guo, D., & Hong, R. Joint multi-grained popularity-aware graph convolution collaborative filtering for recommendation. IEEE Transactions on Computational Social Systems, 2022; 10(1): 72-83.

[4] Li, L., Wang, Z., Li, C., Chen, L., & Wang, Y. Collaborative filtering recommendation using fusing criteria against shilling attacks. Connection Science, 2022; 34(1): 1678-1696.

[5] Vahidnia, M. H. Point-of-interest recommendation in location-based social networks based on collaborative filtering and spatial kernel weighting. Geocarto international, 2022; 37(26): 13949-13972.

[6] Ilangovan, P. Support Vector Machine based a New Recommendation System for Selecting Movies and Music. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 2021;12(10): 1425-1429.

[7] Afchar, D., Melchiorre, A., Schedl, M., Hennequin, R., Epure, E., & Moussallam, M. Explainability in music recommender systems. AI Magazine, 2022;43(2): 190-208.

[8] Papadakis, H., Papagrigoriou, A., Panagiotakis, C., Kosmas, E., & Fragopoulou, P. Collaborative filtering recommender systems taxonomy. Knowledge and Information Systems, 2022;64(1): 35-74.

[9] Wu, L., He, X., Wang, X., Zhang, K., & Wang, M. A survey on accuracy-oriented neural recommendation: From collaborative filtering to information-rich recommendation. IEEE Transactions on Knowledge and Data Engineering, 2022; 35(5): 4425-4445.

[10] Yannam, V. R., Kumar, J., Babu, K. S., & Sahoo, B. Improving group recommendation using deep collaborative filtering approach. International Journal of Information Technology, 2023; 15(3): 1489-1497.

[11] Liu, L. The artistic design of user interaction experience for mobile systems based on context-awareness and machine learning. Neural Computing and Applications, 2022;34(9): 6721-6731.

[12] Borges, R., & Stefanidis, K. Feature-blind fairness in collaborative filtering recommender systems. Knowledge and Information Systems, 2022; 64(4): 943-962.

[13] Zitouni, H., Meshoul, S., & Mezioud, C. New contextual collaborative filtering system with application to personalized healthy nutrition education. Journal of King Saud University-Computer and Information Sciences, 2022;34(4): 1124-1137.

[14] Assuncao, W. G., Piccolo, L. S., & Zaina, L. A. Considering emotions and contextual factors in music recommendation: a systematic literature review. Multimedia Tools and Applications, 2022; 81(6): 8367-8407.

[15] Yao, K., Wang, H., Li, Y., Rodrigues, J. J., & de Albuquerque, V. H. C. A group discovery method based on collaborative filtering and knowledge graph for IoT scenarios. IEEE Transactions on Computational Social Systems, 2021; 9(1): 279-290.

[16] Horasan, F. Latent Semantic Indexing-Based Hybrid Collaborative Filtering for Recommender Systems. Arabian Journal for Science and Engineering, 2022; 47(8): 10639-10653.

[17] Ghosh, S., Tyagi, D., Vashisht, D., Yadav, A., & Rajput, D. A comprehensive survey of personalized music identifier system. The International Journal of Recent Technology and Engineering (IJRTE), 2021; 9(6): 90-96.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Nov 3, 2023
Accepted: Nov 20, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1932–1939, DOI 10.12694/scpe.v25i3.2794

HIGH-PERFORMANCE COMPUTING WEB SEARCH SYSTEM BASED ON COMPUTER BIG DATA

YINGXI KANG[*], BEIPING TANG, AND XIAODONG HU

Abstract. File sharing, streaming media, collaborative computing, and other P2P systems are all unicast to establish the corresponding overlapping network. The superimposed network is generally carried out based on the existing primary network. In this way, the access of each node is random. At the same time, this will cause the topological structure of the upper and lower layers to be inconsistent. This will increase the communication delay between nodes and cause an excellent bandwidth burden to the underlying network. The existing topology matching methods still face problems, such as poor scalability and long node aggregation time. This paper aims to design a topological distributed node aggregation method based on network coordination and distributed hash table (DHT) algorithm. This paper established a two-dimensional mesh model of nodes based on equal-distance concentric circles and divided into two equal areas. The parts of multiple namespaces correspond one by one according to their location. Because nodes are kept close, neighbours can be aggregated through DHT's primary "publish" and "search" primitives. Experimental results show that the TANRA method can match the network's topology under a slight delay and a large number of nodes. The TANRA method can effectively reduce the path delay in structured networks.

Key words: Topological induction; Node proximity; Topology aware node aggregation algorithm; Node cluster; Distributed hash table; Overlay Network

1. Introduction.

Peer-to-peer (P2P) technology has been widely concerned with its excellent scalability and fault tolerance. P2P networks can make full use of the node resources of the network edge system. It is getting more and more attention in the new wave of applications on the Internet. Especially in mass content publishing and streaming media, the resource sharing of intermediate nodes can effectively reduce the consumption of network resources where the data source resides [1]. Therefore, highly concurrent content distribution and media transmission can be achieved. In P2P mode, there are a series of streaming media technologies such as PROMISE and Cool streaming. Although the above systems have good scalability and support ability for high concurrency services, the lack of corresponding topological identification methods causes the topological structure of the upper-layer overlay network and the lower-layer communication network to be inconsistent. In this case, the nodes in the network cannot exchange information, which makes the data transmission efficiency in the high-level network low and affects the overall performance. At the same time, it also brings a lot of bandwidth burden to the underlying physical network. Studies show that P2P has accounted for 60% of the business on the Internet in recent years. Therefore, reducing the occupation of P2P networks under the premise of ensuring the quality of network service is an urgent problem that needs to be studied [2]. Therefore, a highly scalable node aggregation algorithm is designed in this paper. Then, a topologically aware distributed node aggregation algorithm (TANRA) based on the network coordinate algorithm and DHT algorithm is proposed. Unlike other methods, the proposed method can achieve topology consistency between the global and underlying networks. At the same time, it can converge quickly and increase the overhead of nodes. This scheme is suitable for many node configurations in P2P networks.

2. HPC web search system framework.

The architecture of the proposed in-site search system is shown in Figure 2.1 (The picture is quoted from Example of system architecture picture in TikZ). The system consists of two parts: index sub and search sub. It mainly completes the index creation, increment, update, delete, and other functions in the Xapian database and queries it according to users' requirements. Interact with the Xapian database through a web interface. Xapian has a separate index database. It is independent

[*] Hunan Institute of Engineering, Xiangtan, 411104 Hunan, China (Corresponding author, 06158@hnie.edu.cn)

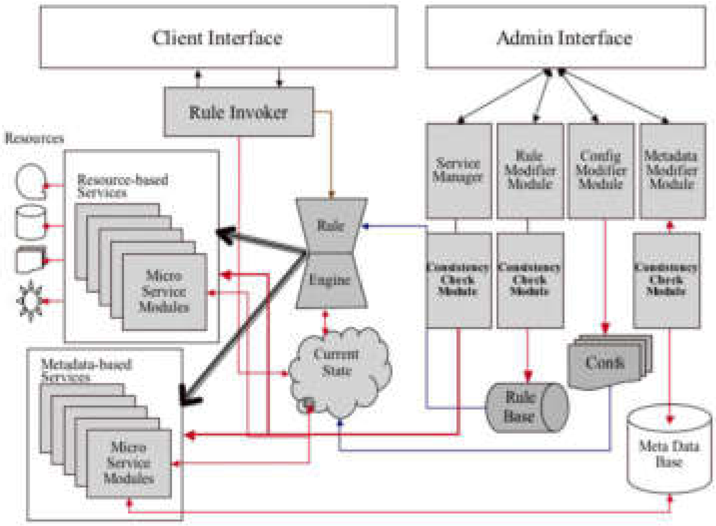


Fig. 2.1: Architecture diagram of the site finding system.

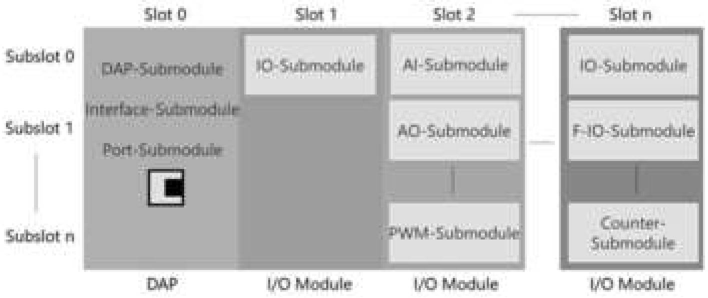


Fig. 2.2: Schematic diagram of index submodule.

of the site's database [3]. This is beneficial for strengthening the independence and flexibility of the Xapian website. An effective large-scale data query system is established based on adequate information query.

2.1. Index submodule. A lot of new data is generated on the site every day. The metrics are not the same. These metric changes when the site is upgraded. To improve the running speed of the system and improve the user experience of the system, the author also designed a site-independent index submodule. The component's architecture is shown in Figure 2.2 (the image is referenced).

The index is created to read the Web page data from the site's database. For the title of the page and the content, call the segmentation module, and then get the segmentation words and create a Xapian file for each page. Each Xapian file has a unique file ID number [4]. Store the split word as a Term in this file. In addition, the Values structure of the file also retains related data such as city ID number and user name so that the file can be easily and quickly filtered, sorted, and deleted duplicate items. This is the relay data attached to the file. Its purpose is fast access to matches, matching, and filtering. Some data in the file can be used to store any data. The system stores web page titles and addresses in the file data module.

When the first index is created, it must be updated periodically. Update the index. These include invalid index removal, changed index updates, and index additions for new pages. When the site data constantly changes, the site and indicator data may change. Some pages have been removed, so the files must be removed from the established database. There are also web pages to be modified, which requires searching the corre-

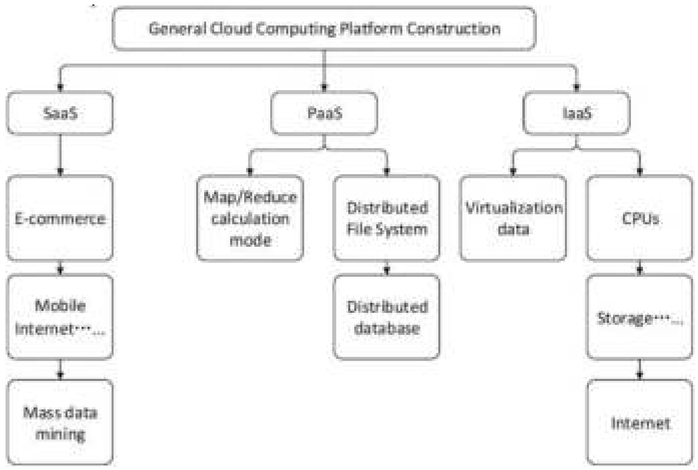


Fig. 2.3: Frame diagram of retrieval submodule.

sponding index library [5]. There are also many newly created web pages. This requires a new index to be created for each new page. Crontab is used to access delta Update under Linux and the site database under Linux. Add, delete, and modify existing databases based on changes in site data. This ensures that the index is consistent with the data on the site.

2.2. Searching for Sub-components. Retrieval submodule is one of the main modules of website search. Conditional queries can be made to the site. Figure 2.3 is a structural diagram of the retrieval subsystem. This article completes an interface similar to standard search engines like Google. Users can type in keywords to search and specify the search criteria [6]. The extraction submodule extracts the keywords entered by the user. Use the segmentation module to divide keywords into several words. According to the user input, the standard Boolean query is constructed. Use collection operations to filter the obtained data. The index database must do further work when a matching set of records is retrieved. This involves deleting duplicate records and classifying records according to time, importance, and other factors. The search results are then returned to the network interface as a list.

2.3. Word segmentation module. Chinese segmentation technology is to divide Chinese into several independent characters. The segmentation module is needed to subdivide the segmented text in the information query, whether in constructing an index or the query [7]. Since Xapian cannot use Chinese automatic segmentation, third-party Chinese automatic segmentation software must be used. There are three kinds of Chinese automatic segmentation methods: automatic segmentation based on a string, automatic segmentation based on understanding and automatic segmentation based on statistics. Currently, Chinese automatic machine-cutting technology mainly includes SCWS, Fudan NLP, ICTCLAS, HTTPCWS, etc. This paper introduces a lexical-based machine Chinese automatic machine segmentation system. Fudan NLP is a software designed by the Java company to support Chinese NLP. ICTCLAS is an open-source Chinese automatic machine independently developed by the Chinese Academy of Sciences [8]. It won the first prize in the National 973 Project evaluation. The splitting speed is about 500 KB/s. The segmentation accuracy reaches 98.45%. HTTPCWS is an open-source automatic Chinese text-cutting system based on HTTP. Its kernel uses ICTCLAS3.02009 to split it.

3. Topology distributed node aggregation method. Firstly, each node's network coordinates are obtained using the distributed network coordinates calculation method based on Vivaldi. Place nodes in a 2D plane (Figure 4). The Euclidean distance can represent the network delay between nodes [9]. The following form of description is taken:

Set the source node to u. ε_i is u concentric circle centered on A. Form cluster $P = \{\varepsilon_i || i = 1, 2, \ldots, n\}$ of concentric circles. Let l_i be the radius of ε_i.

Definition 1. The radii of adjacent concentric circles $\varepsilon_i, \varepsilon_{i-1}$ centered on u are respectively l_i, l_{i-1}. Let's call $l_i - l_{i-1}$ the distance between $\varepsilon_i, \varepsilon_{i-1}$.

Definition 2. Cluster of concentric circles centered on u is cluster of concentric circles centered on $P = \{\varepsilon_i || i = 1, 2, \ldots, n\}$. If $l_i l_{i-1} {=} l_{i-1} - l_{i-2} = \ldots = l_2 - l_1 = l, l$ is the radius of the inner ring ε_1 of the concentric circle. P is called isometric cluster of concentric circles. P is used to divide the two-dimensional plane of the node [10]. If the innermost circular surface area enclosed by ε_i is and the torus area enclosed between G_1, ε_i and ε_{i-1} is $G_i, i \geq 2$. R_{G_i} is the area of G_i. There is the following lemma:

Lemma 1. $\forall \varepsilon_i \in P, i \geq 2$. Make multiple centripetal lines from the point on ε_i to the center of the circle intersecting ε_{i-1}. The series centripetal lines divide the torus region $G_i, i \geq 2$ into j degrees. So, get $\{G_i = \bigcup g_{i,t} | i = 1, 2, \cdots, n, t = 1, 2, \cdots, j\}$. If $R_{g_{i,t}} = R_{g_{i-1,t}}, i > 2, R_{g_{2,t}} = R_{G_1}$ then $j = 2i - 1$.

Proof: Because$R_{G_i} = \pi(il)^2 - \pi[(i-1)l]^2 = \pi l^2(2i-1)$ and

$$\begin{array}{l} R_{G_i,t} = R_{G_{i-1,t}} = \cdots = R_{G_2,t} = R_{G_c} = \pi l^2, \\ R_{G_i} = jR_{G_{i,t}} = jR_{G_1} = j\pi l^2 \end{array},$$

so $j = 2i - 1$.

You've obtained the certificate. $\forall \varepsilon_i \in P, i \geq 2, \varepsilon_i$ is equally divided by $2i - 1$ arcs. If $\varepsilon_i = \bigcup_{t}^{2i-1} arc_{i,t}, arc_{i,t}$ is an t arc over ε_i, then the length of $arc_{i,t}$ when $i \to \infty$ is πl. l is the inner radius in P.

Proof: First find the length of each arc onε_i after $2i - 1$ equal division

$$\begin{array}{l} c_t, t = 1, 2, \ldots, 2i-1 : c_t = 2\pi \cdot il/2i - 1 \\ \lim\limits_{i \to \infty} c_t = 2\pi l \cdot \lim\limits_{i \to \infty} (i/2i - 1) = \pi l \end{array}$$

proof.

According to Lemma 1, if the torus G_i of P concentric cluster A is to be divided in half so that the area of each subring is equal to the region of the inner ring, then the torus should be divided into odd degrees of equality. The coordinate system of a 2D grid is divided into $1 + 3 + \cdots + 2i - 1 = i^2$ subfields according to the area of equal height. The distance difference l between each subring's outer and inner circle of each subring is found [11]. Each node in the network is positioned according to a particular position under given conditions. Lemma 2 gives that the outer arc length of any subregion converges to πl on the boundary of E. It is also necessary to split the maximum distance between any two points on the subring at $i \to \infty$. There are a couple of theorems here.

Theorem 1. $\forall g_{i,t} \in G_i, i \geq 2$ and $\exists n_1, n_2 \in g_{i,t}, t = 1, 2, \ldots, 2i-1, n_1, n_2$ are points. Define σ as the longest line between n_1, n_2. In this case, $\sigma = l\sqrt{1 + \pi^2}, l$ is the radius of the inner ring in the circle, $i \to \infty$.

4. Performance analysis and experimental simulation. In this part, the TANRA algorithm is verified by a node simulation experiment. Select GT-ITM as the tool for network topology generation. They are using Waxman's random graph model. These parameters are set to alpha and beta of 0.5 and 0.5, respectively. The topological nodes are organized according to the hierarchy of transfer roots, and a delay is added to the edge of the nodes. In total, there are 27 transmission nodes and 436 Stub topology nodes. The boundary delay between the regions is equally spaced along the interval [10,15]. The edge delay between the transmission and Stub areas is equally spaced along the area [40,80]. There is no boundary connection between root domain names. The connection delay between topological nodes in each Stub area is equally spaced along [10,30]. The average degree of each topology node calculated by GT-ITM is 3.55. The number of terminal nodes is gradually increased during the test. The maximum number of terminal nodes can be increased to 10,000.

Zipf assigns the number of terminals occupied by topological nodes. The number of adjacencies of each end node conforms to the Zipf assignment [12]. In each trial, nodes were randomly selected as source nodes. Semantic Routing Improvement algorithm (SCSRAA), DHT algorithm and TANRA algorithm are used to compare the performance of the algorithms. Both SCSRAA and TANRA use Bamboo as a DHT routing algorithm. The DHT algorithm uses a hierarchical search method for multiple adjacent nodes. The number

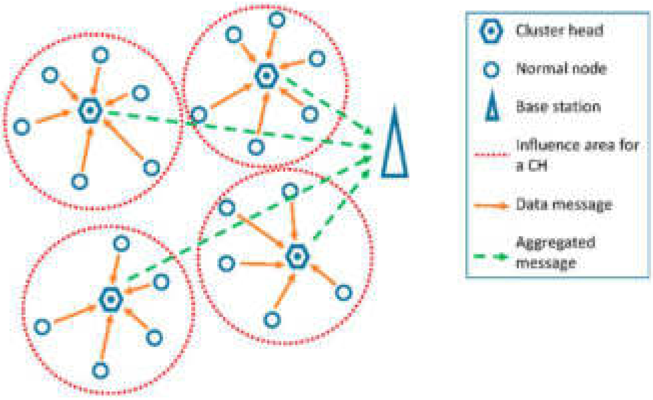


Fig. 3.1: Regional division.

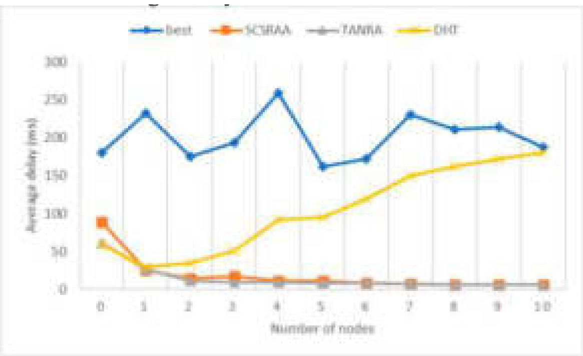


Fig. 4.1: The average latency for the number of adjacent nodes following Zipf (10).

of queries is the same as the previous two methods to ensure openness [13]. Unless otherwise specified, the number of searches is set to 10.

Figure 4.1 shows the average delay in assigning nodes according to Zipf (10) with the number of adjacent nodes at a layer spacing of 30 ms. When the number of access terminal nodes increases, the delay distribution of the SCSRAA system shows prominent fluctuation characteristics. TANRA has good delay convergence. When the number of nodes in the network exceeds 2000, the calculated result is very close to the minimum delay of the network [14]. When the number of networks is less than 1000, the DHT algorithm only exchanges 100 adjacent at a time. Its search algorithm has a significant effect on reducing the average delay. However, when the number of nodes in the network increases, the information interaction between neighbours is insufficient, increasing the network's average delay.

Figure 4.2 shows the average latency of nodes when the number of neighbouring nodes meets the Zipf (40) assignment. With the increase in the number of adjacent nodes, the delay of the TANRA network gradually decreases and tends to the best state [15]. However, SCSRAA delays show significant fluctuations. The DHT algorithm adds 40 neighbours, increasing the system's communication overhead. The average latency increases slightly more slowly than in Figure 5. Still, it continues to rise. When all nodes know only local messages, the network's performance depends on the number of communications between neighbouring nodes and the number of nodes in the entire network. This is consistent with our experimental data.

The introduction of TANRA into the overlay network enables the transformation from 2D to multi-level namespaces. This makes it possible for two adjacent nodes to be split into two sectors corresponding to adjacent segments. This article calls it mismatching [16]. This will affect the success rate of topology matching.

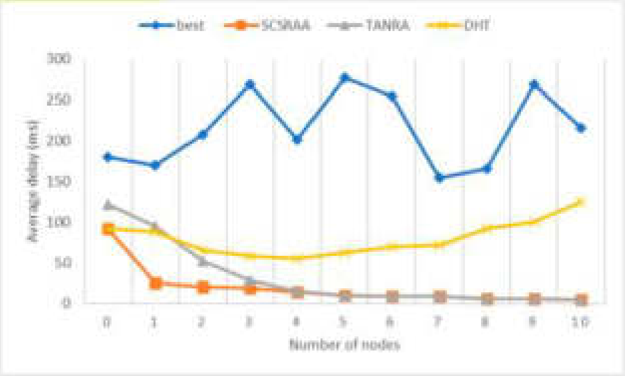


Fig. 4.2: Average latency for the number of adjacent nodes following Zipf (40).

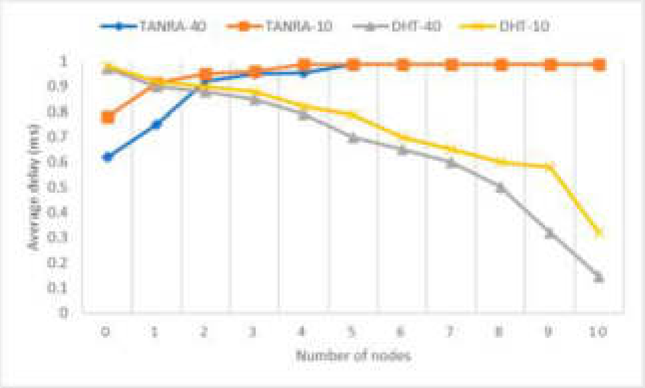


Fig. 4.3: Probability of success for topology comparison.

Figure 4.3 shows the success rate of pairing with a different number of neighbours at a stratified interval of 10 milliseconds. In the case of 1000 nodes, the algorithm's success rate is 65.62% when the neighbour ratio assigned by Zipf (10) is met. At 4000 nodes, the efficiency of the system has reached 98.48%. When the number of nodes in the network meets Zipf (40) allocation and the number of nodes reaches 1000, the transmission success rate of the network can reach 82.29%. In the case of more than 3000 nodes, the success rate of topology comparison of the network is more than 99%. Experimental results show that the TANRA method can effectively improve the topology matching rate of the network when there are a large number of nodes [17]. When only ten adjacent messages interact, the matching rate of the DHT algorithm decreases with the increase in the number of nodes in the network. The reason is insufficient information exchange between adjacent nodes in this method.

Figure 4.4 shows the node has entered the system and 40 neighbour nodes are found in the node and the average latency of the node in the node with a layer spacing of 30 ms. SCSRAA proposes a hierarchical addition method in the case of sparse networks. But when the number of nodes in the network increases, the network moves to a deeper level [18]. This leads to a delay in joining. The TANRA algorithm shows the opposite property. When the network becomes sparse, the nodes have a high probability of no node registration in the region, and their joining delay is relatively high. However, when the number of nodes in the network increases, subsequent nodes join the initial interval with a higher probability. Its additional delay tends to decrease on the whole.

5. Conclusion. In this paper, a distributed node aggregation method, TANRA, based on network coordination and DHT, is designed. The method uses concentric ring clustering with equal spacing to divide the

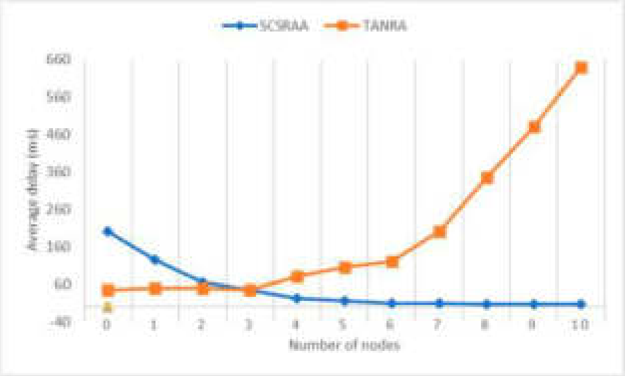


Fig. 4.4: Average delay of increased delay nodes.

space and the area of two nodes on a 2D grid. Then, a local partition method based on multi-level namespaces is proposed. It maintains proximity in each node. In DHT, the two fundamental elements of ”publish” and ”search” are used to add new nodes and find neighbouring nodes. Simulation experiments show the effectiveness of the method. The TANRA method can effectively ensure the consistency of network topology in the case of many nodes while reducing the addition delay. This project will integrate this method with Mesh technology to establish a no-structure overlapping network model to overcome the mismatch problem in the TANRA method. At the same time, the TANRA method is used to optimize the adjacency route in the structured network to reduce the path delay.

Acknowledgments. The authors acknowledge the : This article is a phased research achievement of the 2023 project of Hunan Province’s Education Science ”14th Five Year Plan”: Analysis and Intervention Research on Online Learning Based on Education Big Data Mining (project number: XJK23BXX002).

REFERENCES

[1] Bohu, L., Lin, Z., & Xudong, C. Introduction to cloud manufacturing. Zte Communications, 2020;8(4): 6-9.

[2] Fatemidokht, H., Rafsanjani, M. K., Gupta, B. B., & Hsu, C. H. Efficient and secure routing protocol based on artificial intelligence algorithms with UAV-assisted for vehicular ad hoc networks in intelligent transportation systems. IEEE Transactions on Intelligent Transportation Systems, 2021; 22(7): 4757-4769.

[3] Murshed, M. S., Murphy, C., Hou, D., Khan, N., Ananthanarayanan, G., & Hussain, F. Machine learning at the network edge: A survey. ACM Computing Surveys (CSUR), 2021; 54(8): 1-37.

[4] Mahmood, A., & Wang, J. L. Machine learning for high performance organic solar cells: current scenario and future prospects. Energy & environmental science, 2021; 14(1): 90-105.

[5] Liu, Y., Yuan, X., Xiong, Z., Kang, J., Wang, X., & Niyato, D. Federated learning for 6G communications: Challenges, methods, and future directions. China Communications, 2020; 17(9): 105-118.

[6] Lao, L., Li, Z., Hou, S., Xiao, B., Guo, S., & Yang, Y. A survey of IoT applications in blockchain systems: Architecture, consensus, and traffic modeling. ACM Computing Surveys (CSUR), 2020; 53(1): 1-32.

[7] Zhang, Q., Xin, C., & Wu, H. Privacy-preserving deep learning based on multiparty secure computation: A survey. IEEE Internet of Things Journal, 2021; 8(13): 10412-10429.

[8] Lin, X., Wu, J., Mumtaz, S., Garg, S., Li, J., & Guizani, M. Blockchain-based on-demand computing resource trading in IoV-assisted smart city. IEEE Transactions on Emerging Topics in Computing, 2020; 9(3): 1373-1385.

[9] Ma, Y., Wang, Z., Yang, H., & Yang, L. Artificial intelligence applications in the development of autonomous vehicles: A survey. IEEE/CAA Journal of Automatica Sinica, 2020;7(2): 315-329.

[10] Lindsay, G. W. Convolutional neural networks as a model of the visual system: Past, present, and future. Journal of cognitive neuroscience, 2021; 33(10): 2017-2031.

[11] Ma, L., & Sun, B. Machine learning and AI in marketing–Connecting computing power to human insights. International Journal of Research in Marketing, 2020; 37(3): 481-504.

[12] Lu, H., Zhang, M., Xu, X., Li, Y., & Shen, H. T. Deep fuzzy hashing network for efficient image retrieval. IEEE transactions on fuzzy systems, 2020; 29(1): 166-176.

[13] Jarada, T. N., Rokne, J. G., & Alhajj, R. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. Journal of cheminformatics, 2020; 12(1): 1-23.

[14] Jiang, J., Chen, M., & Fan, J. A. Deep neural networks for the evaluation and design of photonic devices. Nature Reviews Materials, 2021; 6(8): 679-700.

[15] Gai, K., Guo, J., Zhu, L., & Yu, S. Blockchain meets cloud computing: A survey. IEEE Communications Surveys & Tutorials, 2020; 22(3): 2009-2030.

[16] Chen, H., Jiang, B., Ding, S. X., & Huang, B. Data-driven fault diagnosis for traction systems in high-speed trains: A survey, challenges, and perspectives. IEEE Transactions on Intelligent Transportation Systems, 2020; 23(3): 1700-1716.

[17] Xin, W. A. N. G., Zi-Yi, W. A. N. G., Zheng, J. H., & Shao, L. I. TCM network pharmacology: a new trend towards combining computational, experimental and clinical approaches. Chinese Journal of Natural Medicines, 2021; 19(1): 1-11.

[18] Ren, P., Xiao, Y., Chang, X., Huang, P. Y., Li, Z., Chen, X., & Wang, X. A comprehensive survey of neural architecture search: Challenges and solutions. ACM Computing Surveys (CSUR), 2021; 54(4): 1-34.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Nov 3, 2023
Accepted: Nov 22, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1940–1949, DOI 10.12694/scpe.v25i3.2795

RESEARCH ON DYNAMIC OPTIMIZATION ALGORITHM OF WAREHOUSING LOCATION LAYOUT BASED ON NONLINEAR OPTIMIZATION

GUANG CHEN[*], ZHIWEI TU[†], SHENG ZHANG[‡], JING FANG[§], AND FAN SHE[¶]

Abstract. The paper aims to improve the turnover rate and operation efficiency of goods that are shipped out and replenished in the warehouses of electric power enterprises through big data analysis and optimization algorithms.The data is distributed in diverse locations and data nonlinear optimization algorithms certainly helps to understand the patterns for effective management of warehouses.This article focuses on reducing the delay in the operational processes.A multi-objective optimization (MOO) has been proposed which is aiming at improving the efficiency of transition process of commodities, storage, and overall warehouse operations.The study helps in the optimization of the allocation of cargo spaces with the aid of big data analysis optimization technology which collects and manages data in a distributed environment. A multi-objective cargo space optimization algorithm is proposed along with consideration of dynamic constraints.The algorithm is based on the coefficient of variation adaptive differential evolution algorithm.Individual decoding is performed according to the real-time cargo space availability.The simulation results show that the convergence speed of the algorithm is greatly improved.Meanwhile, the efficiency of warehouse transition process, shelf stability and the classification of commodities are remarkably improved.In nutshell, the multi-objective decision-making with the integration of big data analysis optimization technology assists in the effective organization of warehouse allocation system by considering multiple factors and constraints.

Key words: big data analysis; cargo location optimization; dynamic constraints; multi-objective; non-linear optimization

1. Introduction. With the increase in the scale and quantity of power projects, the warehousing management of power materials has become more and more complicated, which is easy to cause problems such as untimely supply of materials and inability to deliver, which affects the smooth development of power projects and the operation of power grids security and stability. In view of the large quantity and variety of power supplies, reasonable storage space allocation can provide higher picking efficiency for the warehouse in operation, reduce the loss of goods in the process of loading, unloading, handling, storage and picking, and effectively reduce the storage in the warehouse. Operating cost.

The optimization of the cargo location layout refers to the process of dynamic adjustment and reconfiguration of the company's inventory settings and the placement of goods according to the characteristics of materials, demand response and changing factors. The optimization of the cargo space requires the cooperation between different equipment, tools and labor. According to the shelf type, the characteristics and classification of the goods, the planning of the cargo space, the artificial factors, etc., the optimal cargo space allocation is jointly realized. Warehouse location optimization can provide higher picking efficiency for operating warehouses, reduce the loss of goods in the process of loading, unloading, handling and storage picking, and effectively reduce operating costs in warehousing. Therefore, there is huge room for improvement in the storage space. The requirements of modern warehouse management systems are complex, and the optimization problem of cargo location decision considering dynamic resource constraints and various objectives has more practical application value.

In this paper, combined with the actual needs of a power company, an improved nonlinear multi-objective adaptive differential evolution algorithm (ADEA) is proposed. By adding secondary optimization links, the

[*] State Grid Corporation of China, Beijing, China

[†] State Grid Electric Power Research Institute, Nanjing, China; Wuhan Nari Limited Liability Company of State Grid Electric Power Research Institute, Wuhan, China

[‡] State Grid Jiangsu Electric Power Co.,Ltd., Nanjing, China

[§] State Grid Electric Power Research Institute, Nanjing, China; Wuhan Nari Limited Liability Company of State Grid Electric Power Research Institute, Wuhan, China (Corresponding author, fjkite@163.com)

[¶] State Grid Fujian Electric Power Co.,Ltd., Fuzhou, China

convergence speed of the algorithm is improved. To meet the three requirements of stability and cargo placement correlation, a comprehensive and feasible solution for cargo location optimization was obtained by building a multi-objective optimization model, improving algorithms, and simulation verification, thereby realizing online dynamic cargo location allocation of random inventory.

2. Related Work. The main objectives of the optimization of the cargo location layout are the frequency of storage and exit, shelf stability, commodity relevance, and space utilization. Yang et al. [1] optimize the picking location of the roadway stacker when exiting the warehouse, a mathematical function model was established by analyzing the picking strategy of the warehouse delivery operation, which was established with the shelf stability and access efficiency as the goal according to the principle of storage space allocation. Wang et al. [2] designed an optimized target model based on the vertical stability of the shelf, which meets the maximum load capacity and maximum limit of the shelf, and minimizes the center of gravity of the goods. And designed a hierarchical genetic algorithm, the calculation result reduces the center of gravity of the shelf. Zhang et al. [3] introduced the concept of the demand correlation pattern to describe the correlation among items, based on which a new model is constructed to address the SLAP. The model is subsequently reduced using the S-shape routing strategy, and a method for determining DCPs from historical data is proposed. Zhou et al. [4] established an optimization model by establishing the relationship between storage products, combining the current distribution strategy, and simulating through software. Quintaniua et al. [5] studied the establishment of an optimization model with the goal of maximizing storage space utilization. Solving this method not only greatly improves the warehouse utilization rate, but also shortens the picking time.

When calculating optimization models with diverse objectives and complex constraints, a reasonable calculation method will make the calculation process faster and the calculation results more accurate. Therefore, how to choose an efficient solution algorithm is also the focus of research in the cargo location optimization problem. Lin et al. [6] established a multi-objective optimization model that considers the determination of retrieval time and retrieval frequency based on genetic algorithm, which effectively improves the search ability under the constraints of frequent entry and exit of different goods. Seval et al. [7] improved the performance of the genetic algorithm, which solved the problem of location allocation based on clustering storage strategy and minimizing picking costs as the goal, and effectively optimized the warehouse layout in the automotive industry. Muppani et al. [8] proposed a linear optimization model of cargo location allocation based on simulated annealing algorithm to improve control space utilization and reduce picking costs. This model is better than traditional dynamic programming algorithms in accuracy. For the storage location allocation problem with grouping constraints, Xie et al. [9] established a two-layer grouping optimization model, and solved it by a multi-stage random search method and a tabu search algorithm. Aiming at the problem of product demand fluctuations over time, Patrick et al. [10] proposed an iterative heuristic method that solves the problems jointly and that takes account of future dynamics in customer demand and their influence on the three planning problems.

However, the above research only solves the optimization of cargo location decision under single-batch operation, and does not further discuss the variation law of the optimization target value of multi-batch operations and whether the continuous optimization capability of cargo location is stable. When constructing constraints, Augustyn et al. [11] considered the influence of environmental factors such as warehouse size on the optimization of storage location decisions, but did not consider the dynamics of factors such as inventory and allocable storage locations. In terms of model algorithm implementation, in recent years, research on cargo space optimization using genetic algorithm [12] and machine learning [13] has been emerging, and differential evolution algorithm (DE) has attracted much attention due to its excellent optimization performance [14] , and also has certain applications in warehouse optimization [15]. There are relatively few researches on the optimization of cargo location decision based on differential evolution and Pareto optimal [16].

The above-mentioned various studies mainly limited the consideration of a certain part in the allocation of goods, leading to some deficiencies in the model, such as insufficient model factors and slow algorithm convergence. This paper uses different principles of cargo location allocation to establish a cargo location allocation optimization model, uses the weight coefficient method to convert it into a single objective function, and solves the model through a nonlinear multi-objective adaptive differential evolution algorithm. Considering constraint conditions such as dynamic inventory and allocable cargo locations, and based on adaptive differential evolution of mutation parameters, a multi-objective cargo location optimization algorithm that responds to dynamic constraints is proposed. The Pareto solution set is further evaluated based on the analytic hierarchy process, and the influence of multi-objective weights on the continuous optimization of multi-batch operations is studied. The optimization algorithm in this paper can jump out of the local minimum extremal region, so as to find the global optimal solution faster, thereby ensuring the convergence of the algorithm. In the establishment of the objective function, this paper takes into account the nearest storage of the goods, the lowest center of gravity, and the correlation criteria of the goods, which is more comprehensive than the previous research.

3. Model. Improper distribution of goods is the primary problem that restricts the efficiency of warehouse inbound and outbound. At present, most power material warehouses use random storage strategies, which not only greatly increase the time for goods in and out of the warehouse, but also increase the difficulty of picking goods accordingly. In addition, the quality of different types of goods and the frequency of their storage will also affect the shelf life and the efficiency of goods storage. Therefore, it is necessary to set an appropriate storage strategy for warehouse space optimization and re-plan it using the distribution principle of the storage space; then use different principles to establish multiple models for optimization, and the final distribution of the storage space can reach the ideal state. The traditional optimization method of cargo location is only optimized according to the frequency of its storage and exit, but this optimization method considers few aspects and is not comprehensive enough.

General scheduling optimization uses Petri nets, expert systems, temporal logic, simulated annealing, neural networks, genetic algorithms, etc. [12][13][14]. The genetic algorithm is the most efficient way to achieve global search, but its coding method will become difficult as its model becomes more complicated. In this paper, the nonlinear algorithm based on multi-objective adaptive differential evolution is used to solve the problem, so that the evolutionary algorithm jumps out of the local extreme value region, finds the global optimal solution, and ensures the convergence of the algorithm.

3.1. Problem assumption. Different power material warehouses have different characteristics. Combining the characteristics of multiple power material warehouses, some special circumstances are not considered for the time being. The following assumptions and explanations are made for the warehouse warehousing operation process[17]: (1) The volume of each cargo space in the warehouse is equal (2) The goods are placed side by side in a single layer in the cargo space; (3) The warehouse adopts a storage strategy of random storage; (4) The warehouse operations are completed by manual handling equipment, and the handling equipment is in the shelf area The walking route is arbitrary; (5) During the warehousing process, forklifts are used as handling equipment, and each forklift has the same load capacity and can carry different types of goods at the same time; (6) The warehouse has enough space to meet all waiting Demand for incoming goods.

3.2. Model establishment. The symbols used in the model are explained as follows: There are a row of shelves in the warehouse, and each row of shelves has column b and layer C. The coordinates (0, 0, 0) indicate the location of the inbound and outbound platform, and the position coordinates of a certain cargo k are (x_k,y_k,z_k) (The value range of x_k is 1 to a, the value range of y_k is 1 to b, and the value range of z_k is 1 to c); v_x represents the transmission speed of the stacker in the x-axis direction; v_y is the transfer speed of the stacker in the y-axis direction; v_z is the transfer speed of the stacker in the z-axis direction; L is the length of the shelf cell; L_0 is the distance between the shelves; $r_k$is the turnover of the k-th product Rate (frequency of in and out of storage); w_k is the unit mass of the k-th product.

3.3. Analysis of optimization goals. The principle of cargo space allocation mainly includes the principle of nearby storage and exit, the principle of the lowest center of gravity, the principle of relevance of goods, and so on. A good distribution principle can not only reduce the distance between goods in and out of the warehouse and shorten the time required for operations, but also make full use of its storage space while meeting the requirements of shelf stability and reduce storage costs. The optimization objectives of the three types of allocation principles are as follows:

(1) The principle of nearby storage. In order to improve the efficiency of warehousing and warehousing, materials with high turnover rate should be closer to the entrance and exit so that the overall warehousing time is shorter. Assuming that the stacker pick-up time is negligible, for a cargo located on the z-layer of the x row, y column, and z layer, its in-out time can be simplified as the total operation time of the stacker.

The running time of the stacker in the x direction is $\sum_{x=1}^{a}\frac{x_k\times(L+L_0)}{v_x}$, and the running time in the y direction is $\sum_{y=1}^{b}\frac{y_k\times L}{v_y}$, the running time in the x direction is $\sum_{z=1}^{c}\frac{z_k\times(L+L_0)}{v_z}$ respectively. The shortest distance that the goods move in and out of the warehouse in a single time is: $2\sqrt{[x_k\times(L+L_0)]^2+(y_k\times L)^2+[z_k\times(L+L_0)]^2}$, so the optimization objective function of the nearby storage principle is as follows:

$$\begin{aligned}\min f_1(x,y,z)=&\sum_{x=1}^{a}\sum_{y=1}^{b}\sum_{z=1}^{c}\left(\frac{x_k\times(L+L_0)}{v_x}+\frac{y_k\times L}{v_y}+\frac{z_k\times(L+L_0)}{v_z}\right)\\&\times 2\sqrt{[x_k\times(L+L_0)]^2+(y_k\times L)^2+[z_k\times(L+L_0)]^2}\times r_k\end{aligned}\tag{3.1}$$

The goal of this optimization function is the sum of the moving distances of goods k.

(2) The principle of the lowest center of gravity. In order to maintain the stability of the warehouse shelves, according to the principle of light up and down and lower center of gravity, the total center of gravity of the warehouse or warehouse should be as low as possible. Its optimization function can be expressed as:

$$\min f_2(x,y,z)=\frac{\sum_{x=1}^{a}\sum_{y=1}^{b}\sum_{z=1}^{c}w_k n_{xyzk}z_k\left(L+L_0\right)}{\sum_{x=1}^{a}\sum_{y=1}^{b}\sum_{z=1}^{c}w_k n_{xyzk}}\tag{3.2}$$

where n_{xyzk} represents the number of k-th goods on the shelf (x, y, z). The objective function represents the position of the center of gravity of the k-th cargo. Therefore, Goal 2 represents the principle of lightness on the shelf and heavy weight. The function value f_2 of Goal 2 reflects the degree of center, so the optimization goal is to make it as smallest as possible.

(3) Cargo-related principles. If there is a certain correlation between the goods, put the goods that need to be out of the warehouse at the same time and put them in the close or adjacent cargo space. Considering the nature of the cargo itself, the cargo location should be carefully arranged. For example, special types of cargo should be placed in a special location and placed together as much as possible. Its optimization function can be expressed as:

$$\min f_3(x,y,z)=\sum_{x=1}^{a}\sum_{y=1}^{b}\sum_{z=1}^{c}\sqrt{(x_k-\bar{x}_k)^2+(y_k-\bar{y}_k)^2+(z_k-\bar{z}_k)^2}\tag{3.3}$$

where $(\bar{x}_k,\bar{y}_k,\bar{z}_k)$ the center coordinate of the k-th category of goods, and its value is calculated based on the weighted average of all such goods. It can be seen that the target 3 represents the distance between the center position coordinates of a certain type of goods and the position coordinates, so the distance between adjacent goods in the entire warehouse is obtained.

According to the analysis of different principles, three objective functions are established. Combining these three objective functions into a whole can form the mathematical model of the main research problem. This is a multi-objective optimization problem. The mathematical model can be expressed as follows:

$$\begin{aligned}&\left\{\begin{array}{l}\min f_1(x,y,z)\\ \min f_2(x,y,z)\\ \min f_3(x,y,z)\end{array}\right.\\ &\text{s.t}\left\{\begin{array}{l}1\le x\le a\\ 1\le y\le b\\ 1\le z\le c\end{array}\right.\end{aligned}\tag{3.4}$$

For multi-objective optimization problems, in order to comprehensively consider the factors of multiple objectives, this paper studies how to set weights for different objectives to obtain weighted single-objective problems. Starting from the practical application of warehousing, a reasonable single solution needs to be selected from the Pareto solution set. Therefore, the weight is calculated by the commonly used AHP method, and the weight is used to solve the multi-objective weighting, and the individual with the smallest comprehensive objective function value is obtained, and the relationship between the multi-objective weight and the continuous optimization ability of the cargo space is observed. Construct the judgment matrix $B = (b_{ij})_{n\times n}$ according to the importance scale, normalize it by column to get the matrix $\mathrm{C} = (\mathrm{c_{ij}})_{\mathrm{n\times n}}$, divide the sum of the row elements of the C matrix by the sum of the elements The weight of each target can be obtained.

$$\mathrm{c_{ij}} = \frac{b_{ij}}{\sum_{k=1}^{n} b_{kj}} \tag{3.5}$$

$$\varepsilon_{\mathrm{i}} = \frac{\sum_{k=1}^{n} c_{ik}}{\sum_{j=1}^{n}\sum_{i=1}^{n} c_{ij}} \tag{3.6}$$

Determine the weight of the principle of the nearest in and out warehouse, the principle of the lowest center of gravity, and the principle of the correlation of goods through the AHP. After solving ADEA to obtain the Pareto solution set, calculate the comprehensive objective function value F of all individuals in the Pareto solution set according to formula (3.7), and select the individual with the smallest F as the optimal individual. The multi-objective optimization objective translates into the following formula:

$$\begin{gathered} \min f(x,y,z) = \varepsilon_1 \min f_1(x,y,z) + \varepsilon_2 \min f_2(x,y,z) + \varepsilon_3 \min f_3(x,y,z) \\ \mathrm{s.\,t}\begin{cases} \varepsilon_1+\varepsilon_2+\varepsilon_3 = 1 \\ 1 \le x \le a \\ 1 \le y \le b \\ 1 \le z \le c \end{cases} \end{gathered} \tag{3.7}$$

4. Optimization algorithm.

4.1. The cargo location optimization algorithm based on adaptive differential evolution. According to the analysis of the problem, solving the optimal cargo location for cargo distribution is essentially a multi-objective optimization process. The optimization function is shown in the above formula. In this paper, a multi-objective adaptive differential evolution algorithm (ADEA) is used to calculate the optimization process, and the evolutionary parameters and evolutionary operators are adjusted through the adaptive process, which can effectively improve the convergence performance of the algorithm. Through adaptive adjustment of mutation parameters, the individual decodes according to the real-time feasible domain response dynamic constraint conditions of the cargo location, and the comprehensive evaluation of the multi-objective Pareto solution set to obtain the optimal operating cargo location. The algorithm flow is as follows:

(1) Initialization and coding. The individual uses the floating point code in the range of [0,1], the length is equal to the job number D, and the individual gene index number is the job number. Initialize the Pareto solution set R_p as an empty set, and randomly generate NP initial individuals with dimension D. (x_k, y_k, z_k) corresponding to each gene is the rank, column, and layer of the goods.

(2) Individual decoding in response to dynamic constraints. Through the gene value xik corresponding to job k and the real-time feasible domain $D_k(x_1,y_1,z_1),(x_2,y_2,z_2),...(x_n,y_n,z_n)$, the corresponding target location (x_k,y_k,z_k). Calculate the objective function value through the set of cargo locations.

(3) Mutation operation. Assuming that the size of the population is NP and the dimension of the solution is N, then the population X ⊂ R^N, the G-th generation individual i can represent the vector $P_i^G = \left(P_{i1}^G, P_{i2}^G, \ldots, P_{iN}^G\right)$, randomly select 3 individuals from the contemporary population as parent individuals for compilation, and generate mutant individuals u_i^G:

$$u_{ij}^G = P_{r1j}^G + F \times \left(P_{r2j}^G - P_{r3j}^G\right) \tag{4.1}$$

where formula (4.1) is the individual variation formula, and u_{ij}^G is the j-th dimension element of the variation individual u_i^G.

The variation scaling factor F is adaptively generated according to formulas (4.2)-(4.4):

$$F_{i1} = F_u e^{\frac{In(F_1/F_u)t}{G-1}} \tag{4.2}$$

$$F_{i2} = \frac{f_i - f_{\min}}{f_{\max} - f_{\min}} \tag{4.3}$$

$$\mathrm{F} = \frac{F_{i1} - F_{i2}}{2} \tag{4.4}$$

where F consists of two parts, F_{i1} and F_{i2}. Among them, F_{i1} exhibits nonlinear adaptive decay according to time, which can ensure that the algorithm can efficiently perform global search in the early process of evolution, and can obtain a stronger ability to locally seek the optimal solution in the later process of evolution; F_{i2} The optimization parameters are dynamically and adaptively adjusted according to the difference between the objective function value of each individual and the optimal individual in the population; r_1, r_2, and r_3 are 3 integers that are completely different from each other, and are not equal to i. The value is a random selection from the set of 1,2,...,NP; F_l is the lower limit of F_i; F_u is the upper limit of F_i; G is the maximum upper limit of the iteration; $g \in [0, G-1]$ is The current number of iterations; f_i is the real-time target value calculated by a single individual; f_{min} is the smallest target value among all individuals in the current population; f_{max} is the largest target value among all individuals in the current population.

(4) Cross operation. The diversity of the population is the best way to improve the efficiency of finding the optimal solution, the experiment vector $v_i^g = \left[v_{1,i}^g, v_{2,i}^g, \ldots, v_{D,i}^g\right]$ is obtained by using the binomial crossover operation , As shown in formula (4.5), $\mathrm{R}_{\text{i, rand}} \in (1, 2, \ldots, D)$, which can ensure that at least one parameter of $v_{j,i}^g$ comes from U; the crossover operator CR $\in [0, 1]$.

$$v_{ij}^g = \begin{cases} u_{ij}^g, & \text{if } \operatorname{rand}(0,1) \leq CR \text{ and } j = \operatorname{rnbr}(i) \\ P_{ij}^G, & \text{otherwise} \end{cases} \tag{4.5}$$

where rand(i,j) represents the estimated value of a random number between i and j, and rnbr() represents a randomly selected sequence. CR is the crossover operator.

(5) Select operation. Based on the dominant relationship in the multi-objective algorithm, the selection operator is shown in formula (4.6), LC $\left(P_i^{g+1}, v_i^g\right)$ represents the congestion entropy in P_i^{g+1} and v_i^g Smaller individuals.

$$\mathrm{LC}\left(U_i^{g+1}, X_i^g\right) = \begin{cases} P_i^{g+1}, & \text{if } f\left(P_i^{g+1}\right) \gtrsim f\left(v_i^g\right) \\ v_i^g, & \text{if } f\left(P_i^{g+1}\right) \preccurlyeq f\left(v_i^g\right) \\ \mathrm{LC}\left(v_i^{g+1}, P_i^g\right), & \text{otherwise} \end{cases} \tag{4.6}$$

(6) Obtain the Pareto solution set. $x^* \in \Omega$ is the Pareto optimal solution, which means $\nexists x \in \Omega$ such that $f(x) < f(x^*)$, Ω is the feasible domain of the variable x, and the Pareto optimal solution set X_k refers to all P-optimal solutions collection. The P-optimal solution set of the t generation and the population of the t+1 generation are merged to solve the Pareto solution set of the t+1 generation to obtain an optimal solution set that is sufficiently diverse and distributed throughout the Pareto front.

4.2. Algorithm flow. The basic flow of the algorithm is shown in Fig. 4.1.

The specific steps are described as follows:

1. Set the population size NP, the maximum number of iterations G, the termination condition, and the crossover operator CR; initialize the population, the individual vector dimension D is equal to the number of jobs n, encode the individuals with random numbers, and initialize the Pareto solution set as an empty set.
2. Determine whether the maximum number of iterations G is reached, if so, select the individual with the smallest comprehensive objective function value from the Pareto solution; if not, proceed to the next step.
3. Perform mutation crossover operation and calculate the objective function value.
4. Determine the dominant relationship between P_i^{g+1} and v_i^g, and select the individual v_{i+1}^g.

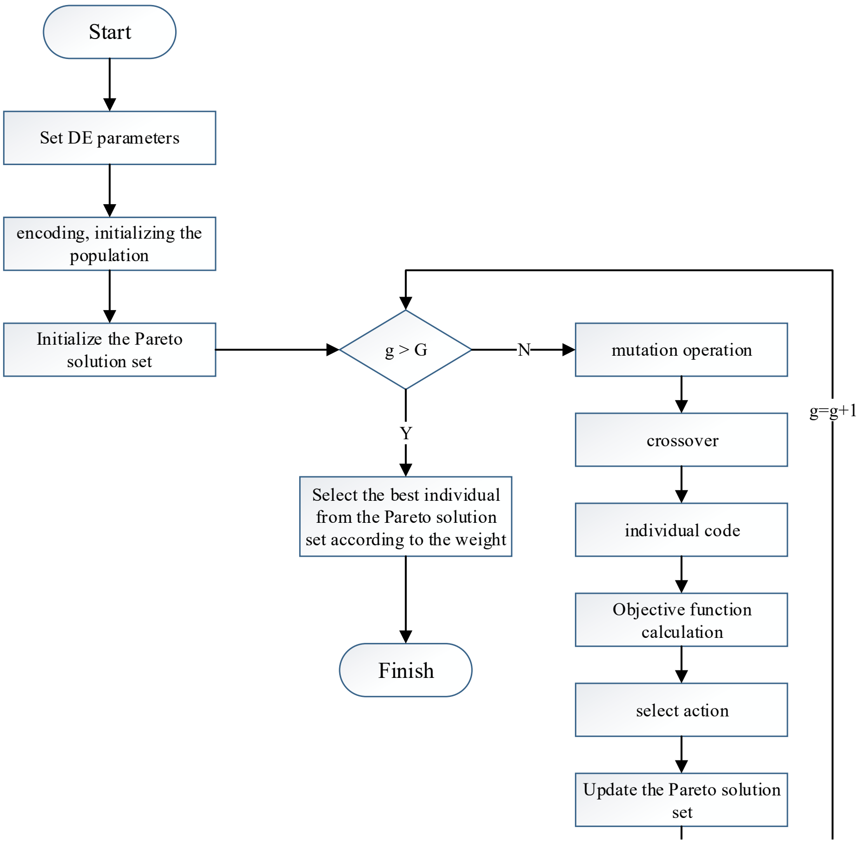

Start
Set DE parameters
encoding, initializing the population
Initialize the Pareto solution set
g > G
N
Y
mutation operation
crossover
individual code
Objective function calculation
select action
Update the Pareto solution set
g=g+1
Select the best individual from the Pareto solution set according to the weight
Finish

Fig. 4.1: Flowchart of ADEA

5. Compare the target vector v_i^g with all vectors in the Pareto solution set v_P, and update the Pareto solution set.
6. The number of iterations g=g+1, merge the Pareto solution obtained in the g-th generation in the population, and go to step 2.

5. Experiment.

5.1. Setting. A simulation experiment has been developed to verify the effectiveness of this algorithm, the experiment taked a warehouse of a power company as a prototype,builded a simulated automated three-dimensional warehouse.there are 100 random storage positions, the warehouse has 10 rows, 10 rows and 4 floors, and the center position coordinates of various types of goods. Set as (2,2,2), (3,2,2), (2,3,2) respectively. The experimental data set in this paper comes from randomly generated simulation data. The warehouse uses conveyor belts in the horizontal direction and tunnel stackers in the vertical direction. The relevant parameters of the warehouse and cargo space are shown in Table 5.1. The simulation experiment platform was windows 10 system, the computer configuration was Intel CPU 2.90 GHz, memory 16G, GPU was GeForce GTX 1650, programming language was Python 3.1, and the integrated open environment was Anaconda 3. The neural network is implemented using the Pytorch open source framework.

5.2. Iterative performance analysis. Fig. 5.1 shows the change curve obtained from the experiment, which represents the optimal fitness function value in the solution process of the multi-objective adaptive differential evolution algorithm and the traditional genetic algorithm. The process of iterating with the population. According to the curve in Fig. 5.1 , we find that the convergence speed of the multi-objective adaptive dif-

Table 5.1: Optimize simulation parameters

Movement speed in X and Y directions	1 m/s	Algorithm iteration times	400
Movement speed in Z direction	0.5 m/s	initial population	100
warehouse slot length	1 m	Number of shelves	10
Number of shelves	10	Shelf layers	4

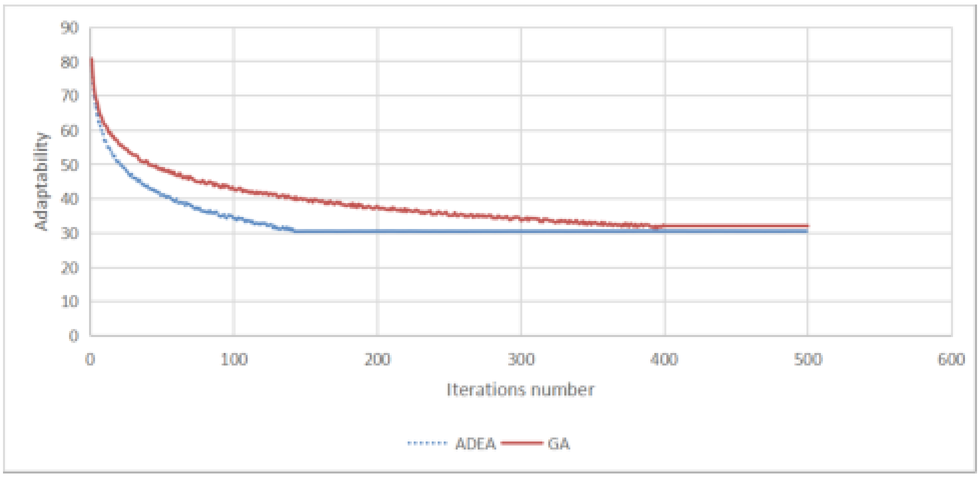


Fig. 5.1: Adaptability curve

ferential evolution algorithm is very fast in the early stage of the iteration; when the population is iterated to 150 times, the result tends to be stable, which reflects that the calculation time and error are better; and Correspondingly, the traditional genetic algorithm needs about 400 iterations to obtain the same result; from this, it can be found that the convergence speed of the multi-objective adaptive differential evolution algorithm is about 2.7 times faster than that of the traditional algorithm.

5.3. Optimization effect analysis. In order to analyze the function of the optimization algorithm in this paper more intuitively, the function values of the three objective functions before and after optimization are solved separately. Table 5.2 shows the comparison of the three objective functions before and after optimization. The three objective functions respectively represent the optimized target value of the nearest storage principle, the optimized target value of the lowest center of gravity principle, and the optimized target value of the cargo-related principle. The smaller the value, the better the effect. The results show that the values of the three objective functions are reduced by 47%, 55%, and 72% respectively compared with before optimization, and the optimization effects are different. Among them, the optimization effect of goods-related storage is the best.

It can be seen from Table 5.2: According to the comprehensive analysis of multi-objective decision-making, the average objective function has dropped by 58%, and the optimization effect is significant. Comparing the distribution status of the cargo space before and after optimization, we can find that the cargo space distribution before optimization is chaotic and the layout is chaotic; the optimized cargo space is mostly concentrated near the exit position, and most of the goods are located on the bottom floor, and the overall layout is reasonable and orderly. Comprehensive analysis of warehouse entry and exit efficiency, rationality of goods classification and storage, and overall stability of shelves have been significantly improved compared with the optimization before.

The experimental results show that the method proposed in this paper can meet the needs of multi-project and multi-material, combined with the cost, time and efficiency of different project locations, as well as the balance problem of multiple types of materials in a single project in different storage locations, and solve the ambiguity of multiple optimization objectives. The matching problem can intelligently determine the optimal

Table 5.2: Object values compare

Objective function	Before optimization	After optimization	Decrease rate
f_1	168	89	47%
f_2	77	34	55%
f_3	298	81	72%

allocation plan based on the global perspective, which can effectively improve the utilization efficiency of materials in the warehouse and the utilization efficiency of materials.

6. Conclusion. Warehouse management includes various processes that make the in-out process of goods convenient and easy by enhancing the efficiency of operations. This paper focuses on the optimization of three factors namely- nearby storage, the CoG, and correlation among the goods for optimization enabled warehousing locations. To optimize these processes and to ascertain the values of optimization functions, a multi-objective ADEA method is proposed. The evolutionary parameters and evolutionary operators in the algorithm are adjusted through the adaptive process. This effectively improves the convergence performance of the algorithm. The results show that the values of the three FFs are reduced by 47%, 55%, and 72% respectively as compared to the values before optimization. It can be concluded that the proposed optimization algorithm can respond to dynamic constraints such as allocating the feasible region of the cargo space optimally. When compared with the simple weighted differential evolution algorithm, the proposed algorithm shows better performance. The proposed methods are suggested to minimize the time of storage operations and to minimize the time for transition of goods. The weight of the optimization target affects the continuous optimization ability of the cargo space. To sum up, the proposed algorithm can effectively solve the optimization problem of dynamic cargo location allocation for power generation inventory. In future, we will enhance the algorithm to consider complex situations such as inbound and outbound mixed batches and consider a case of diverse commodities.

Acknowledgements. This paper is supported by State Grid Corporation Headquarters Technology Project (Research on Key Technologies of Modern Intelligent Warehousing System and Logistics Equipment, No:1400-202118268A-0-0-00).

REFERENCES

[1] Yang, D., Wu, Y., & Ma, W. (2021). Optimization of storage location assignment in automated warehouse. Microprocessors and Microsystems, 80, 103356.

[2] Wang Xuelian, Xu Man, Xiao Jing and Guo Ran. (2014). Optimization of goods locations assignment of automated warehouse on hierarchic genetic algorithm[J]. Applied Mechanics and Materials. 510,265-270

[3] Zhang, R. Q., Wang, M., & Pan, X. (2019). New model of the storage location assignment problem considering demand correlation pattern. Computers & Industrial Engineering, 129, 210-219.

[4] Zhou Guiliang and Mao Lina.(2010). Design and simulation of storage location optimization module in AS/RS based on FLEXSIM[J]. International Journal of Intelligent Systems & Applications, 2(2),1-4.

[5] Quintanilla Sacramento, Perez Angeles, Ballestin Francisco and Lino Pilar. (2012). Heuristic algorithms for a storage location assignment problem in a chaotic warehouse[J]. Engineering Optimization, 47(10), 1405-1422.

[6] Rani V.Uma, Chandra J.LinEby and Jayashree D. (2016). Efficient storage location assignment using genetic algorithm in warehouse management system[J]. Advanced Research, (1),18-24.

[7] Ene Seval and Ozturk Nursel.(2012). Storage location assignment and order picking optimization in the automotive industry[J]. International Journal of Advanced Manufacturing Technology, 60 (5-8), 787-797.

[8] Muppant Reddy Muppani Venkata and Adil Kumar Gajendra. (2008). Efficient formation of storage classes for warehouse storage location assignment: A simulated annealing approach[J]. Omega, 36 (4), 609-618.

[9] Xie Jing, Mei Yi, Ernst T. Andreas, Li Xiaodong and Song Andy. (2016). bi-level optimization model for grouping constrained storage location assignment problems[J]. IEEE Transactions on Cybernetics, 48(1), 385-398.

[10] Kübler, P., Glock, C. H., & Bauernhansl, T. (2020). A new iterative method for solving the joint dynamic storage location assignment, order batching and picker routing problem in manual picker-to-parts warehouses. Computers & Industrial Engineering, 147, 106645.

[11] Lorenc A, Lerher T. Effectiveness of product storage policy according to classification criteria and warehouse size[J]. FME transactions, 2019, 47(1): 142-150.

[12] Wei Yu, Shijun Li, Xiaoyue Tang, Kai Wang. (2019). An efficient top-k ranking method for service selection based on -ADMOPSO algorithm[J]. Neural Computing and Applications. 31(1), 77-92.

[13] Wei Yu, Shijun Li. (2018). Recommender systems based on multiple social networks correlation[J]. Future Generation Computer Systems. 87, 312-327.

[14] Das S, Mullick S S, Suganthan P N. Recent advances in differential evolution–an updated survey[J]. Swarm and evolutionary computation, 2016, 27: 1-30.

[15] Yu Y , Ma H , Mei Y , et al. Multipopulation Management in Evolutionary Algorithms and Application to Complex Warehouse Scheduling Problems[J]. Complexity, 2018, 2018:1-14.

[16] Wisittipanich W , Hengmeechai P . A Multi-Objective Differential Evolution for Just-In-Time Door Assignment and Truck Scheduling in Multi-door Cross Docking Problems[J]. Industrial Engineering & Management Systems, 2015, 14(3):299-311.

[17] J Saderova, Poplawski L , Jr M B , et al. LAYOUT DESIGN OPTIONS FOR WAREHOUSE MANAGEMENT[J]. Polish Journal of Management Studies, 2020, 22(2):443-455.

Edited by: Zhigao Zheng
Special issue on: Graph Powered Big Aerospace Data Processing
Received: Nov 4, 2023
Accepted: Nov 24, 2023

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1950–1962, DOI 10.12694/scpe.v25i3.2686

BLOCKFOG: A BLOCKCHAIN-BASED FRAMEWORK FOR INTRUSION DEFENSE IN IOT FOG COMPUTING

VG PRASUNA,[*] B. RAVINDRA BABU,[†] AND BHASHA PYDALA[‡]

Abstract. In the rapidly evolving domain of the Internet of Things (IoT) and fog computing, maintaining security, scalability, and efficient operation poses significant challenges. Addressing these issues, this study introduces ”BlockFog,” a novel blockchain-based framework designed to bolster intrusion defense in IoT fog computing environments. The core objective of BlockFog is to counteract the vulnerabilities inherent in decentralized IoT ecosystems by leveraging blockchain technology for enhanced security and transparency. The framework's innovative design integrates crucial components such as Device Onboarding & Identity Management, Data Integrity & Logging, Smart Contract-Driven Intrusion Detection, Automated Blockchain Responses, Secure Peer-to-Peer Communication, and a Lightweight Consensus Mechanism. These elements work collectively to ensure the security and functionality of IoT devices within the fog computing paradigm. BlockFog stands out for its meticulous approach to handling high transaction volumes with off-chain computations and layer-2 solutions, ensuring data integrity and facilitating seamless audit processes. The framework's resilience is further demonstrated through its robust response to evolving cyber threats, incorporating Over-the-Air (OTA) updates and advanced data protection mechanisms like zero-knowledge proofs. A comparative analysis highlights BlockFog's superior performance against existing models. The results reveal BlockFog's lower latency rates in normal, high traffic, and attack scenarios, its higher throughput efficiency, and its more effective resource utilization in terms of CPU, memory, and bandwidth usage. Moreover, BlockFog exhibits an enhanced ability to detect and respond to malicious activities, including DDoS attacks, with significantly higher accuracy than its counterparts. These findings underscore BlockFog's potential in redefining security and operational paradigms in IoT fog computing, making it a robust, agile, and transparent framework suitable for the current digital landscape.

Key words: BlockFog, Internet of Things, blockchain, fog computing, Hybridchain-IDS .

1. Introduction.

In the digital age, transformative technologies are redefining possibilities. The Internet of Things (IoT), which imagines a world where every object, from the commonplace to the crucial, is connected and interactive, is driving this technological renaissance. The vastness of IoT is complemented by fog computing, which decentralizes data processing closer to data sources. Theirs is the basis for smarter homes, cities, and industries [1]. There are problems with efficiency, scalability, and security in this digital paradise. Particularly in decentralized systems, existing IoT frameworks struggle to strike a balance between security, scalability, and efficiency [2]. These systems are susceptible to device manipulation and data breaches. How do we keep the connected future of IoT from turning into a cybersecurity nightmare?

This paper introduces a comprehensive framework that tackles the aforementioned problems and establishes a new benchmark for IoT fog computing solutions in response to this conundrum. We created and developed ”BlockFog,” an intrusion defense framework for Internet of Things fog computing that is based on blockchain. BlockFog integrates blockchain technology, which is renowned for its security and transparency, to provide an IoT fog computing solution that is safe, scalable, and efficient.

Some new features of the BlockFog model are as follows. Every element, from smart contract-driven intrusion detection to cryptographic device onboarding, is thoughtfully designed to address challenges within the IoT ecosystem. BlockFog's importance is justified for two reasons. By utilizing blockchain technology, it introduces an unchangeable, transparent, decentralized ledger system that guarantees data integrity and trust. Second, its scalable architecture effectively manages the enormous volumes of transactions generated by IoT.

[*] Professor, Department of CSE, Satya Institute of Technology and Management, Vizianagaram, Andhra Pradesh (Corresponding author, prasunavg@gmail.com)

[†] Faculty of ICT Department, FDRE TVTI, Addis Ababa, Ethiopia (ravindrababu4u@yahoo.com).

[‡] Assistant Professor, Department of Data Science, Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College), Tirupati-517 102,A.P. India (basha.chanti@gmail.com).

Innovative blockchain applications and cryptographic principles were used in the development of BlockFog. These techniques were selected following a careful examination of blockchain and IoT best practices and trends. The integration with the BlockFog framework demonstrates our commitment to a reliable, long-lasting solution.

Following this introduction, related research covers the review of contemporary literature. Further section 2 go over the architecture and significance of BlockFog. A comprehensive experimental study in section 3 contrasts the performance of BlockFog with contemporary models. Further, section 4 concludes the study's findings.

In today's digital world, device communication and data sharing have been transformed by the Internet of Things (IoT). But in order to make these massive networks resistant to changing cyber threats, it is necessary to address the inherent security vulnerabilities of IoT as it expands. Modern research on blockchain's contribution to the revolution in IoT security is reviewed in this review. The in-depth analysis of 25 foundational works will show how researchers and technologists are reinforcing IoT system defenses with the help of blockchain's decentralized ledger and sophisticated cryptography techniques.

Mathew et al. [3] explored the merger of blockchain and collaborative intrusion detection for secure data transactions in industrial IoT. They pinpointed the challenges of multi-layered networks with varied protocol standards, suggesting the integration of collaborative IDS and blockchain as a promising solution. Babu et al. [4] emphasized on safeguarding urban IoT data against DDoS attacks using a permissioned blockchain, introducing the arbiter PUF model for IoT device security.

Li, Wenjuan, Yu Wang et al. [5] presented a blockchain-based filtration mechanism integrated with a collaborative intrusion detection network for IoT security. Their focus was on curbing attacks like DDoS and ensuring efficient traffic management using blockchain and IPFS. Saravanan, V., M. Madiajagan et al. [6] brought forward a Blockchain-based African Buffalo scheme integrated with a Recurrent Neural Network model to enhance cloud-based intrusion detection.

Abou El Houda et al. [7] introduced FedIoT, combining Explainable AI techniques and blockchain to secure Federated Learning-based Intrusion Detection Systems in IoT. Douiba, Maryam et al. [8] addressed security in the healthcare industry, emphasizing the significance of Internet of Health Things. Their solution was a collaborative fog-based intrusion detection system bolstered by blockchain and machine learning.

Siddamsetti, Swapna et al. [9] highlighted a machine blockchain framework for distributed intrusion detection in IoT networks, emphasizing the importance of smart contracts. Meanwhile, Aburas et al. [10] underscored the need for green IoT, discussing how blockchain can bolster its security. Osama Alkadi et al. [11] focused on shared cloud infrastructure and IoT security, presenting a deep blockchain framework.

Hafsa Benaddi et al. [12] emphasized the need for sophisticated IoT intrusion detection systems that can adapt to new attack types, with blockchain integration seen as a potential solution. M. Praveen Kumar et al. [13] highlighted IoT vulnerabilities, proposing blockchain-based solutions with Zero-Knowledge proof techniques for added data protection. Rajesh Kumar Sharma et al. [14] discussed the dangers of malicious attacks in IoT networks and how blockchain can be leveraged for better protection.

Omkar Shende et al. [15] introduced the Collaborative Ensemble Blockchain Model for effective IoT intrusion traffic analysis, employing an ensemble of machine learning models. Eman Ashraf et al. [16] proposed FIDChain, which combines lightweight artificial neural networks and blockchain for healthcare IoT security.

Randhir Kumar et al. [17] put forth a fog computing solution to detect DDoS attacks in blockchain-enabled IoT networks, achieving efficient detection compared to traditional methods. Mohanad Sarhan et al. [18] proposed a hierarchical blockchain-based federated learning framework for IoT intrusion detection, ensuring enhanced security while maintaining data integrity.

Rezvan MAHMOUDIE et al. [19] highlighted the security challenges of decentralized IoT devices, introducing a private blockchain model for IoT intrusion detection that optimizes scalability and minimizes overhead. Salaheddine Kably et al. [20] highlighted the integration of blockchain technology with intrusion detection systems to protect IoT nodes.

Mamunur et al. [21] explored the concept of Federated Learning for IoT security and how it can be boosted using blockchain technology. Jawad Hassan et al. [22] provided a comprehensive overview of blockchain-based intrusion detection for IoT, emphasizing blockchain's potential to redefine IoT security.

Reda Salama et al. [23] proposed the BXAI-IDCUCS model, integrating blockchain with Explainable

Artificial Intelligence for IoT security. AHMED A. M. SHARADQH et al. [24] emphasized HybridChain-IDS, a model that combines blockchain, trusted execution environments, and machine learning for advanced IoT network security.

Hayam Alamro et al. [25] put forward FIDANN, a Federated Artificial Intelligence System of Intrusion Detection for IoT Healthcare System using blockchain. The model leverages machine learning, blockchain technology, and edge computing for efficient intrusion detection. Finally, Priyanka Tyagi et al. [26] addressed the security concerns of IoT-based healthcare systems, proposing FIDANN to enhance intrusion detection mechanisms.

The Internet of Things (IoT) has quickly become known as a challenging field with enormous potential. But there are a lot of security problems as a result of this quick growth. These vulnerabilities are brought to light and the necessity for strong, decentralized, and scalable security solutions is emphasized by recent studies, such as Mathew et al. [3] and Salaheddine Kably et al. [20]. The scholarly contributions underscore the potential of blockchain technology to fortify Internet of Things networks. The decentralized ledger of blockchain is well-known. Every device touchpoint has transparent logging and authentication. Research suggests integrating blockchain to safeguard IoT networks from various cyber threats in light of these advantages.

Industrial IoT was thoroughly examined by Mathew et al. [3]. The need for improved security in intricate, multi-layered networks is highlighted by their research. Network defenses could be completely transformed by fusing blockchain technology with Collaborative Intrusion Detection Systems (CIDS). The threat posed by DDoS attacks on IoT networks is highlighted by parallel research conducted by Babu et al. [4] and Li et al. [5]. Their research shows how successful collaborative detection systems are at stopping these kinds of crimes, particularly when paired with the robust security of blockchain technology. The advantages of blockchain and machine learning algorithms for security are emphasized by Saravanan et al. [6] and Siddamsetti et al. [9]. They claim that harmonizing these can increase the effectiveness and precision of IoT intrusion detection. Aburas et al. [10] and Douiba et al. [8] have pushed blockchain-based solutions for green IoT and healthcare. Their arguments focus on these areas' particular vulnerabilities and the requirement for tailored, efficient security solutions.

The need for BlockFog is highlighted by this research. BlockFog gives hope as blockchain-integrated Internet of Things solutions are discussed. It stands out for its distinct take on blockchain in IoT fog computing and Smart Contract-Driven Intrusion Detection. Its Device Onboarding & Identity Management feature further strengthens its position as the industry leader in IoT security solutions. BlockFog is contrasted with other models, specifically AHMED A. M. SHARADQH [24] and Salaheddine Kably et al. [20]. The performance and potential of BlockFog in IoT security have been evaluated with the aid of these and other references. BlockFog's superiority and innovation in IoT security are highlighted by the contrast between its architecture and empirical performance when compared to these models.

2. Methods and Materials. Navigating the intricate crossroads of the burgeoning realms of the Internet of Things (IoT) and fog computing, it becomes clear that a delicate balance between security, scalability, and streamlined operation is paramount. Enter "BlockFog," a framework birthed precisely at this juncture. It melds blockchain technology seamlessly into the heart of IoT fog computing, marking a significant stride in countering both existing and anticipated challenges of decentralized IoT ecosystems. By strengthening the very peripherals of our digital infrastructures, BlockFog heralds an era of robust, transparent, and agile frameworks for a new wave of interconnected devices. In response to the vastness of the IoT universe, BlockFog is meticulously designed for scale and agility. By harnessing off-chain computations and leveraging layer-2 solutions, it manages to process high transaction volumes with finesse. True to its commitment to transparency, every step—from a device's initial registration to its critical alerts—is intricately logged onto the blockchain. This approach not only fosters trust but also streamlines audit processes. Moreover, as the cyber threat landscape constantly evolves, BlockFog remains ahead, offering timely security enhancements via Over-the-Air (OTA) updates [27]. While the framework's transparency is evident, it is equally dedicated to safeguarding sensitive data, implementing sophisticated mechanisms like zero-knowledge proofs to ensure certain transaction specifics remain confidential. The "BlockFog Framework," its components, and how they relate to IoT devices and the blockchain are depicted in the figure 2.1. Security and functionality are guaranteed by the "Device Onboarding & Identity Management,"[28], "Data Integrity & Logging," [29] and "Smart Contract-Driven Intrusion Detec-

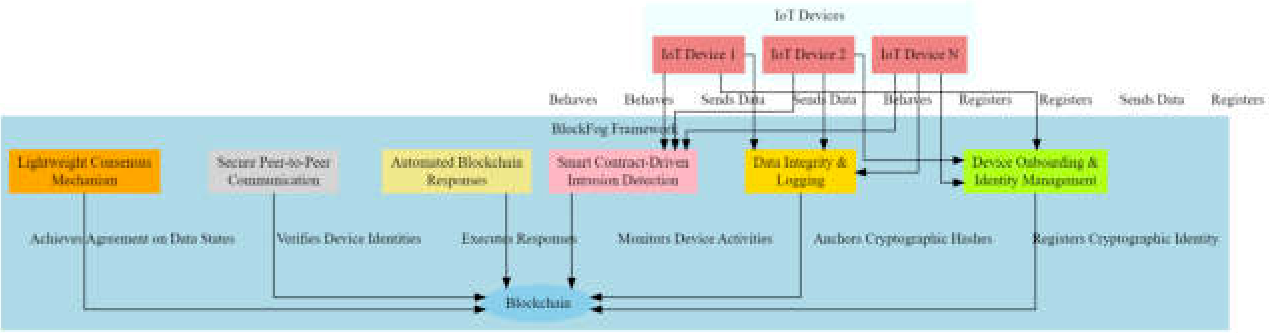

IoT Devices
IoT Device 1
IoT Device 2
IoT Device N
Behaves
Behaves
Sends Data
Sends Data
Behaves
Registers
Registers
Sends Data
Registers
BlockFog Framework
Lightweight Consensus Mechanism
Secure Peer-to-Peer Communication
Automated Blockchain Responses
Smart Contract-Driven Intrusion Detection
Data Integrity & Logging
Device Onboarding & Identity Management
Achieves Agreement on Data Status
Verifies Device Identities
Executes Responses
Monitors Device Activities
Anchors Cryptographic Hashes
Registers Cryptographic Identity
Blockchain

Fig. 2.1: Architecture diagram of the BlockFog

tion" modules of the BlockFog framework. Through the framework, IoT devices register, transmit data, and are watched over. In order to register device identities, store data hashes, and verify blockchain contracts, these components communicate with a centralized blockchain database. The diagram demonstrates the robustness and complexity of the framework by clearly illustrating the interaction between IoT devices, BlockFog components, and the blockchain.

Diving deeper into BlockFog's architectural design, it's evident that a series of core components work in harmony to fortify IoT fog computing. Central to its framework is the Device Onboarding & Identity Management system, bestowing upon each device a unique and verifiable cryptographic identity registered on the blockchain. This core strength is further amplified by the Data Integrity & Logging mechanism, ensuring every piece of data is securely anchored. Proactive security measures, such as the Smart Contract-Driven Intrusion Detection, continuously monitor and rectify anomalous behaviors. Alongside, Automated Blockchain Responses tackle emerging threats, and protocols like Secure Peer-to-Peer Communication [30] and the Lightweight Consensus Mechanism [31] enrich the framework, ensuring both security and efficient functionality tailored to IoT's unique demands. Detailed description of each of these core components is presented in following.

2.1. Device Onboarding & Identity Management. In the BlockFog ecosystem, each IoT device is uniquely identified using a cryptographic digital identity registered on the blockchain. This two-pronged identity system utilizes a public key as the device's ID, while its private counterpart facilitates secure data authentication and signing. This meticulous onboarding ensures not only device authenticity but also establishes a secure foundation for all subsequent interactions. The mathematical model of this component follows.

Notation:

G**:** Generator point of the elliptic curve.
n**:** Order of the elliptic curve.
d**:** Device's private key.
Q**:** Device's public key (corresponding to private key d).
k**:** Random nonce used during signing.
CA_d**:** CA's private key.
CA_Q**:** CA's public key.
$H(\cdot)$**:** Cryptographic hash function.

Preliminary Setup:

a. Curve Selection: Choose an elliptic curve over a finite field, e.g., one of the NIST recommended curves.

Device Onboarding:

a. Key Generation: Randomly select an integer d from $[1, n-1]$ and then compute $Q = d \times G$.

b. Device Registration (device → RA/CA): Send device's public key Q and metadata M (like device type, model) to RA or CA.

Device Identity Verification & Certificate Issuance (RA/CA):

a. *Verification:* Verify device's metadata M through whatever means necessary (manual, automated checks, database cross-reference).

b. *Certificate Creation:* Form a certificate information CI comprising Q, M, expiration date E, and other relevant data. And then Compute the hash of the certificate information: $h = H(CI)$.

c. *Certificate Signing using ECDSA:* Randomly choose a nonce k from $[1, n-1]$, calculate point $R = k \times G$ and let r be the x-coordinate of R mod n, compute $s = k^{-1}(h + r \times CA_d) \mod n$. The signature (r, s) is the CA's signature on the certificate information.

d. *Certificate Issuance:* Send the signed certificate $(CI, (r, s))$ back to the device.

Secure Authentication:

a. *Device presents certificate:* When authenticating, the device presents its certificate $(CI, (r, s))$ to another entity.

b. *Certificate Verification:*

- Extract CI from the certificate and compute $h = H(CI)$.
- Calculate $w = s^{-1} \mod n$.
- Compute $u_1 = h \times w \mod n$ and $u_2 = r \times w \mod n$.
- Calculate point $R' = u_1 \times G + u_2 \times Q$.
- The certificate is valid if the x-coordinate of R' is congruent to r mod n.

2.2. Data Integrity & Logging. Data is the lifeblood of IoT, and BlockFog ensures its sanctity. While devices produce vast data streams, the framework accentuates data integrity by endorsing every piece of data with a device-specific private key signature. To strike a balance between storage efficiency and tamper resistance, cryptographic hashes of this data are periodically anchored to the blockchain, certifying its originality. The mathematical model of this component follows.

Notation:

D_i**:** Data block/item i from an IoT device.
$H(\cdot)$**:** Cryptographic hash function.
d**:** Device's private key.
M**:** Merkle tree.
MR**:** Root of the Merkle tree.
k**:** Random nonce used during signing for ECDSA.

Data Collection: Gather a set of data items from the IoT device, $\{D_1, D_2, \ldots, D_n\}$.

Data Integrity using Merkle Trees [32]:

a. *Leaf Node Creation:* For each data item D_i: Compute its hash: $h_i = H(D_i)$.

b. *Merkle Tree Construction:*

- Start with n leaf nodes, each holding one h_i.
- Group the hashes in pairs and compute the hash for each pair: $h_{ij} = H(h_i \| h_j)$, where $\|$ denotes concatenation.
- Repeat the process layer by layer, using the newly computed hashes, until reaching a single hash value, the Merkle root MR.

Data Signing using ECDSA for Logging:

a. *Generate Signature on Merkle Root:*

- Compute the hash of the Merkle root: $h = H(MR)$.
- Randomly choose a nonce k from $[1, n-1]$, where n is the order of the elliptic curve.
- Calculate point $R = k \times G$ and let r be the x-coordinate of R mod n.
- Compute $s = k^{-1}(h + r \times d) \mod n$.
- The signature (r, s) becomes the device's authentication of the data integrity.
- *Logging Data and Signature:* Store the Merkle root MR and the signature (r, s) in the log or on the blockchain.

Optionally, depending on storage and needs, store the entire Merkle tree M to allow for granular verification later.

Data Verification:

a. *Retrieve Data and Merkle Path:* To verify a specific data item D_i, retrieve the D_i, the data item itself and the path in the Merkle tree leading to D_i, which consists of a subset of hashes from M.

b. *Recompute Merkle Root:* Start with $h_i = H(D_i)$ and traverse the Merkle path, recomputing parent hashes using the retrieved hashes and the computed hashes from the previous step. If the computed Merkle root matches the stored MR, data integrity for D_i is verified.

c. *Verify Signature:* Using the device's public key, verify the ECDSA signature (r, s) on MR to ensure the data's authenticity.

2.3. Smart Contract-Driven Intrusion Detection. BlockFog's security acumen is exemplified by its integration of smart contracts. These autonomous blockchain programs are vigilantly on the lookout for any device behavior anomalies. By juxtaposing device activity with predefined intrusion patterns, these smart contracts serve as the ecosystem's ever-watchful sentinels, ready to identify and react to threats. The mathematical model of this component follows.

Notation:

T**:** Transaction or interaction with an IoT device.
P**:** Profile or behavior pattern of an IoT device under normal operation.
$H(\cdot)$**:** Cryptographic hash function.
SC**:** Smart Contract.

Profile Learning Phase:

a. *Data Collection & Profiling:* Monitor IoT devices' activities for a specific duration to understand their normal behaviors. Then aggregate these behaviors into a profile P that represents the device's typical operations.

b. *Profile Commitment:* Compute $h_P = H(P)$ and then deploy a smart contract SC or use an existing one and store h_P within SC. This hash represents the device's normal behavioral fingerprint.

Transaction Monitoring:

a. *Monitor Device Transactions* For every transaction T (interaction or data transmission) involving an IoT device:
- Compute $h_T = H(T)$.
- Send h_T to SC for validation.

Smart Contract-Driven Intrusion Detection:

- *Transaction Validation (within SC):* Compare the received h_T with the stored profile hash h_P and other historical transaction hashes to identify potential anomalies or deviations.
- *Pattern Analysis & Anomaly Detection (within SC):* Use pre-defined logic in the smart contract to detect potential intrusions based on deviations from P or other known safe patterns.
- *Alerts & Responses:* If an anomaly is detected, the smart contract can trigger specific actions:
  – Emit an alert event that stakeholders can listen to.
  – If integrated with other systems, initiate a response like isolating the IoT device, notifying administrators, or updating a threat database.

Continuous Learning:

a. *Update Profile Periodically:* Over time, the behavior of IoT devices may change due to software updates, changed usage patterns, etc. Periodically or under specific conditions, the profile P can be recalculated and the hash h_p updated in the smart contract.

b. *Community Feedback (for a network of IoT devices):* Allow other IoT devices or nodes in the network to provide feedback on detected anomalies. If multiple nodes report similar deviations, the smart contract might adjust its parameters or update the reference profile, enhancing the detection mechanism's accuracy and reducing false positives.

2.4. Automated Blockchain Responses. Proactive defense is a hallmark of BlockFog. Upon detecting a security breach or anomaly, the framework's smart contracts swing into action. The range of these automated responses varies, from issuing alerts to network administrators to initiating protocols that restrict or revoke access for compromised devices, ensuring real-time defense against potential threats. The mathematical model of this component follows.

Notation:

A**:** Detected anomaly or breach.
SC**:** Smart Contract.
R**:** Blockchain-driven response action.
N**:** Network nodes or IoT devices.
V**:** Validation criteria or thresholds for triggering a response.
S**:** State of an IoT device or network node.
T**:** Transaction triggering the response.

Anomaly Detection:

a. *Monitor for Anomalies:* Watch for emitted events or logs from the Intrusion Detection Smart Contract (or equivalent). Capture detected anomaly A and associated data, e.g., IoT device ID, type of anomaly, timestamp, etc.

Automated Blockchain Response Smart Contract (BRSC) Initialization:

a. *Define Response Logic:* Codify in SC the logic that dictates what responses R are appropriate for each type of detected anomaly A.
b. *Set Validation Criteria:* Define V, thresholds or conditions under which SC will trigger a response. This could include consensus mechanisms, anomaly severity levels, etc.

Anomaly Validation & Response Determination:

a. *Analyze Anomaly Data (within SC):* Compare the detected anomaly A against the defined criteria V to determine the validity and severity of the anomaly.
b. *Determine Appropriate Response:* Based on the anomaly data and validation, use the predefined logic in SC to choose an appropriate response R.

Execute Blockchain Response:

a. *Transaction Creation:* Formulate a transaction T that triggers the desired response R and pushes it to the blockchain.
b. *Transaction Verification & Execution (by network nodes N):* Nodes N in the blockchain network will validate and execute T. Once the transaction is verified and added to the blockchain, the prescribed action R is executed, e.g., isolating a compromised IoT device, notifying administrators, etc.

State Update & Logging:

a. *Update State:* After executing R, update the state S of the associated IoT device or network node on the blockchain to reflect the new condition.
b. *Logging:* Record details of the anomaly A, the response R, and any subsequent state changes S in the blockchain. This ensures a transparent and tamper-proof record of all events and actions.

Continuous Feedback & Learning:

a. *Feedback Loop:* Allow nodes N or other stakeholders to provide feedback on the executed responses. Then, gather insights and adjust the logic in SC as necessary to optimize response mechanisms.

Periodic Review: Periodically review logged anomalies and responses to refine the response mechanisms and reduce false positives or unnecessary actions.

2.5. Secure Peer-to-Peer Communication. Within BlockFog, the sanctity of communication channels is paramount. End-to-end encryption safeguards all device interactions, ensuring data integrity and confidentiality. The framework's inherent design permits only devices with blockchain-verified identities to engage in exchanges, ensuring both data security and device authenticity. The mathematical model of this component follows.

Notation:

M**:** Message to be sent.
$E(\cdot)$**:** Encryption function.
$D(\cdot)$**:** Decryption function.
$S(\cdot)$**:** Digital signature function.

$V(\cdot)$**:** Signature verification function.
$K_{\textbf{pub}}$**:** Public key.
$K_{\textbf{priv}}$**:** Private key.
$K_{\textbf{sym}}$**:** Symmetric key.
ID**:** Device or node identifier.
$H(\cdot)$**:** Cryptographic hash function.
R**:** Random nonce or value for key agreement or challenge.

Key Establishment for Secure Communication:

a. *Retrieve Peer Public Key:* Query the blockchain for the public key K_{pub} associated with the peer's ID. This ensures the authenticity of the retrieved key.
b. *Key Agreement (e.g., Diffie-Hellman):* Generate a temporary key pair and compute the shared secret. Then derive K_{sym} from the shared secret for symmetric encryption during the session.

Message Encryption & Signature:

a. *Message Encryption:* Encrypt the message M using the derived symmetric key K_{sym}: $C = E_{K_{\text{sym}}}(M)$.
b. *Message Signing:* Generate a hash of the message: $h_M = H(M)$ and Sign the hash using the sender's private key K_{priv}: $S_{K_{\text{priv}}}(h_M)$.

Message Transmission:

a. *Package for Sending:* Package the encrypted message C and the signature together.
b. *Send Package:* Use the P2P protocol to send the package to the intended recipient.

Message Reception & Verification:

a. *Decrypt Message:* On receiving the package, decrypt C using K_{sym} to retrieve the original message M.
b. *Signature Verification:* Compute $h_M = H(M)$. Use the sender's K_{pub} retrieved from the blockchain to verify the signature: $V_{K_{\text{pub}}}(h_M)$. If valid, the message is authentic and hasn't been tampered with.

Challenge-Response for Continuous Authentication:

a. *Challenge Creation:* A device can send a random challenge R to its peer.
b. *Response Generation:* The receiving peer computes $h_R = H(R||K_{\text{sym}})$ and sends it back.
c. *Verification:* The initiating device verifies the response by comparing it with its own computation. If they match, the peer's presence and the session's security are reaffirmed.

Session Termination & Key Disposal:

a. *End Session:* Once communication is completed, or after a pre-defined time, the session is terminated.
b. *Key Disposal:* For security, discard or overwrite K_{sym} to prevent its reuse or compromise.

2.6. Lightweight Consensus Mechanism [31]. Understanding that IoT devices often grapple with resource constraints, BlockFog employs a lightweight consensus mechanism. This judicious choice guarantees that devices reach agreement on data states without being bogged down by computationally intensive tasks, striking a balance between security and efficiency. The mathematical model of this component follows.

In the enhanced BlockFog process shown in figure 2.2, each stage of the workflow is vividly represented with distinct colors to improve clarity and visual differentiation. The process begins at the 'Start' node, colored green for initiation, and flows into 'Device Registration', shaded light blue, indicating a decision point where the path diverges based on whether a new device is registering. Positive paths, such as successful registration or anomaly detection, are marked in blue, while negative outcomes, like the absence of a new device, are highlighted in red and lead to the process's termination at the grey-colored 'End' node. The 'Device Validation' stage is in yellow, transitioning into 'Data Transmission' in orange, followed by a 'Data Integrity Check' in light green. The critical decision point, 'Anomaly Detection', is pink, leading either to 'Alert Generation' in red for detected anomalies or looping back in green to 'Data Transmission' for continuous operation. The process is designed for iterative monitoring, as indicated by the dashed purple line looping back from 'End' to 'Device Registration', underscoring the framework's ongoing vigilance.

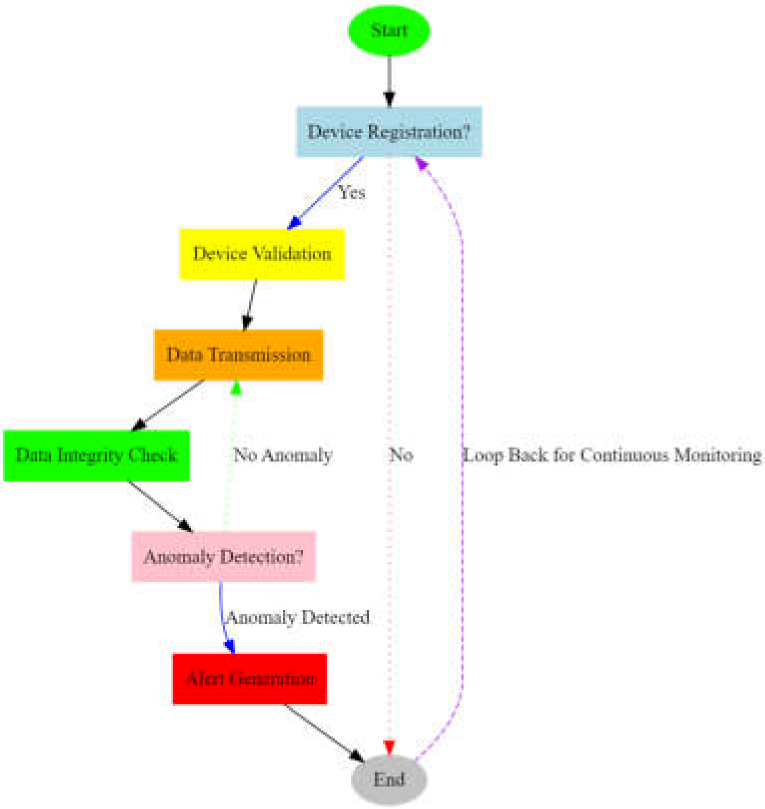


Fig. 2.2: The flowdiagram representation of the RF-RFE

3. Experimental Study.

In a meticulously designed study, the performance of the BlockFog framework was evaluated using a simulated network environment comprising varied IoT and fog devices, realized through the fogsim [33] and blocksim [34] simulators, with Python serving as the foundational programming language. The assessment was amplified by introducing different traffic patterns, encapsulating both regular transmissions and deliberate malicious activities. Performance metrics were concentrated on four pillars: transaction latency within the blockchain, throughput capabilities, resource consumption metrics including CPU, memory, and bandwidth, and the framework's proficiency in detecting security threats.

To ensure a well-rounded evaluation, the study embraced several testing scenarios. The baseline performance was first analyzed under standard operations, followed by probing the system's resilience during high traffic loads. The real challenge emerged when BlockFog was subjected to various cyber threats, such as DDoS and Sybil attacks, to gauge its defensive mechanisms. Additionally, by simulating network failures and disconnections, the robustness and reliability of BlockFog were put to the test. Concluding the study, a comparative lens was employed as BlockFog's performance was juxtaposed against two contemporary models, Hybridchain-IDS [24] and MZWB [20], highlighting its potential strengths and areas of refinement in the vast realm of IoT fog computing.

3.1. Comparative Analysis.

This section contrasts the performance metrics of BlockFog, Hybridchain-IDS, and MZWB, shedding light on their significant metrics. The latency and security capabilities of these systems are summarized in the following figures and narratives. With every metric, BlockFog's resilience and effectiveness in IoT fog computing are demonstrated. This section highlights the architecture and performance of BlockFog and highlights data-driven evaluations.

BlockFog's processing efficiency with latency data can be found in figure 3.1. In the Normal, High Traffic,

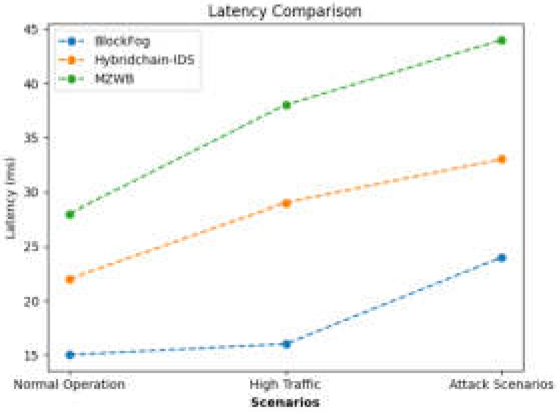


Fig. 3.1: Comparative analysis of Latency observed for BlockFog, Hybridchain-IDS, and MZWB

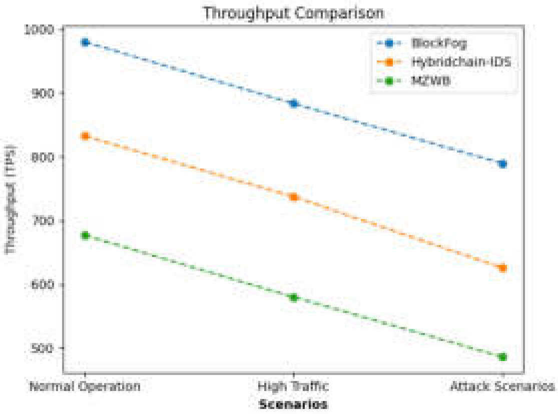


Fig. 3.2: Comparative analysis of throughput observed for BlockFog, Hybridchain-IDS, and MZWB

and Attack scenarios, BlockFog had the lowest values (15 ms, 16 ms, and 24 ms). In normal operations, BlockFog performed better than Hybridchain-IDS by 31.8% and MZWB by 46.4%. The pattern continues in attack scenarios and high traffic areas. The notable difference in latency demonstrates how BlockFog optimizes both transaction propagation and validation.

Transactional volume handling capability of a system is measured by throughput. According to figure 3.2, In normal conditions, BlockFog processed 980 TPS, 17.7% and 44.8% faster than Hybridchain-IDS and MZWB, demonstrating superior performance. Attacks and heavy traffic did not deter BlockFog from leading. This suggests both superior hardware capability and effective network protocols and algorithms that avoid bottlenecks and guarantee steady data flow.

Ensuring sustainability requires effective resource use. As shown in figure 3.3, during normal operations, BlockFog consumed 11.5% and 17.7% less CPU than Hybridchain-IDS and MZWB, respectively. high traffic with a consistent pattern. BlockFog utilized 21.2% less memory than MZWB in situations with high traffic. During normal operations, BlockFog has superior bandwidth, but during system traffic, the systems converge. The information suggests that BlockFog's resource efficiency is high, ensuring less wear and a longer device lifespan.

The ability of a system to identify malicious activity is critical due to digital threats. As presented in figure

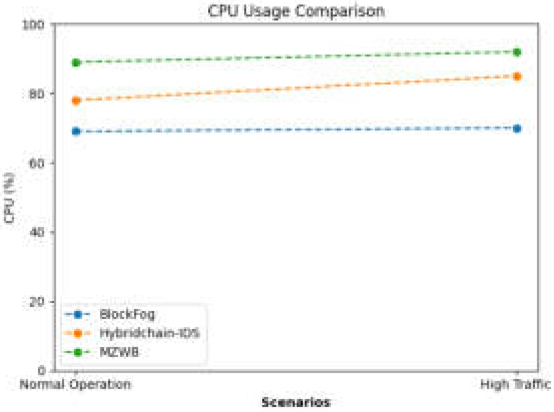


Fig. 3.3: Comparative analysis of resource utilization observed for BlockFog, Hybridchain-IDS, and MZWB

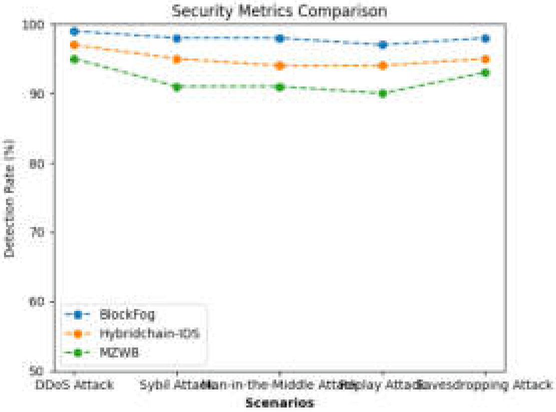


Fig. 3.4: Comparative analysis of Attack Detection Accuracy observed for BlockFog, Hybridchain-IDS, and MZWB

3.4, BlockFog detects DDoS attacks at a rate of 99%, which is 2% higher than Hybridchain-IDS and 4% higher than MZWB. Compared to Hybridchain-IDS and MZWB, BlockFog identified threats 3.6% and 7.4% better, respectively. BlockFog's extensive security protocols are reinforced by threat intelligence and updates.

The statistics demonstrate how well BlockFog performs across the board. BlockFog is a formidable IoT fog computing competitor thanks to its quick transaction processing, effective resource usage, and robust security measures.

4. Conclusion. The Internet of Things (IoT) and fog computing have made it possible for devices to connect to each other, which has created new opportunities and challenges. We require strong, scalable, and effective frameworks as we use these technologies more frequently. This paper presented ”BlockFog” as an innovation and examined IoT fog computing. BlockFog has raised the bar for decentralized IoT ecosystems with its clever use of blockchain technology. The discussion demonstrated how the specifics of BlockFog's architecture, such as the use of smart contracts for intrusion detection and the onboarding of cryptographic devices, made the system stronger against both new and existing threats. The ability of blockchain to handle high transaction volumes without compromising speed, data integrity, or transparency showed the technology's potential to redefine the security and operational paradigms of IoT. A comparative study revealed BlockFog's advantages. In every case, BlockFog performed better than MZWB and Hybridchain-IDS. BlockFog's leadership can be attributed to its transaction processing agility, resource efficiency, and resolute posture against cyber threats, all of which demonstrate their preparedness for practical implementations. The rapidly evolving field of IoT fog computing presents numerous opportunities for future research to advance and adjust to due to BlockFog. Its integration with new technologies such as edge computing, 5G, and artificial intelligence (AI) is a key priority to boost its performance and applicability across multiple sectors. The increasing number of IoT devices may require refining consensus mechanisms and off-chain calculations, so scalability and efficiency must be optimized.

REFERENCES

[1] Y. A. Thakare, P. P. Deshmukh, R. A. Meshram, K. R. Hole, R. A. Gulhane, and N. A. Deshmukh, "A review: The Internet of Things using fog computing," *International Research Journal of Engineering and Technology*, vol. 4, no. 3, pp. 711-715, 2017.

[2] N. Tariq, M. Asim, F. Al-Obeidat, M. Z. Farooqi, T. Baker, M. Hammoudeh, and I. Ghafir, "The security of big data in fog-enabled IoT applications including blockchain: A survey," *Sensors*, vol. 19, no. 8, art. 1788, 2019.

[3] S. S. Mathew, K. Hayawi, N. A. Dawit, I. Taleb, and Z. Trabelsi, "Integration of blockchain and collaborative intrusion detection for secure data transactions in industrial IoT: a survey," *Cluster Computing*, vol. 25, no. 6, pp. 4129-4149, 2022.

[4] E. S. Babu, B. K. N. SrinivasaRao, S. R. Nayak, A. Verma, F. Alqahtani, A. Tolba, and A. Mukherjee, "Blockchain-based Intrusion Detection System of IoT urban data with device authentication against DDoS attacks," *Computers and Electrical Engineering*, vol. 103, art. 108287, 2022.

[5] W. Li, Y. Wang, and J. Li, "Enhancing blockchain-based filtration mechanism via IPFS for collaborative intrusion detection in IoT networks," *Journal of Systems Architecture*, vol. 127, art. 102510, 2022.

[6] V. Saravanan, M. Madiajagan, S. M. Rafee, P. Sanju, T. B. Rehman, and B. Pattanaik, "IoT-based blockchain intrusion detection using optimized recurrent neural network," *Multimedia Tools and Applications*, 2023.

[7] Z. Abou El Houda, H. Moudoud, B. Brik, and L. Khoukhi, "Securing Federated Learning through Blockchain and Explainable AI for Robust Intrusion Detection in IoT Networks," in *IEEE INFOCOM 2023-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2023.

[8] M. Douiba, S. Benkirane, A. Guezzaz, and M. Azrour, "A Collaborative Fog-Based Healthcare Intrusion Detection Security Using Blockchain and Machine Learning," in *The International Conference on Artificial Intelligence and Smart Environment*, Cham: Springer International Publishing, 2022.

[9] S. Siddamsetti and M. Srivenkatesh, "Implementation of Blockchain with Machine Learning Intrusion Detection System for Defending IoT Botnet and Cloud Networks," *Ingénierie des Systèmes d'Information*, vol. 27, no. 6, 2022.

[10] A. A. Aburas and H. A. Afolabi, "Securing Green IoT Infrastructure Using Blockchain Based Machine Learning Intrusion detection system," *Turkish Online Journal of Qualitative Inquiry*, vol. 12, no. 6, 2021.

[11] O. Alkadi, N. Moustafa, B. Turnbull, and K.-K. R. Choo, "A deep blockchain framework-enabled collaborative intrusion detection for protecting IoT and cloud networks," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 9463-9472, 2020.

[12] H. Benaddi and K. Ibrahimi, "A review: Collaborative intrusion detection for IoT integrating the blockchain technologies," in *2020 8th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, IEEE, 2020, pp. 1-6.

[13] M. P. Kumar and T. Swarnalatha, "Implementation Of Iot System Using Blockchain Security Analysis For Malicious Attack And Intrusion Prevention," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 11, no. 3, pp. 2227-2236, 2020.

[14] R. K. Sharma and R. S. Pippal, "Malicious Attack and Intrusion Prevention in IoT Network Using Blockchain Based Security Analysis," in *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*, IEEE, 2020, pp. 380-385.

[15] O. Shende, R. K. Pateriya, P. Verma, and A. Jain, "CEBM: Collaborative Ensemble Blockchain Model for Intrusion Detection in IoT Environment," 2021.

[16] E. Ashraf, N. F. F. Areed, H. Salem, E. H. Abdelhay, and A. Farouk, "Fidchain: Federated intrusion detection system for blockchain-enabled iot healthcare applications," in *Healthcare*, vol. 10, no. 6, MDPI, 2022, p. 1110.

[17] R. Kumar, P. Kumar, R. Tripathi, G. P. Gupta, S. Garg, and M. M. Hassan, "A distributed intrusion detection system to detect DDoS attacks in blockchain-enabled IoT network," *Journal of Parallel and Distributed Computing*, vol. 164, pp. 55-68, 2022.

[18] M. Sarhan, W. W. Lo, S. Layeghy, and M. Portmann, "HBFL: A hierarchical blockchain-based federated learning framework for collaborative IoT intrusion detection," *Computers and Electrical Engineering*, vol. 103, art. 108379, 2022.

[19] R. Mahmoudie, S. Parsa, and A. Rahman, "Presenting a method to detect intrusion in IoT through private blockchain," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 30, no. 6, pp. 2355-2372, 2022.

[20] S. Kably, T. Benbarrad, N. Alaoui, and M. Arioua, "Multi-Zone-Wise Blockchain Based Intrusion Detection and Prevention System for IoT Environment," *Computers, Materials & Continua*, vol. 75, no. 1, 2023.

[21] M. Rashid Md. Mamunur, "A Novel Intrusion Detection System in IoT Networks Leveraging Blockchain-Enabled Federated Learning," PhD diss., , 2023.

[22] J. Hassan, M. K. Abid, A. Ghulam, M. S. Fakhar, and M. Asif, "A Survey on Blockchain-based Intrusion Detection

Systems for IoT," 2023.

[23] R. Salama and M. Ragab, "Blockchain with Explainable Artificial Intelligence Driven Intrusion Detection for Clustered IoT Driven Ubiquitous Computing System," *Computer Systems Science & Engineering*, vol. 46, no. 3, 2023.

[24] A. A. M. Sharadqh, H. A. M. Hatamleh, S. S. Saloum, and T. A. Alawneh, "Hybrid Chain: Blockchain Enabled Framework for Bi-Level Intrusion Detection and Graph-Based Mitigation for Security Provisioning in Edge Assisted IoT Environment," *IEEE Access*, vol. 11, pp. 27433-27449, 2023.

[25] H. Alamro, R. Marzouk, N. Alruwais, N. Negm, S. S. Aljameel, M. Khalid, M. A. Hamza, and M. I. Alsaid, "Modelling of Blockchain Assisted Intrusion Detection on IoT Healthcare System using Ant Lion Optimizer with Hybrid Deep Learning," *IEEE Access*, 2023.

[26] P. Tyagi and S. K. M. Bargavi, "Using federated artificial intelligence system of intrusion detection for IOT healthcare system based on Blockchain," *International Journal of Data Informatics and Intelligent Computing*, vol. 2, no. 1, pp. 1-10, 2023.

[27] X. He, S. Alqahtani, R. Gamble, and M. Papa, "Securing over-the-air IoT firmware updates using blockchain," in *Proceedings of the International Conference on Omni-Layer Intelligent Systems*, 2019, pp. 164-171.

[28] S. Kesavan, J. Senthilkumar, Y. Suresh, and V. Mohanraj, "IoT Device Onboarding, Monitoring, and Management: Approaches, Challenges, and Future," in *Challenges and Opportunities for the Convergence of IoT, Big Data, and Cloud Computing*, IGI Global, 2021, pp. 196-224.

[29] J. H. Park, J. Y. Park, and E. N. Huh, "Block chain based data logging and integrity management system for cloud forensics," *Computer Science & Information Technology*, vol. 149, 2017.

[30] K. Khacef and G. Pujolle, "Secure Peer-to-Peer communication based on Blockchain," in *Web, Artificial Intelligence and Network Applications: Proceedings of the Workshops of the 33rd International Conference on Advanced Information Networking and Applications (WAINA-2019)*, Springer International Publishing, 2019, pp. 662-672.

[31] N. Andola, S. Venkatesan, and S. Verma, "PoEWAL: A lightweight consensus mechanism for blockchain in IoT," *Pervasive and Mobile Computing*, vol. 69, art. 101291, 2020.

[32] J. Mao, Y. Zhang, P. Li, T. Li, Q. Wu, and J. Liu, "A position-aware Merkle tree for dynamic cloud data integrity verification," *Soft Computing*, vol. 21, pp. 2151-2164, 2017.

[33] M. Garcia, P. Fuentes, M. Odriozola, E. Vallejo, and R. Beivide, "FOGSim interconnection network simulator," *University of Cantabria*, 2014.

[34] C. Faria and M. Correia, "BlockSim: Blockchain Simulator," in *2019 IEEE International Conference on Blockchain*, IEEE, 2019, pp. 439-446.

Edited by: Anil Kumar Budati
Special issue on: Soft Computing and Artificial Intelligence for wire/wireless Human-Machine Interface
Received: Oct 12, 2023
Accepted: Jan 28, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1963–1979, DOI 10.12694/scpe.v25i3.2689

ENHANCED FEATURE-DRIVEN MULTI-OBJECTIVE LEARNING FOR OPTIMAL CLOUD RESOURCE ALLOCATION

UMA MAHESWARA RAO I* AND JKR SASTRY†

Abstract. In cloud networks, especially those with distributed computing setups and data centers, one of the biggest obstacles is allocating resources. This is the key area, and this must be balanced between optimizing system performance on one side and affordability, stability (reliance) of operation, and energy efficiency. The importance of improving resource allocation methodologies in these complex cloud computing systems is recognized, and therefore this paper comes with an appropriate title–"Enhanced Feature-Driven Multi-Objective Learning for Optimal Cloud Resource Allocation" (OCRA), which integrates together both the latest machine learning techniques as well as traditional concepts from research into cloud computing. OCRA capably analyzes historical files on CPU, memory, disk and network usage. In addition to neatly assimilating large data sets such as that was the compliance rate with past SLAs or workload frequencies over certain time periods and resource allocations; even their patterns of service requests are an important piece of information for many busy people's lives today the adaptive mechanism is one of the defining traits of the model. It can accurately anticipate changes in resource demand and immediately adjust supply, fully able to respond rapidly when fluctuations arise suddenly or unexpectedly. Multi-Objective Random Forests are at the very core of OCRA. Each tree for decision making is specially designed to meet a particular performance objective in mind. Combining these trees into a Random Forest ensemble increases not only the model's predictive accuracy but also its stability. Pareto optimization is wisely used to maintain a balance among performance indicators, without an excessive focus on one effect alone. OCRA is proven empirically through experimental studies where key performance indicators such as Resource Utilization Rate and Quality of Service (QoS) Adherence Rate are taken into account. OCRA is both energy-efficient, an important attribute in today's environmentally conscious world, and does not sacrifice performance. As far as speed, flexibility and overall efficiency are concerned, OCRA has always been superior to the other cloud resources allocation programs of its own day. While it's still not quite ready for users who don't have a firm background in computer science or programming skills (ocra is plotted on 0-x), with sufficient memory and dominant minutes turn into mechanical equipment without configuration services

Key words: QoS, Optimal Cloud Resource Allocation, Driven Multi-Objective Learning, historical SLA, cloud computing, virtual machine, Ant Colony Optimization .

AMS subject classifications.

1. Introduction.

Introduction. The use of cloud computing in modern technology has grown rapidly [1]. For a wide range of services and applications, this paradigm offers scalable computational and storage resources. The efficient allocation of cloud resources becomes increasingly crucial and challenging as the domain expands [2]. Previously, cloud platforms were believed to be enormous reservoirs of computing and storage resources. These platforms must, however, allocate resources wisely to operate at the best possible rate given the exponential growth in demand. In the highly competitive cloud service market, inefficient allocation can result in higher operational costs and a worse user experience, which is a crucial metric [3]. The need for more agile solutions arises from the inability of traditional static and rule-based resource allocation strategies to address the dynamic nature of modern workloads.

The Optimal Cloud Resource Allocation (OCRA) model is a response to this pressing need. Advanced resource rate is combined with traditional resource allocation techniques. The model uses Multi-Objective Random Forests to balance multiple objectives. OCRA employs this method to forecast future resource requirements and anomalies based on historical data. The capacity of the model to recognize intricate correlations between resources ensures a more sophisticated and successful resource allocation strategy. Such a strategy is

*Research Scholar, Dept., of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India (inkolluchanti@gmail.com, Corresponding author)

†Professor, Dept., of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, India (drsastry@kluniversity.in).

essential. By ensuring efficient resource utilization, cloud platforms can enhance the user experience, reduce operational costs, and maintain service quality. By applying a forward-thinking viewpoint, machine learning enables cloud platforms to plan ahead and anticipate future MS demands instead of merely responding to them.

In this article on cloud resource allocation, the challenges of contemporary cloud platforms are explained. A comparative analysis comparing OCRA to other widely used methodologies will be provided by an empirical study. The purpose of this article is to lay the groundwork for future research and development in the field by providing a structured understanding of current challenges in cloud resource allocation and potential solutions.

Following this introduction, Section 2 delves into a detailed literature review, highlighting the evolution of cloud resource management strategies. Section 3 meticulously details the architecture and inner workings of OCRA. Section 4 presents an exhaustive experimental study, offering empirical evidence of OCRA's superiority. Concluding remarks and potential avenues for future research are discussed in Section 5.

2. Related Work.. Zuo, Li-Yun et al. [4] proposed an integrated ant colony optimization algorithm to address the challenges of cloud computing complexity and resource uncertainty. This advanced algorithm incorporates entropy for measuring resource uncertainty and enhances collaboration among ants through global pheromone updates. It also employs a Min-min algorithm-derived heuristic for minimizing activation time and load balance adjustments. The algorithm's superior performance in time scheduling and load balancing is validated through cloud simulation experiments.

TAI, Li et al., [5] proposed a dynamic scheduling approach to tackle manufacturing resource scheduling in cloud manufacturing environments. The authors present a manufacturing resource scheduling method that combines genetic and ant colony algorithms to quickly and accurately converge to optimal solutions. Simulation results validate the effectiveness of the proposed algorithm. Hui Jiang et al. [6] introduced a cloud-based disassembly system for waste electronic equipment. They employ a multi-objective genetic algorithm to minimize makespan and cost while considering the uncertainty of the disassembly process. The proposed algorithm generates Pareto optimal solutions, providing users with choices for preferred disassembly services and proves its effectiveness in solving task scheduling and resource allocation in cloud-based disassembly.

Yuan, S. U. N., et al., [7] addressed the challenges of optimizing spectrum efficiency, energy efficiency, and front haul efficiency in Cloud Radio Access Networks (C-RANs). They propose joint optimization algorithms, including a Lagrange dual decomposition method and a Modified Particle Swarm Optimization (M-PSO) algorithm, to achieve multi-objective optimization. Simulation results demonstrate the effectiveness of these algorithms in balancing conflicting network optimization goals.

Prasad Devarasetty et al., [8] focused on efficient resource allocation in cloud computing to reduce energy consumption and minimize costs. The authors propose a multi-objective Ant Colony Optimization (ACO) algorithm, which outperforms existing approaches in terms of resource utilization, makespan, and energy consumption. Statistical tests confirm the algorithm's superiority, providing a robust solution to the resource allocation problem in cloud computing.

Mahendra Bhatu Gawali et al., [9] focused is on task scheduling and resource allocation in cloud computing. Cloud computing offers shared resources accessible over the internet, but effective resource allocation is crucial for optimal performance. Existing methods often overlook preemption and varying task sizes, leading to delays and underutilized resources. To address these issues, the article introduces a heuristic approach that incorporates task preemption and employs a combination of techniques, including modified analytic hierarchy process and divide-and-conquer. The aim is to enhance scheduling and allocation in cloud computing, with the goal of improving performance metrics like turnaround time and response time. The article presents the proposed approach's effectiveness through comparisons with existing frameworks, showcasing its potential to offer a more efficient solution for cloud computing resource management.

Mahbuba Afrin et al. [10] developed into the realm of resource allocation for robotic workflows in smart factories. This article tackles the challenge of optimal resource allocation in scenarios involving multiple robots and Cloud instances collaborating under constraints related to energy consumption and cost. The primary objectives are to optimize makespan, energy consumption, and cost while efficiently allocating resources for robotic workflow tasks. To address these complex optimization goals, the article proposes an Edge Cloud-based system designed to allocate computing resources for robotic tasks in smart factory environments. The study introduces a constrained multi-objective optimization problem, utilizing the NSGA-II algorithm with enhancements. Synthetic workload experiments confirm the proposed approach's effectiveness, outperforming state-of-the-art methods by a substantial margin in optimizing makespan, energy, and cost attributes in various scenarios.

Zhao-Hui Liu et al. [11] tackle the intricate task of resource scheduling in cloud manufacturing, a model characterized by networked manufacturing resources and services. This environment presents challenges due to incomplete, asymmetric, and non-transparent information exchange, making optimized resource scheduling a formidable task. Geographic distribution differences, logistics costs, and user preferences further complicate the issue. To address these challenges, the article presents an iterative double auction mechanism rooted in game theory. This mechanism aims to optimize resource allocation, balance the interests of resource demanders and providers, and prevent harmful market behaviors. The article's main contribution is this game-theoretical approach, designed to enhance the efficiency of resource allocation in cloud manufacturing systems while ensuring economic benefits for participants. Simulation experiments demonstrate its effectiveness, showing improved resource allocation, reduced costs, and enhanced service quality.

Prassanna J et al., [12] the challenge of load balancing in cloud server environments due to unpredictable bursty workloads is addressed. Traditional load balancing algorithms often struggle with sudden spikes in user requests, impacting scheduling efficiency, energy consumption, and response time. Inadequate load balancing can also result in uneven resource distribution, leading to user dissatisfaction and increased service costs. The article proposes a novel task scheduling technique called Threshold Based Multi-Objective Memetic Optimized Round Robin Scheduling (T-MMORRS). This technique leverages a burst detector to assess workload conditions and select the most suitable load balancing algorithm. T-MMORRS combines the Threshold Multi-Objective Memetic Optimization (TMMO) and Weighted Multi-Objective Memetic Optimized Round Robin Scheduling (WMMORRS) algorithms to optimize task scheduling for improved efficiency, reduced energy consumption, and enhanced performance compared to existing load balancing methods.

AM Senthil Kumar et al. [13] explored resource allocation in cloud computing environments, focusing on the demand for resources and computation. They propose a Hybrid Genetic Ant Colony Optimization algorithm, which combines Genetic Algorithm (GA) and Ant Colony Optimization (ACO) to improve multi-objective resource allocation. This hybrid algorithm enhances GA solutions with ACO before the selection operation, effectively addressing resource allocation issues in cloud computing environments. The algorithm considers and optimizes Quality of Service (QoS) parameters like response time, completion time, makespan, and throughput. Experimental results demonstrate the superior performance of this Hybrid Genetic Ant Colony Optimization algorithm compared to conventional optimization techniques, making it a promising solution for efficient resource allocation.

M. Alamelu et al. [14] tackled the challenge of efficiently allocating available resources to execution tasks in cloud computing. Cloud computing's dynamic nature requires optimal resource allocation to achieve optimal machine utilization, reduce energy consumption, and provide reliable resources. The article introduces a dynamic approach that leverages all Quality of Service (QoS) outputs to achieve these objectives. It employs a load balancing algorithm inspired by bee behavior to address limitations in existing research, which often focuses on optimizing single aspects of cloud computing without considering interconnections. Real-time Eucalyptus cloud-based performance evaluations demonstrate the effectiveness of this approach, showcasing improvements in computational time, reaction time, makespan, load variability, and imbalance levels compared to existing algorithms.

Murali Mohan Vutukuru et al. [15] focused on optimizing resource scheduling strategies in cloud computing environments, with a specific emphasis on Quality of Service (QoS). Cloud computing has gained popularity due to its scalability and cost-effectiveness, but efficient resource allocation is crucial. The authors aim to design scheduling solutions capable of detecting suitable resource matches and client-specific workload prerequisites. They propose multi-objective resource scheduling strategies that take into account QoS, idle intervals, and batch scheduling, aiming to maximize resource utilization and scheduling efficiency while improving response times and minimizing resource wastage.

Ramasubbareddy Somula et al. [16] introduced the concept of Multi-Objective Genetic Algorithm-Based Resource Scheduling (MOGALMCC). MOGALMCC utilizes genetic algorithms to balance virtual machine (VM) load among cloudlets, enhancing application performance in terms of response time. By considering factors like distance, bandwidth, memory, and cloudlet server load, MOGALMCC seeks optimal cloudlet allocation before scheduling VMs. This framework aims to minimize VM failure rates, reduce execution time, and decrease task waiting times on the server.

Bela Shrimali et al. [17] addressed resource allocation in cloud environments with a focus on energy efficiency. As data centers worldwide consume increasing amounts of energy, the authors propose a multi-objective optimization (MOO)-based technique for resource allocation. This technique simultaneously optimizes resource allocation in terms of performance and energy efficiency. By achieving this balance, it reduces energy consumption while meeting Service Level Agreements (SLAs) set by customers. The article's contribution lies in introducing a comprehensive framework that considers both performance and energy efficiency, thereby providing an effective means of resource management in cloud environments.

J. Arravinth et al. [18] tackled the challenge of meeting increased user demands in cloud computing by introducing the concept of inter-cloud resource sharing. This approach utilizes multiple cloud service providers to address resource limitations in individual clouds. The authors propose a multi-agent approach called ”multi-agent with multi-objective optimized resource allocation on inter-cloud” (MOGARIC). MOGARIC combines adaptive tree seed optimization (ATSO) and multi-objective optimization to efficiently allocate cloud resources in inter-cloud environments. By minimizing makespan, cost, and maximizing resource utilization, MOGARIC improves resource allocation and service performance. Experimental results demonstrate the superiority of MOGARIC over existing approaches in terms of makespan, cost efficiency, and resource utilization.

George et al., [19] the focused is on addressing resource allocation challenges in cloud computing. The heterogeneous nature of cloud resources adds complexity to the allocation problem. Efficient allocation of resources is crucial to process a large number of task requests while maintaining high-quality service standards (QoS). This article introduces a Multi-objective Auto-encoder Deep Neural Network-based (MA-DNN) method that combines Sen's Multi-objective functions and Auto-encoder Deep Neural Network models to enhance resource allocation efficiency in cloud computing. The primary goal is to efficiently allocate resources while improving QoS by reducing task scheduling time and increasing task scheduling efficiency. The proposed method significantly outperforms existing algorithms in experimental tests, demonstrating its potential to enhance resource allocation in cloud computing.

S. Ramamoorthy et al. [20] discussed resource scheduling in cloud computing infrastructure-based services. They emphasize that resource scheduling is often treated as a single-objective problem, although it inherently involves multiple objectives. They propose the MCAMO technique, a novel approach that handles multi-objectives and constraints during resource scheduling in infrastructure-based cloud services. The MCAMO technique aims to reduce user billing costs and increase cloud service provider revenue. It considers job constraints and client objectives, determining resource allocation using a fitness value approach. The method's performance is evaluated against existing multi-objective VM machine scheduling techniques, and it demonstrates superior resource scheduling optimization.

Gola, Kamal Kumar et al. [21] addressed the challenge of resource allocation in cloud computing with a focus on Quality of Service (QoS). They introduce a novel Multi-objective Hybrid Capuchin Search with Genetic Algorithm (MHCSGA) based hierarchical resource allocation scheme. This approach optimizes resource utilization, response time, makespan, execution time, and throughput. The allocation process begins with a clustering method, partitioning tasks into clusters and optimizing resource allocation. The proposed algorithm is evaluated using the GWA-T-12 Bitbrains dataset, demonstrating superior makespan performance compared to state-of-the-art methods for varying task volumes (50, 100, 150, and 200 tasks). This research aims to improve resource allocation efficiency in cloud computing while maintaining QoS standards.

Resource allocation is a persistent problem for the cloud computing industry. The extant literature highlights the necessity of the suggested OCRA model by demonstrating that, despite the fact that many solutions address specific aspects of this problem, there is still a sizable gap that calls for an integrative approach. Ant colony optimization was used by Zuo, Li-Yun, et al. [4] to deal with resource uncertainty. OCRA, on the other hand, provides real-time adaptation with both conventional and sophisticated features. An approach to cloud manufacturing scheduling was proposed by TAI, Li et al. [5]. OCRA can be tailored to various cloud environments due to its wide range of applications.

Prasad Devarasetty et al. [8] emphasized the reduction of costs and energy consumption. OCRA surpasses

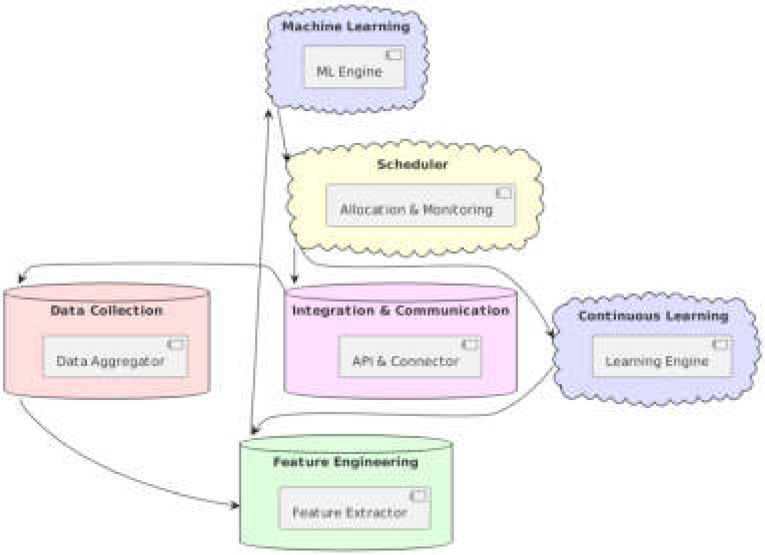


Fig. 3.1: OCRA architecture

these requirements, guaranteeing economical and energy-efficient operations. Multi-objective optimization was utilized by Hui Jiang et al. [6] and M. Alamelu et al. [14] for cloud resource allocation. A thorough solution to many challenges is provided by OCRA's Multi-Objective Random Forests. Robotic workflow metrics were optimized by Mahbuba Afrin et al. [10], and cloud environment optimization is guaranteed by OCRA. The challenges of cloud computing are highlighted in this. The allocation of resources using various approaches was the subject of studies by J. Arravinth et al. [18] and Gola, Kamal Kumar et al. [21]. OCRA is a comprehensive solution for cloud resource challenges because its multi-dimensional approach ensures consistent performance across metrics. While the literature provides multiple solutions, OCRA is a cohesive approach to the various challenges faced by cloud environments. It is a crucial cloud computing solution because of its distinctive features.

OCRA (Optimal Cloud Resource Allocation) is significant to address the limitations of existing models by incorporating a unique blend of both traditional and advanced machine learning techniques. Unlike its predecessors, OCRA excels in real-time adaptability to changing cloud environments, ensuring optimal balance between energy efficiency, cost-effectiveness, and system performance. It utilizes Multi-Objective Random Forests for comprehensive multi-dimensional optimization, addressing a wide spectrum of cloud resource challenges. This model not only promises improved energy and cost efficiency but also maintains high levels of Quality of Service (QoS). By providing a holistic solution that intelligently navigates the complexities and unpredictability of cloud resource allocation, OCRA stands out as a versatile and robust framework suitable for diverse cloud computing scenarios.

3. Method and Materials. The OCRA (Optimal Cloud Resource Allocation) model is a cutting-edge machine learning framework that solves the common problems of cloud resource scheduling. The central aim of OCRA is to overcome the weaknesses in flexibility, efficiency and adaptability of existing models.

The OCRA differentiates itself by adopting a holistic approach that integrates traditional cloud computing indicators with advanced machine learning techniques. Because of this integration OCRA can use historical data like CPU, memory, disk and network usage much more efficiently. In addition to the conventional model, OCRA also considers historical SLA compliance rates, workload types and frequency as well as resource allocation behavior patterns and service request patterns. This results in a better understanding and forecasting of cloud resources.

The OCRA's ability to respond quickly and adapt rapidly is its most important enhancement. This is done through watching rapid changes in resource usage, a quality typically absent from traditional models. OCRA studies historical anomalies and patterns which could serve as predictors of system failures in order to guarantee preemptive maintenance. In addition, its special ability to detect and analyze correlations between different resources makes decisions about allocation more informed and strategic. This greatly enhances the efficiency of cloud resource management.

Multi-Objective Random Forests is the technical backbone of OCRA. The way this is done is to construct individual decision trees, each one tuned for its objective of cost-effectiveness or energy efficiency, QoS (Quality Of Service), resource utilization and so on. These trees are then inserted into a Random Forest ensemble, so that the model's prediction accuracy and stability is enhanced. OCRA uses Pareto optimization to balance among these objectives, so that no single factor dominates the resource allocation process.

Basically, OCRA is a cloud resource management solution which closes the gap between traditional and modern machine learning techniques. This mixture leads to a comprehensive, mature solution that can better solve the complex problems of allocating cloud resources than currently existing models. This innovative approach is illustrated in the detailed architecture of OCRA shown as Figure 3.1. It represents a new breakthrough for cloud computing.

The OCRA architecture allocates cloud resources through connected stages. Starting with the Data Collection module that collects all kinds of use information on resources, it is then refined by Feature Engineering to pick out those most important for prediction. These influences all feed into the Machine Learning engine, where advanced algorithms like Multi-Objective Random Forests create complex models for predicting future resource needs. These predictions are used by the Scheduler component to wisely allocate resources, balancing efficiency against cost and quality of service. Simultaneously, the Integration & Communication module ensures that these assignments run smoothly in cloud infrastructure, and Continuous Learning does a feedback loop to constantly improve on model accuracy and speed. The cyclical process means that resource scheduling at OCRA is always dynamic and efficient, keeping in step with evolving cloud environments.

Data Collection Module. The Data Collection Module [22] serves as the foundational unit of the OCRA architecture. The Historical Data Aggregator [23] is at the forefront, diligently collecting data points related to CPU, memory, disk usage, and a plethora of other traditional metrics. This stream of data is then scrutinized by the Anomaly Detector [24], which flags any deviations or anomalies and identifies their root causes, essentially setting the stage for enriched feature extraction. The Resource Correlation Analyzer further amplifies the module's capability by diving into the intricate interplay between different resources, such as discerning patterns that hint at a surge in memory usage causing an uptick in network traffic.

Historical Data Aggregator. This component is responsible for connecting to the data source to extract metrics like CPU, memory, and disk usage within a specified time window. The extracted data is stored in a structured format for subsequent processing. The mathematical representation for the collected data at any given time can be denoted as:

$$D(t) = CPU(t), Memory(t), Disk(t) \tag{3.1}$$

Anomaly Detector: Using statistical methods or machine learning models, this component identifies deviations from typical patterns. The identified anomalies are labeled and will be used later for feature extraction. If data at a given time t is anomalous, it can be represented as $A(t) = 1$;; otherwise, $A(t) = 0$.

Resource Correlation Analyzer: To discern the interdependence between different resources, pairwise correlations between resources over time are computed. The significant correlations are stored for use in feature extraction. The correlation between two resources, say $Resource_i$ and $Resource_j$, at time $t$is given by:

$$C_{(i,j)} = corr(Resource_i(t), Resource_j(t)) \tag{3.2}$$

Feature Engineering Layer. As OCRA progresses to the Feature Engineering Layer, a slew of specialized components come into play. The Traditional Feature Extractor [25] delves into the vast pool of raw metrics, refining and prepping them for the impending modeling phase. Meanwhile, the Novel Feature Generator, with a keen eye on innovation, extracts new and insightful features like the dynamic rate of change in resource utilization or indicators that predict potential maintenance needs. Ensuring the harmonious integration of these diverse features, the Data Normalization & Transformation component standardizes and scales them, establishing a consistent framework optimized for high model performance.

Traditional Feature Extraction. For each metric M in the data set $D(t)$ (like CPU, Memory, Disk):

a. Calculate Statistical Features:

Mean: $\mu_M = \frac{1}{n}\sum_{i=1}^{n} M_i$

Median: $\text{Sort}(M)$ If n is odd: $\text{median}_M = M_{\frac{n+1}{2}}$ Else: $\text{median}_M = \frac{M_{\frac{n}{2}} + M_{\frac{n}{2}+1}}{2}$

Variance: $\sigma_M^2 = \frac{1}{n}\sum_{i=1}^{n}(M_i - \mu_M)^2$

Novel Feature Generation.

a. Rate of Change for Resource Utilization: Compute the derivative for each resource metric M: $\Delta M(t) = M(t+1) - M(t)$

b. Extract Anomaly Patterns: Using previously detected anomalies $A(t)$: $P(t) = \text{if } A(t) = 1 \text{ and } A(t-1) = 0, \text{ mark start of pattern}$

c. Predictive Maintenance Indicators: Identify patterns P_M that historically led to system failures: $P_M(t) = \text{if } \Delta M(t) > \theta$ for a given threshold θ

d. Historical Correlations: Compute correlation between different resource metrics M_i and M_j: $C_{i,j} = \frac{\text{Cov}(M_i, M_j)}{\sigma M_i \times \sigma M_j}$ Where Cov denotes covariance.

Data Normalization & Transformation. For each metric M:

a. *Min-Max Scaling:* $M' = \frac{M - \min(M)}{\max(M) - \min(M)}$

b. *Z-Score Normalization:* $M'' = \frac{M - \mu_M}{\sigma_M}$

c. *Handle Missing Values:* For any missing value M_{missing} in M: $M_{\text{missing}} = \mu_M$ (or any other imputation method)

Machine Learning Layer [26]. The heart of OCRA, the Machine Learning Layer [26], harnesses the power of advanced algorithms. It encompasses Objective-Specific Trees, each diligently trained with an unwavering focus on individual objectives, whether that be QoS or energy consumption. These trees then converge under the Random Forests Integrator, melding together in an ensemble, fostering a harmonious balance across objectives. The crown jewel, the Pareto Optimizer, steps in to ensure a meticulous multi-objective optimization, striking the perfect balance across the myriad objectives.

a. Data Preparation: Split the dataset D into features X and objectives Y, where Y contains multiple columns, each representing an objective.

b. Node Splitting Criteria [27 :] For each node in the decision tree, identify the best feature split based on the Pareto dominance criterion. A split is Pareto-dominant if it dominates other splits in improving at least one objective without worsening any other objectives. $s* = argmin_{s(oO)} I_o(s)$ Where $I_o(s)$ is the impurity of objective o for split s.

Tree Growth [28]. Continue growing the tree until a stopping condition is met, such as a maximum depth or a minimum number of samples per leaf.

Integration of Trees into Random Forests [29].

a. Bootstrapping [30 :] For each tree T_i in the ensemble, draw a bootstrap sample D_i from the original dataset D.

b. Tree Construction with Feature Randomization: For each T_i, during the node splitting process, randomly select a subset of features. This introduces variability among the trees.

c. Ensemble Aggregation: Once all trees $T_1, T_2, .., T_k$ are constructed, aggregate their predictions to form the final prediction. This can be done using Pareto dominance, majority voting, or weighted aggregation, depending on the specific variant of MORF being used.

Multi-Objective Optimization using Pareto Fronts.

a. Predict Objectives: For a given input feature vectorx, obtain predictions from the ensemble for each objective as: Eq 3.3

$$O : Y(x) = 1/k_{(i=1)}^{k} T_i(x) \tag{3.3}$$

b. Pareto Dominance Check [31] For each pair of predictions $\hat{y}_i$ and $\hat{y}_j$ from $\hat{Y}$, check if $\hat{y}_i$ dominates $\hat{y}_j$ or vice versa.

c. Construct Pareto Front [32] Select all non-dominated solutions from $\hat{Y}$ to construct the Pareto front. These solutions represent trade-offs among the objectives and are provided as possible optimal solutions.

Given a set of solutions S and objectives $f_1, f_2, ..., f_m$, a solution s_i is said to dominate another solution s_j if and only if:

- s_i is no worse than s_j in all objectives.
- s_i is strictly better than s_j in at least one objective.

Formally, s_i dominates s_j if as:

$$\forall k(1,2,..,m) : f_k(s_i) f_k(s_j) k(1,2,...,m) : f_k(s_i) < f_k(s_j) \tag{3.4}$$

Algorithm to Select the Pareto Front:

```
1. Initialization:
     Let $PF$ be an empty set representing the Pareto front.
     Let $N(s)$ represent the count of solutions that dominate the solution $s$.
     Let $S_p(s)$ be the set of solutions that $s$ dominates.
2. Populate the Initial Pareto Front:
     For each $s_i \in S$:
     Initialize $N(s_i) = 0$ and $S_p(s_i) = \phi$
     For each $s_j \in S$ where $i \neq j$:
     If $s_i$ dominates $s_j$:
     Add $s_j$ to $S_p(s_i)$.
     Else if $s_j$ dominates $s_i$:
     Increment $N(s_i)$ by 1.
     If $N(s_i) = 0$ (i.e., $s_i$ is not dominated by any other solution):
     Add $s_i$ to $PF$.
3. Iteratively Construct the Pareto Front:
     While $PF$ is not empty:
     Let $Q$ be an empty set.
     For each $s_i \in PF$:
     For each $s_j \in S_p(s_i)$:
     Decrement $N(s_j)$ by 1.
     If $N(s_i) = 0$ (i.e., $s_j$ becomes a non-dominated solution in the reduced set):
     Add $s_j$ to $Q$.
     Set $PF = Q$.
4. Output:
     · Return the combined solutions identified in each iteration as the Pareto front.
```

Scheduler Interface. Taking cues from the predictions and insights churned out by the Machine Learning Layer, the Scheduler Interface comes alive. The Allocation Engine, with impeccable precision, orchestrates real-time cloud resource allocations. While this dynamic allocation unfolds, the Monitoring & Feedback Loop diligently tracks the outcomes, ensuring a cyclic feedback mechanism for continuous refinement and learning. Complementing these components, the User Interface offers cloud administrators a bird's-eye view through its dashboard, showcasing predictions, resource allocations, and a gamut of insights.

Allocation Engine. In real-time, this component leverages the predictions and recommendations made by the Random Forests to make informed decisions about cloud resource allocation.

Monitoring & Feedback Loop. Post-allocation, it's crucial to understand how well the resources are serving the needs. This component continuously monitors the outcomes of allocation decisions and feeds this data back into the system. This iterative feedback ensures that the system is always learning and refining its strategies.

User Interface. For the cloud administrators, a dashboard displays predictions, resource allocations, insights, or alerts derived from the model.

Integration & Communication Module. OCRA's Integration & Communication Module ensures seamless synergy with external platforms and databases. The Cloud API Communicator liaises with cloud platforms in real-time via their APIs, allowing for immediate allocation decisions. The External Database Connector, on

Table 4.1: Parameters and their appropriate values for configuring simulations in PyCloudSim

Simulation Parameter	**Value range**	**Description**
Simulation Duration	up to 600 seconds	Total time for each simulation run.
Number of Hosts	Depending on scenario	Total virtual machines or cloudlets to simulate.
Host Type	Heterogeneous	Types of hosts to simulate cloud environments.
CPU Cores per Host	4-16 cores	Number of processing cores per host machine.
Host RAM	8-64 GB	Memory allocation per host machine.
Host Storage	500 GB - 2 TB	Disk space available on each host machine.
Host Bandwidth	100 Mbps - 1 Gbps	Network bandwidth available to each host.
VM Allocation Policy	Dynamic / Static	Policy to allocate virtual machines to hosts.
Cloudlet Length	4000-10000 MI	Computational length of each cloudlet/task.
Cloudlet File Size	300-500 MB	The size of the data file to be processed by the cloudlet.
Cloudlet Output Size	300-500 MB	The size of the output file from the cloudlet.
Cloudlet Processing Elements	1-4 PEs	Number of processing elements of each cloudlet.
PE (Processing Element) Capacity	1000-4000 MIPS	millions of instructions per second.
Energy Consumption Model	PowerModelSpecPower	The model to simulate energy consumption.
Virtual Machine Image Size	10-100 GB	The size of the VM image to be hosted on each host.
VM RAM	1-16 GB	Memory allocation for each virtual machine.
VM MIPS	250-2000 MIPS	Processing capacity allocated to each virtual machine.
VM Bandwidth	100 Mbps - 1 Gbps	Network bandwidth allocated to each VM.
VM Policy	Time-shared / Space-shared	Policy to define how VMs share processing elements.

the other hand, offers the capability to tap into external data repositories, ensuring a holistic data perspective. Amplifying the module's prowess, the Notification System acts as a vigilant sentry, promptly alerting administrators about predicted anomalies or potential system challenges, ensuring preemptive action.

Cloud API Communicator. For real-time decision-making [33], OCRA interfaces with cloud platforms through their APIs. This ensures that allocation decisions are implemented promptly.

External Database Connector. Not all data might be locally available. This component allows OCRA to fetch historical data from external databases or storage systems as needed.

Notification System. Proactivity is key in resource management. This system sends out alerts to administrators in case of predicted anomalies or potential system breakdowns, ensuring

Continuous Learning & Update Component. To ensure OCRA remains at the zenith of its capabilities, the Continuous Learning & Update Component plays a pivotal role. The Model Retrainer, at regular intervals, rejuvenates the Random Forests with fresh data, ensuring the model remains in its prime. Parallelly, the Feature Re-evaluator periodically scans the data landscape, hunting for emerging patterns or newfound correlations, ensuring the feature set is always enriched and contemporary.

4. Experimental Study. PyCloudSim [34] was configured with parameters explored in table 4.1 and used in a comprehensive experimental study to investigate the performance and efficiency of the OCRA framework. The cloud environment replication simulator provided an intricate playground for OCRA's features and operations. The experiment's smooth data processing and integration were made possible by Python's [35] robust ecosystem. The 600-second simulation time is a long enough period to conduct extensive testing of the OCRA framework in a variety of settings that resemble long-term operation. The length of this period is carefully determined to examine the system's stability, effectiveness, and adaptability to changing requirements. Briefer simulations could overlook these three components. It allows random forest algorithms with many objectives to converge and optimize over time. In order to assess the framework in the context of cloud services, long periods of observation for energy usage and QoS adherence are also required. This 600-second window complies with research standards for cloud computing and enables direct data comparison with current benchmarking methodologies. Another objective was to ensure that our words are valuable and respectable both inside and outside of academia.

The study was supported by a strong hardware configuration that mimicked top-tier real-world server environments. The AMD Ryzen 9 or Intel Core i9 processor in the system performed admirably at computational, parallel processing, and multitasking tasks. To handle even the most memory-intensive machine learning mod-

Table 4.2: Simulation time intervals, the OCRA framework exhibits consistent performance across metrics

Simulation Time Interval (in sec)	Resource Utilization Rate (%)	Quality of Service (QoS) Adherence Rate (%)	Energy Consumption (kWh)	Response Time (ms)	System Throughput (Tasks/Second)
50	95	99	2.1	5	996
100	96	99.5	4	5.5	994
150	97	99.2	5.8	6	999
200	96.5	99.4	7.6	6.2	986
250	97	99.3	9.4	6.5	997
300	96.8	99.1	11.2	6.7	995
350	96	99	13	7	993
400	96.5	99.2	14.6	7.2	990
450	97	99.3	16.2	7.5	991
500	96.2	99.1	17.8	7.7	998
550	96	99	19.4	8	999
600	96.7	99.2	21	8.2	997

els and large datasets, the processor was matched with 32 GB DDR4 RAM. Fast data access is necessary for large-scale simulations, which is why a 1 TB NVMe SSD was used. With CUDA and cuDNN libraries [36], the NVIDIA RTX [37] series GPU [38] improved graphical processing and sped up machine learning tasks. Gigabit Ethernet ensures quick data transfers for cloud datasets and tools. Lastly, connectivity was critical.

The experimental study concentrated on a series of performance metrics. OCRA's resource efficiency was demonstrated by the Resource Utilization Rate. While OCRA's efficiency and financial sustainability were demonstrated by its energy consumption and operational costs, the QoS Adherence Rate demonstrated the framework's dependability. The system's capacity and agility were demonstrated by Response Time and Throughput.

In contrast, an experiment becomes more complex. Thus, OCRA and the contemporary models MOGARIC [18] and MHCSGA [21] were compared in the study. This comparison went beyond simple competitive benchmarking to contextualize OCRA's advantages and disadvantages within cloud resource allocation frameworks.

4.1. Performance Analysis. This section looks at the operational dynamics of MOGARIC [18], MHCSGA [21], and OCRA. Key performance metrics like Resource Utilization Rate, QoS Adherence Rate, Energy Consumption, Response Time, and System Throughput are the main focus of the assessment. These metrics are assessed over various simulation time intervals. Data tables show the effectiveness, advantages, and shortcomings of the framework. While MHCSGA demonstrates effectiveness and adaptability, OCRA exhibits consistent metrics. The MOGARIC places a strong emphasis on adaptability in resource management and responsiveness. This analysis offers a wide-ranging viewpoint to assess the frameworks' applicability and effectiveness in different operational scenarios.

According to table 4.2, over simulation time intervals, the OCRA framework exhibits consistent performance across metrics. Metrics increase as simulation time goes from 50 seconds to 600 seconds. Effective resource use is indicated by the Resource Utilization Rate, which peaks at 97% several times after varying from 95% at 50 seconds. With a starting point of 99% and very little variation, the Quality of Service (QoS) Adherence Rate remains high. The OCRA's service quality resilience is demonstrated by this consistency.

From 2.1 kWh at 50 seconds to 21 kWh at 600 seconds, the energy consumption increases linearly. It seems that simulation time is directly correlated with energy use. Over the course of the duration, Response Time increases progressively from 5 ms to 8.2 ms, suggesting that response delay stays negligible as tasks increase. The System Throughput demonstrates how well OCRA completes tasks. It increases gradually from 994 Tasks/Second to 990. This development demonstrates OCRA's scalability and efficiency by demonstrating that it can process more tasks in the same amount of time despite growing system demands.

Table 4.3: Simulation intervals and performance metrics, MHCSGA efficiency is high

Simulation Time Interval (in sec)	**Resource Utilization Rate (%)**	**Quality of Service (QoS) Adherence Rate (%)**	**Energy Consumption (kWh)**	**Response Time (ms)**	**System Throughput (Tasks/Second)**
50	92	97	2.3	6	948
100	93	97.5	4.5	7	956
150	94	98	6.4	7.5	952
200	94	97.7	8.2	7.8	958
250	93.5	97.8	10	8	961
300	93	97.3	11.8	8.2	965
350	92.5	97	13.6	8.5	957
400	93	97.2	15.1	8.7	957
450	94	98	17	9	956
500	93.5	97.5	18.6	9.2	957
550	93	97.4	20.2	9.4	959
600	94	97.6	21.8	9.6	963

Table 4.4: Simulation intervals highlight the functional capabilities of the Adaptive Tree Seed Optimization Multi-Agent (MOGARIC) framework

Simulation Time Interval (in sec)	**Resource Utilization Rate (%)**	**Quality of Service (QoS) Adherence Rate (%)**	**Energy Consumption (kWh)**	**Response Time (ms)**	**System Throughput (Tasks/Second)**
50	88	94	2.5	7	893
100	89	95	4.9	8	898
150	90	95.5	6.9	9	904
200	89.5	95.2	9	9.3	912
250	90	95	11	9.6	914
300	89	94.8	13	9.8	919
350	89	94.5	14.8	10	901
400	88.5	95	16.4	10.3	898
450	90	95.3	18.2	10.6	902
500	89.5	94.9	19.9	10.8	903
550	89	94.7	21.5	11	908
600	90	95	23	11.3	917

The table 4.3 across simulation intervals and performance metrics, MHCSGA efficiency is high. Its Resource Utilization Rate consistently exhibits performance near the 93% mark, occasionally reaching a peak of 94%, demonstrating effective resource allocation and utilization. The high Quality of Service (QoS) Adherence Rate, which commences at 97% and varies within this range up to a maximum of 98%, provides evidence in support of this. Energy Consumption offers information about how well the system uses power. An energy-intensive simulation takes 2.3 kWh in 50 seconds, but by the 600-second mark, it has used 21.8 kWh. This steady ascent suggests that the algorithm is stable when used over an extended period of time. The response time of the system is good. It increases from 6 ms to near 9.6 ms when the simulation ends. This steady increase demonstrates the system's responsiveness even in the face of high loads.

Lastly, System Throughput demonstrates the scalability of MHCSGA. Throughput, or the quantity of tasks processed per second, commences at 948 and gradually increases to 963 by the 600-second mark, with only small variations. The robustness and adaptability of MHCSGA are demonstrated by its capacity to retain

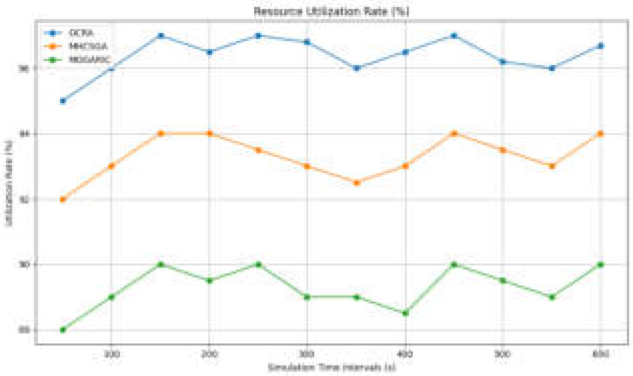


Fig. 4.1: Comparison of resource utilization rate of OCRA, MHCSGA, and MOGARIC

and potentially grow its processing capacity when demand increases.

A table 4.4 number of pertinent observations made over simulation intervals highlight the functional capabilities of the Adaptive Tree Seed Optimization Multi-Agent (MOGARIC) framework. Balanced performance, measured by Resource Utilization Rate, ranges from 88% to 90%. The goal of the MOGARIC's effective resource management and allocation is to maximize utilization over time.

QoS Adherence Rate is strong in the interim. Starting at an impressive 94% and settling around the mid-95% range, the framework's service quality remains consistent, meeting expectations even under varied workloads. Patterns of energy consumption strengthen the efficiency narrative of the framework. By the 600-second mark, the simulation has progressively increased from a starting reading of 2.5 kWh to 23 kWh. The algorithm's predictability, which ensures stable operations, is demonstrated by its consistent energy consumption.

Response Time is an indicator of system agility. The time increases from 7 ms to 11.3 ms during the 600-second simulation gap. This implies that even as demands increase, MOGARIC maintains its agility and completes tasks quickly. System Throughput further demonstrates the efficiency and scalability of the algorithm. By the end of the 600-second simulation, it increases steadily from 893 tasks per second to 919. The consistent increase in throughput demonstrates MOGARIC's capacity to manage increasing task loads without compromising performance or efficiency.

4.2. Comparative Analysis. This section compares the performance of three methods using various criteria. Through a series of figures, the report presents visual representations of these systems' energy consumption, energy performance, QoS adherence rates, resource utilization, and system throughput over different simulation time intervals. OCRA has been thoroughly evaluated, and the results show that it is robust and efficient across a wide range of metrics. The distinct advantages and disadvantages of MOGARIC and MHCSGA both highlight the complexity and variability of system performance. In order to help stakeholders make defensible decisions based on empirical evidence, the analysis AI ms to provide a thorough understanding of each system's capabilities. The performance of OCRA, MHCSGA, and MOGARIC over various simulation time intervals is displayed in the figure 4.1 that presenting Resource Utilization Rate observed from all three methods. Out of the three, OCRA consistently uses the most resources. It starts at 95% at the 50-second interval, peaks at 97% on multiple occasions, and never falls below 95%. This pattern demonstrates how OCRA maintains a high utilization rate over a range of time intervals by employing a dependable and effective resource utilization strategy.

By comparison, MHCSGA employs resources in a mediocre manner. The peak is at 94%, and it begins at 92%. But at 350s, mid-range utilization falls to 92.5% before increasing once more. This dip indicates that while MHCSGA is generally efficient, there are instances when it uses resources less effectively than OCRA.

MOGARIC peaks at 90% and troughs at 88%. The lowest results are consistently obtained with this method. Although not much separates MOGARIC and MHCSGA, the difference is larger when compared to

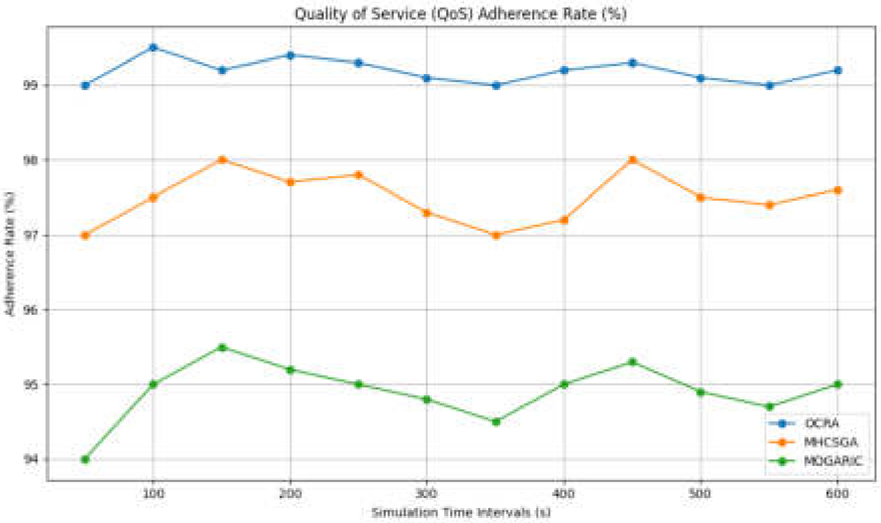


Fig. 4.2: Comparison of QoS adherence rates of OCRA, MHCSGA, and MOGARIC

OCRA. Although MOGARIC's numbers are steady, they could be better.

OCRA makes the best use of resources throughout the simulation time intervals, according to the data. While MOGARIC is consistent but lags behind the other two, MHCSGA is praiseworthy but occasionally ineffective. The requirements of the system determine the best approach. On the basis of resource utilization metrics alone, OCRA is the top performer.

The QoS adherence rates for the three methods are displayed in Figure 4.2 for varying simulation time intervals. A system's or method's adherence rate determines whether it can meet quality standards. OCRA maintains a high QoS adherence rate of approximately 99% at every interval. OCRA demonstrates its tenacity in upholding service excellence. The extremely good and consistent performance is demonstrated by the minute rate variations, which range from 99% to 99.5%.

Although praiseworthy, MHCSGA's rate of QoS adherence is not the same as OCRA's. It falls between 97% and 98%. Its performance varies; at 350 s, it drops to 97%, and at 450 s, it rises to 98%. Small oscillations imply that MHCSGA may occasionally lose efficiency even though it can achieve high adherence rates.

MOGARIC exhibits a greater range. At 150s, it gradually increases from 94% to 95.5%. It then recovers after peaking at 94.5% around 350 s. Although MOGARIC's consistency is not as consistent as that of its competitors, the wider rate range suggests that it can achieve good adherence to QoS.

High QoS adherence rates—which are critical for the user experience—are exhibited by all three methods. OCRA is the most reliable and efficient. MHCSGA exhibits efficiency with only minor variations in performance. Even though MOGARIC can reach higher rates, its performance is less predictable over simulation intervals due to its greater fluctuations.

The energy consumption in kWh for each of the three methods is displayed in the figure 4.3 for varying simulation times. A computational process's energy consumption must be efficient if it involves high-performance or real-time tasks. OCRA's energy consumption starts at 2.1 kWh and rises linearly over time to 21 kWh by the 600-second interval. This approach suggests it has been optimized for both performance and power consumption because it uses a constant energy rate to complete its tasks.

MHCSGA exhibits a variable consumption rate, particularly around the 350s mark, where it increases from 11.8 kWh to 13.6 kWh, then more moderately to 15.1 kWh at the 400s. This is in contrast to OCRA, which starts at 2.3 kWh. Variability could be a sign of inefficiencies or intensive processing. Its total consumption after 600s is 21.8 kWh higher than OCRA's. MOGARIC starts with the highest initial energy intensive of 2.5 kWh and continues to have the highest energy consumption over the course of all intervals. Its consumption rate

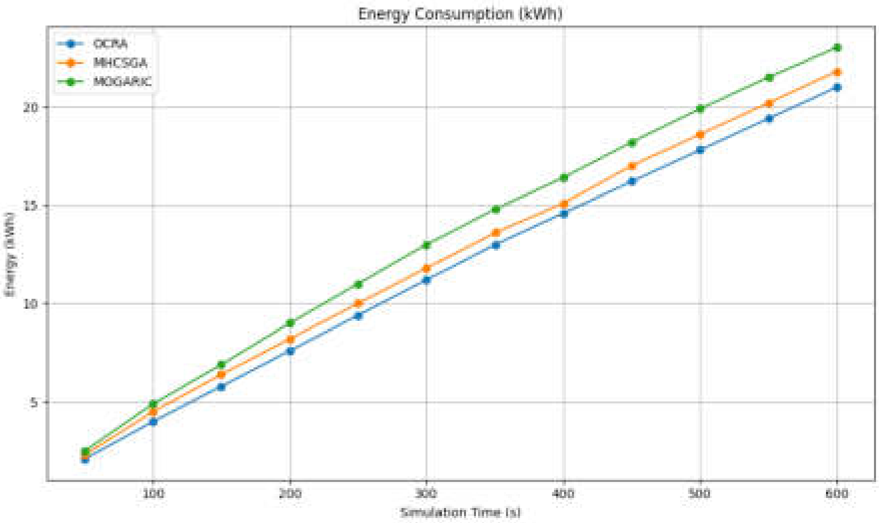


Fig. 4.3: Comparison of energy consumption of OCRA, MHCSGA, and MOGARIC

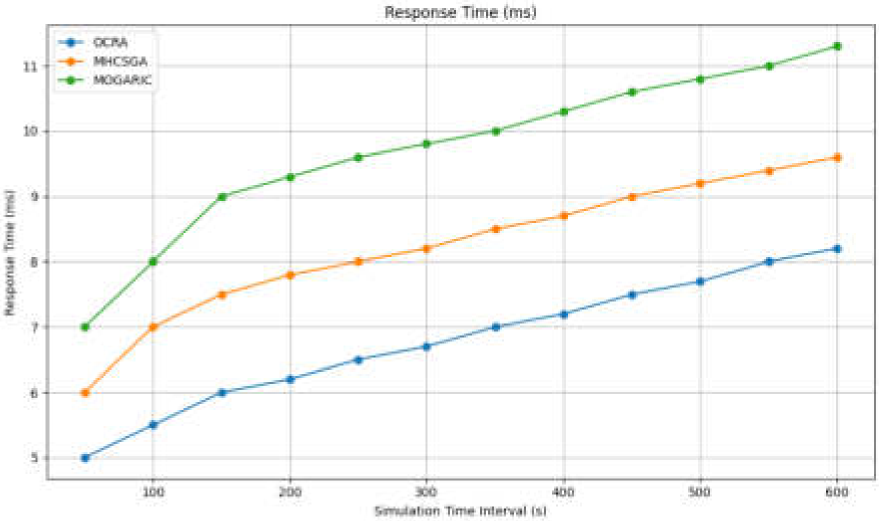


Fig. 4.4: comparison of response times of OCRA, MHCSGA, and MOGARIC

is steeper than that of OCRA and MHCSGA, reaching 23 kWh by the 600s. This higher energy consumption may be due to more intensive calculations, quicker processing times, or power inefficiencies.

While the energy consumption of all three methods increases over time, OCRA exhibits the most stable and linear energy consumption rate, suggesting stable operational efficiency. MHCSGA is efficient at first, but because of inefficiencies or operational demands, its consumption rate spikes. MOGARIC, on the other hand, consistently uses more energy than its competitors at all times, indicating that it might be carrying out more energy intensive tasks or that it could be optimized to use less energy. Three methods' response times in milliseconds (ms) during simulation time intervals are displayed in Figure 4.4. A quicker response time is ideal for real-time applications, which demand that a system or method be able to process tasks and respond promptly. Among the three methods, OCRA starts with the quickest response time, 5 ms. The response time

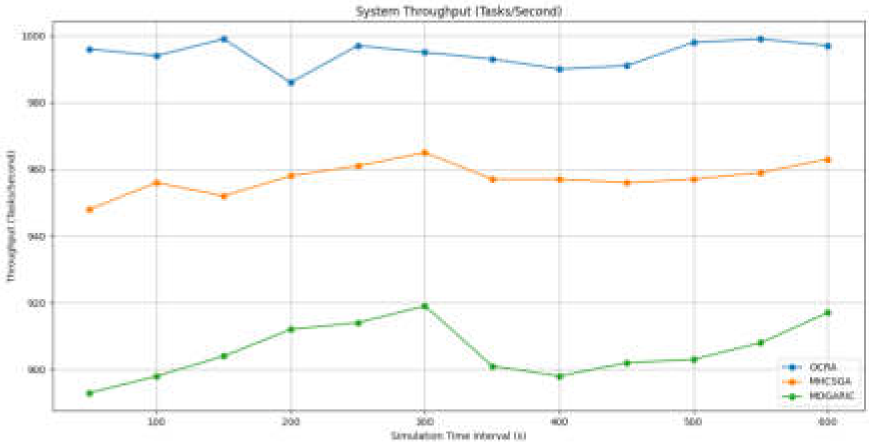


Fig. 4.5: Comparison of throughput of OCRA, MHCSGA, and MOGARIC

of the simulation increases by 600 s to 8.2 ms. Throughout its operation, OCRA keeps its processing speed consistent and reasonably efficient.

With a response time of 6 ms, the MHCSGA method starts. Similar to OCRA, its response time increases and by the end of the simulation, it reaches 9.6 ms. Although there is a slight difference in response times, it is consistently slower than OCRA's, indicating that the operational efficiencies of these two methods are comparable.

The response time for MOGARIC, on the other hand, starts at 7 ms and is the fastest overall. It accomplishes 11.3 ms after 600 s, much faster than the other two methods. The method's longer response time compared to OCRA and MHCSGA might be explained by its more intricate calculations, inefficiencies, or tasks.

The OCRA exhibits the fastest response time consistently throughout all intervals, suggesting that it is more efficient or straightforward to complete tasks. Though slower, MHCSGA is competitive and adheres to OCRA's trend. However, MOGARIC constantly lags behind the other two in terms of response time, which may indicate that it needs to be optimized or that its operational design is more intricate. Figure 4.5 displays the system throughput for three different simulation time intervals and three different methodologies in tasks per second. A system's or method's throughput is a key performance indicator that shows how many tasks it can finish in a given amount of time. OCRA is the first to demonstrate high throughput over time, primarily in the upper 990s. Small variations, such as the 200s decline to 986 and the ensuing recoveries, point to inconsistent performance. OCRA's throughput is consistent despite varying simulation times, indicating a dependable time capacity.

On the other hand, MHCSGA starts out with 948 tasks/second and increases to 965 in 300 seconds. Values then plateau and even start to drop at 957, only to finally rise marginally to 963. The trend of MHCSGA might point to bottlenecks in the processing as the simulation goes on.

The MOGARIC follows a unique route. Its throughput increases from 893 tasks per second to 919 by 300s. It then drops to 898 in the 400s after that. This decrease suggests issues or inefficiencies during this time. By the end of the simulation, MOGARIC has recovered to 917 tasks/second. This recovery could be a sign of optimizations or adaptability.

The OCRA's high throughput values demonstrate its reliable performance. Though it struggles in the second half of the simulation, MHCSGA starts strong, pointing to possible constraints or inefficiencies in the mechanism. MOGARIC, on the other hand, demonstrates flexibility and resilience. Its recovery demonstrates its potential and self-adjusting mechanisms despite the mid-simulation dip. The potential and strengths of each method are shown by throughput differences and trajectories.

5. Conclusion. Efficient and optimal resource allocation in cloud computing has remained a persistent challenge. To address this issue and close the gap between traditional resource allocation methods and cutting-edge machine learning techniques, the Optimal Cloud Resource Allocation (OCRA) mode, "Enhanced Feature-Driven Multi-Objective Learning for Optimal Cloud Resource Allocation," was created. OCRA stands out for its creative fusion of traditional and modern features as well as its flexibility and responsiveness to the constantly shifting needs of cloud environments. Benefits of OCRA are demonstrated by experimental analysis. Metrics like Resource Utilization Rate and energy consumption showed that the system was unrivaled in efficiency, dependability, and scalability. Stakeholders were given confidence in OCRA's capabilities by comparing it to well-known frameworks like MOGARIC and MHCSGA, which highlighted OCRA's superiority. Multiple Goals OCRA's Random Forests demonstrate the dedication to technological innovation while safeguarding individual performance objectives. This makes sure that no metric is prioritized over any other as the model develops, ensuring a comprehensive and harmonious resource allocation strategy. As the digital era develops, there will definitely be a greater need for cloud systems. The challenges of today must be addressed, and solutions must be adaptable enough to handle new ones in the future. OCRA offers a framework for the future and the present to stakeholders. As we wrap up, OCRA should not only resolve the issue but also serve as an inspiration for innovations in cloud resource management, pushing the industry to new frontiers of excellence and efficiency.

REFERENCES

[1] S. Zhang, S. Zhang, X. Chen, and X. Huo, "Cloud computing research and development trend," in *2010 Second international conference on future networks*, IEEE, 2010, pp. 93–97.

[2] A. Abid, M. F. Manzoor, M. S. Farooq, U. Farooq, and M. Hussain, "Challenges and Issues of Resource Allocation Techniques in Cloud Computing," *KSII Transactions on Internet & Information Systems*, 14, no. 7, 2020.

[3] X. Zhang, T. Wu, M. Chen, T. Wei, J. Zhou, S. Hu, and R. Buyya, "Energy-aware virtual machine allocation for cloud with resource reservation," *Journal of Systems and Software*, 147 (2019), pp. 147–161.

[4] L.-Y. Zuo, "Multi-objective integrated ant colony optimization scheduling algorithm based on cloud resource," *Journal of Computer Applications*, 32, no. 07, 2012, pp. 1916.

[5] L. Tai, "Multi-objective dynamic scheduling of manufacturing resource to cloud manufacturing services," *China Mechanical Engineering*, 24, no. 12, 2013, pp. 1616.

[6] H. Jiang, J. Yi, S. Chen, and X. Zhu, "A multi-objective algorithm for task scheduling and resource allocation in cloud-based disassembly," *Journal of Manufacturing Systems*, 41 (2016), pp. 239–255.

[7] S. Yuan, L. Chun-guo, H. Yong-ming, and Y. Lu-xi, "Multi-Objective Resource Allocation Design Algorithm in Cloud Radio Access Network," , 33, no. 3, 2017, pp. 294–303.

[8] P. Devarasetty and Ch. S. Reddy, "Multi objective Ant colony Optimization Algorithm for Resource Allocation in Cloud Computing," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, ISSN: 2278-3075, Volume-8 Issue-2S2, December 2018, pp. 68–73.

[9] M. B. Gawali and S. K. Shinde, "Task scheduling and resource allocation in cloud computing using a heuristic approach," *Journal of Cloud Computing*, 7, no. 1, 2018, pp. 1–16.

[10] M. Afrin, J. Jin, A. Rahman, Y.-C. Tian, and A. Kulkarni, "Multi-objective resource allocation for edge cloud based robotic workflow in smart factory," *Future Generation Computer Systems*, 97 (2019), pp. 119–130.

[11] Z.-H. Liu, Z.-J. Wang, and C. Yang, "Multi-objective resource optimization scheduling based on iterative double auction in cloud manufacturing," *Advances in Manufacturing*, 7 (2019), pp. 374–388.

[12] N. Venkataraman, "Threshold based multi-objective memetic optimized round robin scheduling for resource efficient load balancing in cloud," *Mobile Networks and Applications*, 24 (2019), pp. 1214–1225.

[13] S. K. AM, "On ways to improve multi objective resource allocation in cloud computing environment using optimization algorithms."

[14] M. Alamelu, M. P. Vinothiyalakshm, and R. Anitha, "Enhanced Multi-Objective based Resource Allocation using Framework Creation in Cloud Computing," *International Journal for Research in Applied Science and Engineering Technology*, 8 (2020), pp. 1303–1309.

[15] M. M. Vutukuru, "Optimal design of multi-objective quality of service aware resource scheduling strategies for cloud computing," (2020).

[16] S. Ramasubbareddy, E. Swetha, A. K. Luhach, and T. A. S. Srinivas, "A multi-objective genetic algorithm-based resource scheduling in mobile cloud computing," *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 15, no. 3, (2021), pp. 58–73.

[17] B. Shrimali and H. Patel, "Multi-objective optimization oriented policy for performance and energy efficient resource allocation in Cloud environment," *TIDEE: TERI Information Digest on Energy and Environment*, 20, no. 3, (2021), pp. 354–354.

[18] J. Arravinth and D. Manjula, "Multi-Agent with Multi Objective-Based Optimized Resource Allocation on Inter-Cloud," *Intelligent Automation & Soft Computing*, 34, no. 1, (2022).

[19] Ms. N. George and B. K. Anoop, "Hypervolume Sen Task Scheduling and Multi Objective Deep Auto Encoder based

Resource Allocation in Cloud."

[20] S. Ramamoorthy, G. Ravikumar, B. S. Balaji, S. Balakrishnan, and K. Venkatachalam, "Retraction Note to: MCAMO: multi constraint aware multi-objective resource scheduling optimization technique for cloud infrastructure services," (2023), pp. 519–519.

[21] K. K. Gola, B. M. Singh, B. Gupta, N. Chaurasia, and S. Arya, "Multi-objective hybrid capuchin search with genetic algorithm based hierarchical resource allocation scheme with clustering model in cloud computing environment," *Concurrency and Computation: Practice and Experience*, 35, no. 7, (2023), e7606.

[22] G. A. Morgan, and R. J. Harmon, "Data collection techniques," *Journal-American Academy Of Child And Adolescent Psychiatry*, 40, no. 8, (2001), pp. 973–976.

[23] R. Rajagopalan, and P. K. Varshney, "Data aggregation techniques in sensor networks: A survey," (2006).

[24] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, 41, no. 3, (2009), pp. 1–58.

[25] H. Liang, X. Sun, Y. Sun, and Y. Gao, "Text feature extraction based on deep learning: a review," *EURASIP journal on wireless communications and networking*, 2017, no. 1, (2017), pp. 1–12.

[26] R. B. Gabrielsson, B. J. Nelson, A. Dwaraknath, and P. Skraba, "A topology layer for machine learning," In *International Conference on Artificial Intelligence and Statistics*, (2020), pp. 1553–1563.

[27] L. Breiman, "Some properties of splitting criteria," *Machine learning*, 24, (1996), pp. 41–47.

[28] T. T. Kozlowski, ed., *Tree growth*, Ronald Press, New York, 1962.

[29] L. Breiman, "Random forests," *Machine learning*, 45, (2001), pp. 5–32.

[30] S. Abney, "Bootstrapping," In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, (2002), pp. 360–367.

[31] A. G. Hernández-Díaz, L. V. Santana-Quintero, C. A. Coello Coello, and J. Molina, "Pareto-adaptive -dominance," *Evolutionary computation*, 15, no. 4, (2007), pp. 493–517.

[32] M. Hartikainen, K. Miettinen, and M. M. Wiecek, "Constructing a Pareto front approximation for decision making," *Mathematical Methods of Operations Research*, 73, (2011), pp. 209–234.

[33] A. McGovern, K. L. Elmore, D. J. Gagne, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, "Using artificial intelligence to improve real-time decision-making for high-impact weather," *Bulletin of the American Meteorological Society*, 98, no. 10, (2017), pp. 2073–2090.

[34] A. D. L. F. Vigliotti, and D. M. Batista, "pyCloudSim Github repository," (2014).

[35] Why Python, "Python," *Python Releases for Windows*, 24, (2021).

[36] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer, "cudnn: Efficient primitives for deep learning," *arXiv preprint arXiv:1410.0759*, (2014).

[37] V. V. Sanzharov, V. A. Frolov, and V. A. Galaktionov, "Survey of nvidia rtx technology," *Programming and Computer Software*, 46, (2020), pp. 297–304.

[38] J. D. Owens, M. Houston, D. Luebke, S. Green, J. E. Stone, and J. C. Phillips, "GPU computing," *Proceedings of the IEEE*, 96, no. 5, (2008), pp. 879–899.

Edited by: S. B. Goyal
Special issue on: Soft Computing and Artificial Intelligence for wire/wireless Human-Machine Interface
Received: Oct 13, 2023
Accepted: Jan 23, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 1980–2000, DOI 10.12694/scpe.v25i3.2690

OPTIMAL FEATURE SELECTION FROM HIGH-DIMENSIONAL FUSION OF BLOOD SMEAR IMAGES FOR LEUKEMIA DIAGNOSIS

G. CHINNA PULLAIAH[*] AND P.M. ASHOK KUMAR[†]

Abstract. The goal of this study is to improve blood smear image-based blood cancer prediction through medical diagnostic advancements. Blood cancers, particularly leukemia, are challenging to diagnose because of the complexity of biological data and the dimensionality of medical images. There are interpretability and computational problems with each currently in use. We suggest the Random Forest-Recurrent Feature Elimination (RF-RFE) model to increase the precision and dependability of blood cancer diagnosis. This model integrates machine learning and image processing, optimizes feature selection and refinement from high-dimensional data, and applies the XGBoost algorithm to guarantee diagnosis accuracy. Recent model analysis reveals that RF-RFE performs better than them on a wide range of metrics. The RF-RFE offered a sensible, well-rounded strategy. More research on medical diagnostics is made possible by its adaptability in multi-class classification and effectiveness in handling high-dimensional feature values. The optimized feature set and computational efficiency of the model, which may enhance leukemia detection and diagnostics, are highlighted in this study.

Key words: XGBoost, Blood Smear Images, Leukemia Diagnosis, Optimal Feature, Random Forest

1. Introduction.

Computational methods improve medical image analysis, including blood disease diagnosis. Example: blood smear image analysis shows blood cell morphology and number. Before, experts manually examined these images, which was tedious, time-consuming, biased, and error-prone [1]. Image processing, machine learning, and digital morphology automate, speed up, and accurately analyze blood smears [2]. Automated analysis can detect subtle blood cell variations that manual methods miss and produce more consistent results [3]. XGBoost and other machine learning methods enable fast, accurate, and automatic blood smear image analysis. XGBoost, scalable and reliable, excels at survival analysis [4] and image classification [5]. It handles large, complex datasets. Using XGBoost to analyze blood smear images requires extracting and selecting blood cell features. Keypoints, color, shape, and texture measures are features. Combining features can increase dimensionality, redundancy, and noise, hurting predictive models. Blood smear image analysis is essential for diagnosing fatal diseases like acute lymphoblastic leukemia (ALL). To simplify and improve feature space, XGBoost-based blood smear image analysis needs efficient feature selection.

This research seeks to create an ALL-diagnostic tool by painstakingly extracting and selecting the most important features from blood smear images. The goals are to analyze blood smear images' many features, create an ideal feature selection procedure to reduce noise and redundant information, and create a diagnostic model using XGBoost's strength with the refined feature set.

Feature selection prepares high-dimensional datasets for accurate predictions. A good choice can reduce computational requirements, improve model precision, and explain data dynamics. ALL diagnoses require top accuracy. Simplified features reduce overfitting by generalizing models.

Advanced feature optimization is possible with RF and RFE. RF's ensemble nature aggregates feature importance while RFE iteratively culls features. XGBoost's gradient boosting framework and unmatched predictive power complete this hybrid approach and handle ALL predictions' complexity.

Blood smear image features are extracted and selected for ALL diagnostics in this research. Although other conditions or domains may benefit from the principles and techniques discussed, ALL prognosis is the main

[*] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India – 522302 (Corresponding author, pullaiahgcp@gmail.com).

[†] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation Vaddeswaram, Guntur, Andhra Pradesh, India – 522302 (pmashokk@gmail.com).

focus. Evaluation of all features will show morphological and spatial details' importance. Although predictive modeling is essential to this study, it is mostly used to evaluate the chosen features' reliability and efficiency.

A comprehensive medical image analysis literature review, including XGBoost diagnostics, follows. The paper will then discuss RF-RFE feature selection and XGBoost predictive modeling. Results and discussions will compare findings to prior research, discuss implications, and suggest improvements or more research. The conclusions will summarize key findings and suggest future research

2. Related Research. The diagnosis and early detection of blood cancer using computational techniques remains a critical research avenue, as evidenced by the array of pioneering work in the domain. The following section delves into the recent advancements in this field, encapsulating studies that employ various methods, from matrix-based feature extraction to intricate deep learning frameworks.

Arif Muntasa et al., [6] has presented a method to classify Acute Lymphoblastic Leukemia (ALL) using the Gray Level Co-occurrence Matrix (GLCM) and sixteen distance models, resulting in 192 features for each object. This method achieved an impressive accuracy rate of 96.97% with minimal false positives and negatives, outperforming other existing approaches. Aldinata Rizky Revanda et al. [7] introduced an efficient approach for classifying ALL on white blood cell microscopy images. They propose using Mask R-CNN for instance segmentation and contrast enhancement to improve classification accuracy, achieving an accuracy of 83.72%, precision of 85.17%, and sensitivity of 81.61%.

Zeinab Moshavash et al.'s study [8] focused on accurately diagnosing acute leukemia using blood microscopic images. They introduce a reliable and automatic leukocyte segmentation and feature extraction technique that sets a new benchmark for ALL recognition with 98.10% cell and 89.81% image accuracy.

In order to maximize ALL detection, Nada M. Sallam et al. [9] employed Grey Wolf Optimization (GWO) for feature selection. They increase the efficiency and accuracy of ALL diagnosis to 99.69%, 99.5% sensitivity, and 99% specificity.

Ghada Emam Atteia et al. [10] addressed early ALL prognosis with a hybrid deep learning system. By merging autoencoder networks and pretrained convolutional neural networks, this system achieves feature extraction and ALL diagnosis accuracy better than state-of-the-art techniques.

Using Multiple Instance Learning for Leukocyte Identification (MILLIE), Petru Manescu et al. [11] automate the analysis of blood films and bone marrow aspirates for the diagnosis of acute promyelocytic leukemia through the use of deep learning. MILLIE's high accuracy in identifying APL in bone marrow aspirates and blood films made clinical evaluations easier in environments with limited resources.

Segu Praveena et al. [12] introduced the segmentation and classification of acute lymphoblastic leukemia (ALL) using Deep CNN, Grey Wolf-based Jaya Optimization Algorithm (GreyJOA), and Sparse Fzzy C-Means (Sparse FCM). With its promising sensitivity, specificity, and accuracy, this ALL diagnosis method has the potential to lower patient mortality.

A social spider optimization-based computer-aided diagnosis system for acute lymphoblastic leukemia was proposed by Ahmed T. Sahlol et al. [13]. This innovative model surpasses previous approaches and may help with early ALL diagnosis thanks to its unique integration of multiple features and 95.67% classification accuracy.

Bayesian-optimized CNNs were applied to microscopic blood smear images by Ghada Atteia et al. [14] in order to diagnose ALL. Their model outperformed other cutting-edge techniques with a 96.81% accuracy rate on the test set, demonstrating its enormous potential for ALL detection.

A computer system based on image analysis was created by Ahmed M. Abdeldaim et al. [15] to diagnose ALL. With the K-NN classifier in particular, the system segments and classifies cells as normal or affected with good accuracy. Although statistical results are not provided in the article, the suggested system might aid in the diagnosis of ALL.

The inefficiencies of manually identifying acute lymphoblastic leukemia (ALL) were investigated by Adel Sulaiman et al. [16]. ResRandSVM increases the accuracy of automated diagnosis by utilizing Random Forest for feature selection, ResNet50 for feature extraction, and Support Vector Machine for classifier. Three methods are used to refine the deep features that multiple models extract. The improved features for blood smear leukemia detection are tested by four classifiers. ResRandSVM performs well when using InceptionV3 for feature extraction, Random Forest for feature refinement, and SVM for classification. ResRandSVM performs better in experiments than in other comparisons, indicating that it has the potential to expedite ALL diagnosis.

The recent advancements in Leukemia diagnosis present a range of methodologies, each contributing valuable insights to the domain. Arif Muntasa et al.'s work [6] utilizes the Gray Level Co-occurrence Matrix (GLCM) for feature extraction, whereas other studies, like that of Ghada Emam Atteia et al. [10], delve into deep learning frameworks.

One notable trend from the related research is the focus on robust feature extraction and optimization. Arif Muntasa et al.'s approach [6], while achieving high accuracy rates, extracts 192 features for each object. Such a comprehensive feature space, while detailed, might introduce redundancy, potentially leading to computational inefficiencies and overfitting. This is where our RF-RFE model's strength becomes evident. By streamlining feature sets and removing non-essential attributes, RF-RFE offers an optimized and relevant feature set for diagnosis.

Deep learning models, such as Ghada Atteia et al.'s Bayesian-based CNNs [14] and Petru Manescu et al.'s MILLIE [11], are powerful but often demand significant computational resources. Moreover, their complexity can sometimes challenge interpretability, which is essential in medical applications.

Optimization techniques also find representation in this array of research. Nada M. Sallam et al.'s work [9] introduces the Grey Wolf Optimization algorithm for feature selection. Their method, with its impressive accuracy metrics, illustrates the potential of nature-inspired algorithms. In contrast, our RF-RFE, rooted in Random Forests, offers a method that seeks to understand the inherent structure of the data.

Furthermore, Adel Sulaiman et al.'s ResRandSVM [16] shares similarities with our approach by integrating feature extraction, refinement, and classification. However, our RF-RFE stands apart in its explicit focus on addressing redundancy in high-dimensional data, ensuring the most optimal feature set powers the subsequent XGBoost mechanism.

Considering the diverse methodologies in blood cancer diagnosis, our RF-RFE model emerges as a balanced approach that emphasizes precision, computational efficiency, and clarity, marking its significance in the field.

3. Methods and Materials.

The sequential phase architecture of the RF-RFE framework is depicted in Figure 3.1. Microscopic blood smear features are used by the novel machine learning-based blood cancer prediction model RF-RFE. Preprocessing techniques are prominently featured in this first stage to enhance image clarity and quality. Next, combine the texture, color, and morphology to create a comprehensive feature vector. RF-RFE, designed for high-dimensional data, controls refinement. Through a series of iterations, this process carefully eliminates less significant attributes to produce an ideal feature set for analysis. The gradient-boosting algorithm XGBoost performs well with complex biological data. XGBoost [17] starts off with this precisely calibrated feature set. This machine learning model trains, adapts, and improves its predictive capabilities using gradient-boosted tree algorithms [16]. Throughout the modeling process, performance is monitored to guarantee the best possible outcomes. This system detects leukemia-variant blood cancers early by using machine learning and image analysis.

3.1. Preprocessing.

Image Resizing. Image resizing standardizes the dimensions of all images to a consistent size. This ensures that features extracted from each image are comparable and consistent. Images acquired from different sources or devices can have varying dimensions. Resizing them to a consistent dimension helps in managing the computational cost and ensuring uniformity in feature extraction.

Let I be the input image with dimensions (h, w). Resizing it to dimensions (h', w') is achieved by a spatial transformation function T, such that $I' = T(I)$ where I' is the resized image.

Noise Reduction. Noise reduction involves filtering the image to remove unwanted artifacts and noise, which could distort the image's actual content. Medical images might contain noise due to various reasons like electronic interference, transmission errors, or imperfect sensors. Removing this noise is essential for clear visualization and accurate feature extraction. A common method is Gaussian blurring, represented as: $I' = G * I$ where I' is the denoised image, $*$ denotes convolution, and G is a Gaussian kernel.

Contrast Enhancement. Contrast enhancement amplifies the differences between pixel values in an image, making features more distinguishable. Some medical images might have low contrast due to the nature of the tissue or the acquisition process. Enhancing contrast aids in better visualization and differentiation of regions of interest.

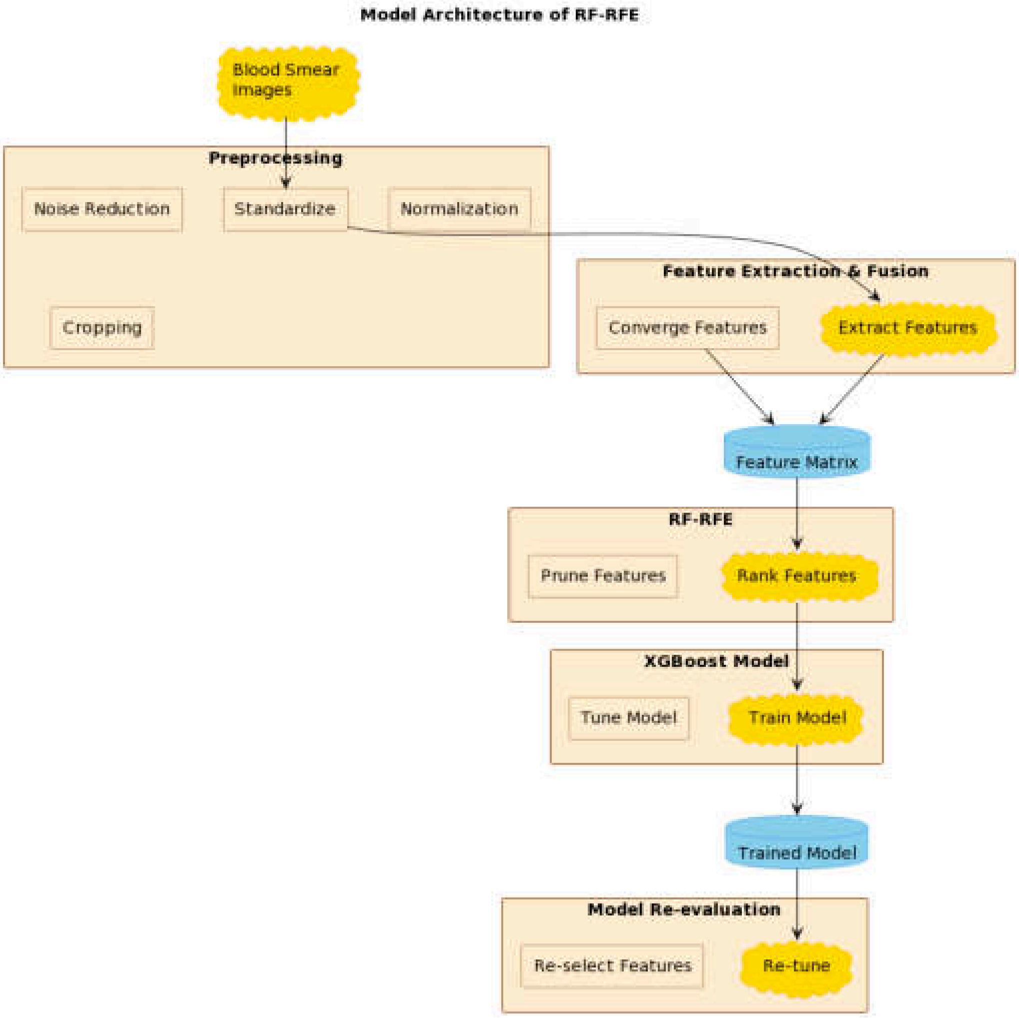


Fig. 3.1: Architecture of RF-RFE

Histogram equalization is one technique:

$$p_r(_k) = n_k/MN \tag{3.1}$$

where p_γ is the normalized histogram, γ_k are pixel intensities, n_k is the number of pixels with intensity γ_k, and $M \times N$ is the image size.

Image Segmentation. Image segmentation partitions an image into multiple segments or regions, often separating objects of interest from the background. In medical imaging, segmenting out regions of interest, like tumors or specific organs, allows for targeted analysis and reduces computational costs.

One method is the Otsu's thresholding:

$$\sigma_\omega^2(t) = \omega_0(t)\omega_1(t)[\mu_0(t) - \mu_1(t)]^2 \tag{3.2}$$

3.2. Feature Engineering. Feature engineering is a cornerstone in machine learning applications for predicting ALL from blood smear images. Fusion of diverse features – including keypoints and key descriptors, morphological attributes, color distributions, texture patterns, spatial relationships, boundary contours, and the

Nucleus to Cytoplasm Ratio – provides a comprehensive representation of the intricate details present in blood smear images. Each of these feature types captures a unique aspect of cellular structures and their potential abnormalities, ensuring the model receives a holistic understanding of the image content. The significance of this fusion lies in its ability to enhance the model's robustness and predictive capability; while some features, like color, might capture staining intensity variations indicative of ALL, others, such as morphological features, can hint at cell structure anomalies. By integrating these diverse features, we harness the collective strength of each feature type, thereby justifying their fusion for an optimal and accurate ALL prediction model.

3.2.1. The Features. Keypoints and Descriptors: These features highlight the salient features of keypoints found in images. Scale [18], orientation [19], location [20], contrast [21], edge response, Harris response, main orientation, descriptor vector [22], Laplacian sign [23], and magnitude [24] are some of them. For locating specific areas of interest in images and understanding their characteristics, these features are essential. While the descriptor vector encodes local appearance, scale, orientation, and location are particularly crucial for spatial information.

Physical characteristics [25]. Morphological features identify characteristics of the object's size and shape. Area, perimeter, compactness, major and minor axis lengths, eccentricity, convexity, area, solidity, extent, orientation, and equivalent diameter are important characteristics. The morphology and structure of objects, such as cells in medical images, can be described using these features. Fundamental size and shape descriptors include area and perimeter, while eccentricity and convexity shed light on any irregularities in an object's shape.

Features of color. Color features concentrate on the color data present in objects. For the red, green, and blue channels, they include mean intensities and standard deviations as well as chroma, hue, value (brightness), color variance, and color entropy. For distinguishing objects based on their color attributes, these features are crucial. The distribution and variation of color can be better understood using mean intensities and standard deviations.

Details of the texture. The spatial arrangement of pixel intensities within objects is described by texture features. They consist of Haralick textures [26], Gabor filters [27], energy (uniformity), entropy [28], homogeneity, correlation, dissimilarity, second moment, and fractal dimension [29]. Understanding the minute details and patterns within objects requires these features, which are essential. For instance, contrast and entropy quantify the complexity and randomness of a texture.

Features of spatial relationships. The arrangement and relationships between the objects in an image are described by spatial relationship features. They include the following metrics: the nearest neighbor distance, pairwise distance statistics (mean and standard deviation), the clustering coefficient, the convex hull area ratio, the object separation index, the object density, the object orientation, the object eccentricity, the object area ratio, and the object perimeter ratio. Analysis of object spatial distributions and clustering patterns benefits greatly from these features.

Features of the boundary and contour. The shape and boundary characteristics of objects are the focus of boundary and contour features. They consist of the following: perimeter, compactness, aspect ratio, circularity, solidity, convexity, bending energy, curvature, and skeletonization. Insights into the object's general shape, roundness, and curvature are provided by these features, which are crucial for identifying object classes.

Ratio of Cytoplasm to Nucleus. The interaction between the nucleus and cytoplasm in cells is quantified by these features. They include the nucleus to cytoplasm area ratio, the nucleus to cytoplasm perimeter, the nucleus to cytoplasm roundness ratio, the nucleus to cytoplasm eccentricity ratio, the nucleus to cytoplasm eccentricity ratio, the nucleus to cytoplasm eccentricity ratio, and the nucleus to cytoplasm eccentricity. The balance between the properties of the nucleus and the cytoplasm, which is a feature of these features, can be a sign of the health of a cell in the context of medical image analysis.

3.2.2. Feature Extraction. Feature extraction from blood smear images is a process tailored to capture intricate cellular details pivotal for diagnostics. It begins with key-points and key-descriptors, identifying unique patterns within the cells that are invariant to image transformations. Morphological features elucidate the shape and structure of cells, highlighting any irregularities. While texture features depict subtle patterns and variations within the cell structures, the color spectrum captures the variation in staining intensities, which can be indicative of pathological changes. Context is provided by spatial relationship features, which show how cells are positioned and distributed in relation to one another. The Nucleus to Cytoplasm Ratio provides information about cellular composition, which is frequently disturbed in conditions like ALL, while boundary and contour features highlight the edges and outline of cells. With their combined ability to provide a multidimensional view of blood cells, these features are crucial for sophisticated diagnostic machine learning models.

Key Ideas and Key Descriptives A well-known method for extracting key details and their descriptors from images that guarantees invariance to scale, rotation, and lighting changes is called SIFT (Scale-Invariant Feature Transform) [30]. These key points can indicate particular unique patterns or anomalies in cells in the context of blood smear images. SIFT is a viable option for thorough blood cell analysis because of its capacity to recognize and describe these unique features, despite possible variations in image capture conditions.

Scale-space Extrema Detection. The first step in SIFT is generating a scale space. This is achieved by convolving the original image $I(x, y)$ with Gaussian functions $G(x, y, \sigma)$ over a range of scales. This can be represented as: $L(x, y, \sigma) = G(x, y, \sigma) * I(x, y)$. Here, '*' denotes the convolution operation.

Key point Localization. After creating the scale space, we search for potential keypoints. These are identified at the maxima and minima of the difference-of-Gaussians (DoG) [31] function. The DoG is formed as: $D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y)$. Here '$k$' is a multiplicative constant.

Orientation Assignment. For rotation invariance, each key point is given an orientation based on the local gradient directions of the image. The gradient magnitude $m(x, y)$ and direction $\theta(x, y)$ at each pixel are given by:

$$m(x,y) = \sqrt{((L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2)} \tag{3.3}$$

$$\Theta(x,y) = arctan((L(x,y+1) - L(x,y-1))/(L(x+1,y) - L(x-1,y)))) \tag{3.4}$$

Keypoint Descriptor. Finally, a descriptor for each keypoint is formed by accumulating gradient magnitudes and orientations in a localized region around the keypoint. This step ensures the descriptor's robustness to changes in appearance, such as lighting or affine transformations.

Morphological Features [32]. Morphological features are crucial in differentiating various cell structures within blood smear images. The method of Watershed segmentation, aided by gradient information, is paramount in delineating these attributes. It effectively discriminates between cells that are adjacent or slightly overlapping, making it apt for precisely defining boundaries. Given the essential nature of cell morphology in diagnostics, this technique's accuracy and versatility make it indispensable.

Gradient Computation. Given an image $I(x, y)$, the gradient magnitude is:

$$G(x,y) = ((I_x(x,y))^2 + (I_y(x,y))^2) \tag{3.5}$$

Distance Transform. For a binary image $B(x, y)$, the distance $D(x, y)$ to the nearest zero pixels is:

$$D(x,y) = min_{(i,j)}((x-i)^2 + (y-j)^2) \tag{3.6}$$

Watershed Segmentation. Using $G(x, y)$, basins are formed meeting at watershed lines, indicating cell boundaries.

Color Features. Color histograms and moments are foundational for extracting color features from blood smear images, capturing variations in staining intensities. These intensities can hint at abnormalities, making the method invaluable for diagnostics. The approach is justified as it's computationally efficient and offers a broad representation of cellular color distribution. Histogram Computation: Given image I, histogram H for a color channel c is:

$$H_c(k) = |(x,y)|I_c(x,y) = k| \tag{3.7}$$

Moments. The n^{th} moment of a histogram H_c is:

$$M_n = {}_{(k=0)}^{255} k^n H_c(k) \tag{3.8}$$

Texture Features. Gray Level Co-occurrence Matrix (GLCM) [33] stands out for texture feature extraction. By analyzing pixel pair frequencies at specific positions, GLCM encapsulates patterns and textures in cells, pivotal for discerning abnormalities. Its efficacy in capturing local variations makes it a justified choice.

GLCM Computation. For an offset $(\Delta x, \Delta y)$, $GLCM_P(i,j)$ is the frequency of pixel pairs with intensities i and j.

Spatial Relationship Features. To understand the spatial positioning and orientation of cells, Delaunay triangulation is optimal. This method creates triangles connecting nearby cells, offering insights into cell distribution and proximity. Given the importance of cell relationships in diagnostic contexts, this approach is vital.

Delaunay Triangulation. Given a set of points P, a triangle (p,q,r) belongs to the Delaunay triangulation if no other point in P lies within the circumcircle of the triangle.

Circumcircle condition. For each triangle ΔABC in T, let O be the center of the circumcircle passing through A, B, and C. The triangulation T is Delaunay if and only if no point P from the set lies inside the circle with O as the center.

Empty Circle Property. For every triangle ΔABC in the Delaunay Triangulation, the circumcircle of ΔABC does not contain any other point of P in its interior.

Boundary and Contour Features. Active Contour Model or Snakes is a potent method for boundary and contour feature extraction. By iteratively evolving curves based on internal and external forces, it clings to cell boundaries, ensuring precise contour delineation. This method's adaptability to subtle boundary nuances justifies its adoption. Snake Evolution: The snake $v(s) = [x(s), y(s)]$ evolves according to:

$$F_{total}() = {}_{0}^{1}F_{int}() + F_{image}() + F_{con}()ds \tag{3.9}$$

Spatial Relationship Features. To understand the spatial positioning and orientation of cells, Delaunay triangulation is optimal. This method creates triangles connecting nearby cells, offering insights into cell distribution and proximity. Given the importance of cell relationships in diagnostic contexts, this approach is vital.

Boundary and Contour Features. Active Contour Model or Snakes is a potent method for boundary and contour feature extraction. By iteratively evolving curves based on internal and external forces, it clings to cell boundaries, ensuring precise contour delineation. This method's adaptability to subtle boundary nuances justifies its adoption.

Nucleus to Cytoplasm Ratio [34]. Thresholding and region-based segmentation are crucial for delineating the nucleus and cytoplasm in cells. By computing their areas separately and determining their ratio, insights into cellular health are gleaned. This feature is critical given its prominence in many pathological conditions, including ALL.

Thresholding: For image I, binary image $B(x,y)$ is:

$$(1ifI(x,y) > T; 0otherwise) \tag{3.10}$$

Ratio Computation: For segmented nucleus area A_N and cytoplasm area A_C:

$$Ratio = AN/AC \tag{3.11}$$

3.3. Optimal Feature Selection. Modern medical image analysis relies heavily on extracting comprehensive features from images to improve the accuracy of disease predictions. When dealing with blood smear images, especially in the context of ALL prediction, a fusion of various features — including keypoints and key descriptors, morphological features, color features, texture features, spatial relationship features, boundary and contour features, and the Nucleus to Cytoplasm Ratio — provides a rich representation of data. Yet, such fusion can also introduce redundancy and noise. Therefore, there's a pressing need for an optimal selection process to retain only the most significant features, ensuring efficient and accurate diagnosis models.

The combination of Random Forest (RF) [35] with Recursive Feature Elimination (RFE) [36] offers a systematic approach to tackle this challenge. RF inherently ranks features based on their importance, providing an aggregated measure of their significance in the classification task. This prioritization becomes critical when handling a diverse set of features, ensuring the model focuses on the most relevant attributes. RFE, on the other hand, is a recursive method that eliminates less important features step by step, thereby refining the feature set.

1. Holistic Data Representation [37]. Given the fusion of diverse features, there's a mix of linear and non-linear data patterns. RF's nature to cater to both ensures that no crucial data pattern is overlooked.

2.Redundancy Reduction [38]. The fusion of multiple feature sets often leads to overlapping information. RF's intrinsic feature ranking, combined with the iterative removal process of RFE, ensures that redundant features are systematically eliminated.

3. Interpretability. Medical diagnostics requires not just accuracy but also the ability to understand the decision-making process. RF offers insight into feature importance, aiding researchers and medical practitioners in discerning the key features driving predictions.

4. Optimal Performance. RF's ensemble nature, utilizing multiple decision trees, ensures a balance between bias and variance, leading to robust and stable predictions. When combined with the refined feature set from RFE, it results in enhanced model performance.

5. Efficiency in Training. By focusing only on the most significant features, the computational burden during model training is reduced, leading to faster and more efficient model training without compromising accuracy.

Optimal feature selection using RF-RFE involves a synergistic approach to refine a fusion of diverse features - keypoints and key descriptors, morphological features, color characteristics, texture patterns, spatial relationship attributes, boundary and contour details, and the Nucleus to Cytoplasm Ratio. This amalgamation offers a comprehensive representation of data, capturing both global and local nuances. RF-RFE stands out in this context due to its inherent ability to rank features based on their ensemble importance. By systematically and recursively eliminating less impactful features, RF-RFE ensures the retention of only the most significant ones, enhancing model performance. This methodology leverages the strengths of Random Forest, such as handling non-linear patterns and feature interactions, to provide a robust and justified selection of optimal features from the intricate fusion.

Filter techniques prioritize features using individual statistical metrics, often overlooking their interactions or their relevance to the target variable. They might consider measures like variance or outcome correlation. While efficient, they can miss intricate relationships in a diverse feature set. RF-RFE, leveraging Random Forest, excels in recognizing inter-feature relationships, giving a richer assessment. With each tree evaluating features across different scenarios, RF-RFE adeptly navigates complex and non-linear relationships, proving more suitable for critical ALL prediction features.

Wrapper techniques, encompassing methods like backward elimination, depend on specific classifiers to assess feature subsets. They take into account feature interplays and can produce classifier-specific feature sets. However, they're resource-heavy, especially for vast feature sets derived from image fusion. RF-RFE smartly amalgamates wrapper and embedded strengths. Random Forest's ranking captures intricate patterns, and RFE's recursive procedure facilitates streamlined, efficient feature pruning. This blend enhances scalability and adaptability, especially beneficial for the nuanced feature set in ALL prediction.

Techniques like LASSO merge feature selection with model training. While efficient and often clear-cut, they may not always grasp complex relationships, especially with a broad fusion of features from blood smear images. RF-RFE, utilizing Random Forest's ensemble strength, delivers a robust feature significance assessment. Paired with RFE's methodical approach, it ensures a thorough yet focused feature exploration, making RF-RFE particularly suitable for the multifaceted feature landscape of ALL prediction.

Theoretical foundation. Random Forest-Recursive Feature Elimination (RF-RFE) combines the robust classification capabilities of Random Forest (RF) [39] with the systematic feature pruning of Recursive Feature Elimination (RFE) [40]. RF builds multiple decision trees on varied data subsets and averages their predictions, offering reduced overfitting and high interpretability. Each tree's construction uses a random subset of features, emphasizing different attributes across trees. RFE, on the other hand, iteratively trains the model, ranks features by their importance, and removes the least significant ones. When fused, RF-RFE leverages RF's feature importance metrics to efficiently and recursively prune irrelevant features, optimizing the model for both performance and interpretability, especially vital in intricate tasks like medical diagnostics.

Random Forest (RF). The strength of RF comes from aggregating (or ”bagging”) the results of numerous decision trees, each trained on a subset of the data. The variability among trees decreases the model’s variance, reducing the likelihood of overfitting.

Let’s denotes:
D: The original dataset.
D_i: A bootstrap ample of D
F: The bull set of features.
F_j: A random subset of features at node split j.

Lemma 1. Every tree in the forest is built on a bootstrap sample (a random sample with replacement) from the original data. This bootstrap sampling introduces variability among the trees:

$$D_i = (x_1^*, y_1^*), (x_2^* y_2^*), ..., (x_n^* y_n^*) \tag{3.12}$$

where each $(x_k^* y_k^*)$ is a random sample with replacement from D.

Lemma 2. At each node split, only a random subset of features is considered, further introducing variability among the trees. This randomness ensures that the trees are uncorrelated, making the averaging process more effective at reducing variance. For each node split j: $F_j \subset F$ where F_j is a randomly selected subset of features from F at that node.

The forest’s final prediction, Y_{RF} for regression can be an average of the individual trees’ predictions, and for classification, it can be a majority vote. If T represents the total number of trees:

$$Y_{RF} = \frac{1}{T} \sum_{i=1}^{T} Y_{\text{tree}_i}$$

where Y_{tree_i} is the prediction of the i^{th} tree.

Recursive Feature Elimination (RFE). RFE is a wrapper-based feature selection algorithm that fits the model multiple times, each time eliminating the least important features.

Let’s denote:
Φ: A function which ranks features based on importance after training the model.
F_k: Set of features retained in the k^{th} iteration.

Lemma 3:. At each iteration, after the model (in this case, RF) is trained, features are ranked based on their importance. The least important features are more likely to add noise than provide value. After training on F_k features:

$$(F_k) = {}_{1}, {}_{2}, ..., {}_{k} \tag{3.13}$$

where${}_1$ is the most important and ${}_k$ is the least important.

Lemma 4. By recursively training the model and eliminating the least important features at each step, the model becomes more focused on the most significant features. This stepwise refinement ensures that the final feature subset is optimal or near-optimal for model performance.

Given a step size after each iteration:

$$F_{(k+1)} = F_k - {}_{k}, {}_{(k-1)}, ..., {}_{(k-+1)} \tag{3.14}$$

That is, the least important δ features from F_k are removed to form F_{k+1}. This recursive process continues until a desired number of features is retained, or until model performance meets a specified criterion.

3.4. RF-RFE Algorithm.

Initialization: Start with the full dataset D and the complete feature set F.
Set T as the number of trees for the RF model.

Set δ as the number of features to remove in each iteration of RFE.
Set $F_{\text{current}} = F$.

Random Forest Training: For $i = 1$ to T
(a) Bootstrap Sampling:
$D_i = \{(x_1^*, y_1^*), (x_2^*, y_2^*), ..., (x_n^*, y_n^*)\}$
Where each (x_k^*, y_k^*) is a random sample with replacement from D.
(b) Construct Tree: For each node split j:
Select a random subset of features:
$F_j \subset F_{\text{current}}$
Split the node using the best feature in F_j based on an impurity criterion (e.g., Gini impurity or entropy).

Feature Importance Evaluation: After training the RF model on F_{current}:
$\Phi(F_{\text{current}}) = \{\phi_1, \phi_2, ..., \phi_m\}$
where ϕ_1 is the most important feature, and ϕ_m is the least important, and m is the size of F_{current}.

Feature Elimination: Remove the least important δ features:
$F_{\text{next}} = F_{\text{current}} - \{\phi_m, \phi_{m-1}, ..., \phi_{m-\delta+1}\}$
Set $F_{\text{current}} = F_{\text{next}}$

Recursive Iteration: Repeat the above 4 steps from initialization to feature elimination until the desired number of features is retained, or another stopping criterion such as model performance on a validation set reaches a threshold is met.

Final Model Training: Train the RF model on the dataset D using the final selected feature subset from F_{current}.
Model Evaluation and Prediction: Evaluate the model's performance using out-of-bag samples. Out-of-bag (OOB) [41] error estimation is a unique property of the bootstrap aggregating (bagging) procedure, which is central to the Random Forest algorithm. When a specific data instance is not used for building a particular tree during bootstrap sampling, it becomes an OOB sample for that tree. Given the nature of bootstrapping, roughly one-third of the data are left out of the bootstrap sample and not used in the construction of the k^{th}tree.

- Let D be the dataset of size N.
- Let T be the number of trees in the random forest.
- For each instance x_i in D, let $\text{Trees}(x_i)$ be the set of trees for which x_i is an OOB sample.
- Let $\text{Pred}_{\text{tree}_j}(x_i)$ be the prediction of the j^{th} tree for the instance x_i:

OOB Prediction for a Data Instance: For each instance x_i, the OOB prediction, $\text{Pred}_{\text{OOB}}(x_i)$, is given by the majority vote (classification) or average (regression) of the predictions of the trees for which x_i is an OOB sample. $\text{Pred}_{\text{OOB}}(x_i) = \text{MajorityVote}(\{\text{Pred}_{\text{tree}_j}(x_i) | \text{tree}_j \in \text{Trees}(x_i)\})$ (For classification)
$\text{Pred}_{\text{OOB}}(x_i) = \frac{1}{|\text{Trees}(x_i)|} \sum_{\text{tree}_j \in \text{Trees}(x_i)} \text{Pred}_{\text{tree}_j}(x_i)$ (For regression)

OOB Error: The OOB error is the proportion of instances that are misclassified (for classification) or the mean squared error (for regression) based on the OOB predictions:

$$Err_O OB = 1/N_{(i=1)}^{N} I[Pred_{OOB}(x_i) y_i] \tag{3.15}$$

(For classification, where is the indicator function, which is 1 if the condition is true and 0 otherwise) $Err_{OOB} = 1/N_{(i=1)}^{N}(Pred_{OOB}(x_i) - y_i)^2$ (For regression)

The Err_{OOB} provides an unbiased estimate of the rest error without the need for cross-validation or a separate rest set, making it highly efficient for model evaluation in bagging-based methods like random forest.

3.5. Model Building with XGBoost. Utilizing XGBoost with features selected by RF-RFE addresses the complexity and high dimensionality inherent in biological data, such as blood smear images for ALL prediction. By combining Random Forest's capacity to discern feature importance with XGBoost's gradient-boosting mechanism, this approach offers an enhanced predictive accuracy and efficiency. The synergy between the ensemble techniques of RF-RFE and XGBoost together ensures robust feature selection, reduced overfitting, and a model fine-tuned for performance, making it particularly vital for the precise and critical domain of medical diagnoses like ALL.

Let D be the dataset of blood smear images.
Let F be the fusion of features extracted from D, where:
$F = \{$keypoints and key descriptors,
morphological features, color characteristics,
texture patterns,
spatial relationship attributes,
boundary and contour details,
Nucleus to Cytoplasm Ratio$\}$

Step 1: Feature Extraction and Fusion

- For each image i in D:
- Extract each feature set f in F
- Create a combined feature vector v_i for image i

Step 2: RF-RFE for Optimal Feature Selection

- Train a Random Forest classifier on D with all features in F.
- Use the feature importance scores provided by the RF classifier to rank the features.
- Initialize a subset S with all features from F.
- While S has more than one feature:
- Remove the least important feature (based on RF's ranking) from S.
- Retrain the RF classifier with the reduced feature set S.
- Update the feature ranking based on the retrained RF classifier.
- The final feature subset S^* is the one that achieves the highest classification performance on a validation set.

Step 3: Model Building with XGBoost

- Let $L(y, \hat{y})$ be the logistic loss function where y is the true label and $\hat{y}$ is the predicted probability.
- Initialize model with: $\hat{y}_i^{(0)} = \frac{1}{2}\ln\left(\frac{\sum_{y_i=1}\omega_i}{\sum_{y_i=0}\omega_i}\right)$ where ω_i is the instance weight.
  For each boosting round $t = 1$ to T:
  Compute the gradient and hessian for each instance i:
  $g_i = \frac{\partial L(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}$
  $h_i = \frac{\partial^2 L(y_i, \hat{y}_i^{(t-1)})}{\partial (\hat{y}_i^{(t-1)})^2}$
- Build a regression tree to predict the gradients using feature from S^*
- For each leaf j of the tree, compute: $\omega_j = -\frac{\sum_{i\in \text{leaf } j} g_i}{\sum_{i\in \text{leaf } j}(h_i+\lambda)}$ where λ is a regularization parameter.
- Update the predictions for each instance i: $\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta\omega_j$ where η is the learning rate and i belongs to leaf j.

3.6. Model Tuning and Re-evaluation. Stagnation isn't acceptable. The model, post its initial training, enters a phase of continuous evaluation. Through regular hyper-parameter tuning and occasional feature re-selection, it ensures its predictions remain sharp and relevant.

Let D_{train} be the training dataset of blood smear images, and D_{al} be the validation dataset.
Let Prepresent hyperparameters for XGBoost, including:

- Learning rate
- Maximum tree depth$D$

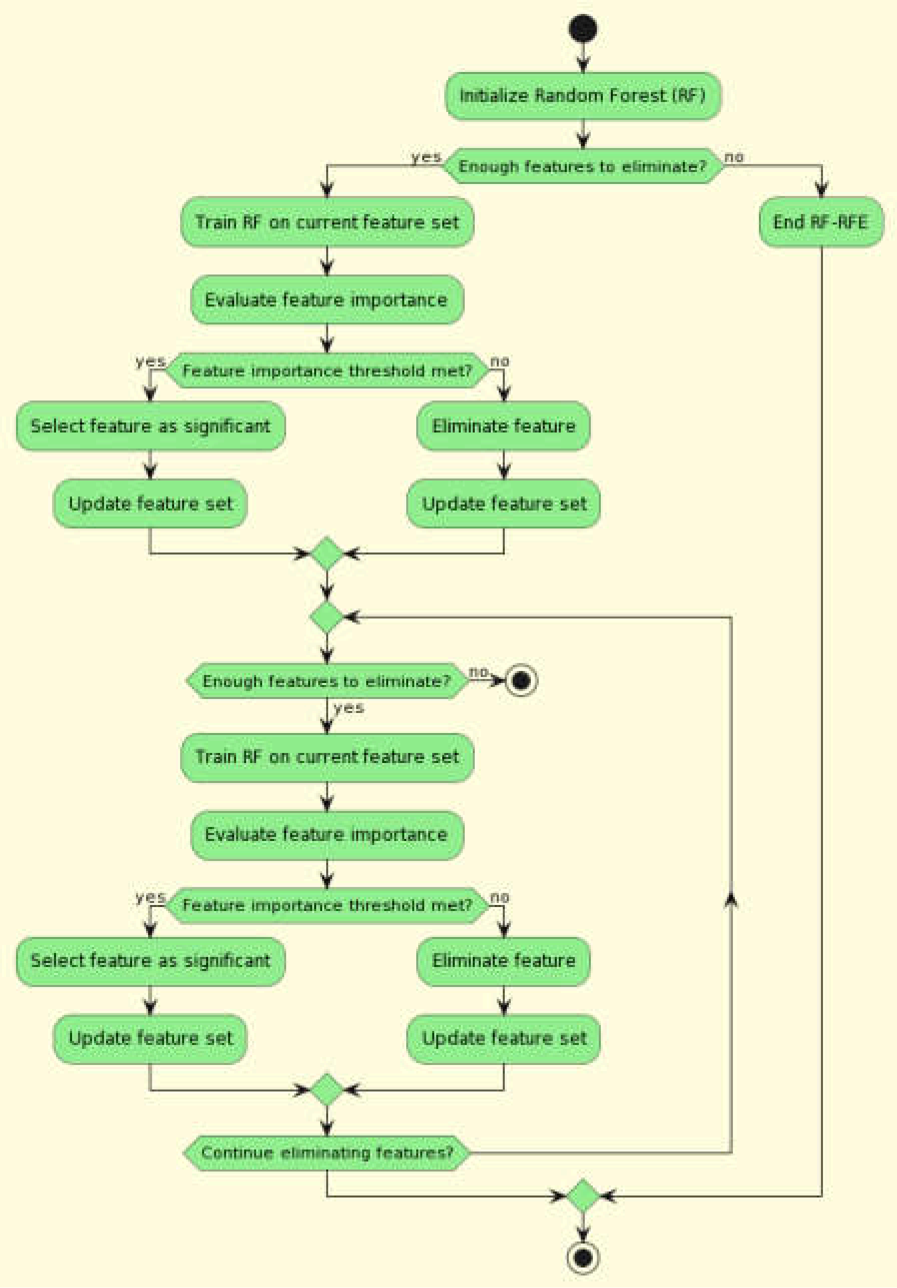

Initialize Random Forest (RF)
yes
Enough features to eliminate?
no
Train RF on current feature set
End RF-RFE
Evaluate feature importance
yes
Feature importance threshold met?
no
Select feature as significant
Eliminate feature
Update feature set
Update feature set
Enough features to eliminate?
no
yes
Train RF on current feature set
Evaluate feature importance
yes
Feature importance threshold met?
no
Select feature as significant
Eliminate feature
Update feature set
Update feature set
Continue eliminating features?

Fig. 3.2: Flow Diagram of the RF-RFE

- Minimum child weight$_{min}$
- Subsample ratio
- Column (feature) sample rate
- Regularization term

Initial Training and Evaluation. Train the XGBoost model on D_{train} using features selected by RF-RFE and initial hyperparameters P_0. Evaluate the model on D_{al} to obtain performance metric M_0.

Hyper-parameter Tuning. For each hyperparameter pin$P$:

Table 4.1: List of Assumptions related to 10 fold cross validation perform

Assumption	**Description**
Data Source and Quality (C_NMC_2019 dataset)	The C_NMC_2019 dataset [43] is assumed to be a reliable and representative source of blood smear images for Acute Lymphoblastic Leukemia (ALL) diagnosis. It is assumed that the dataset has been carefully curated and contains images indicative of various disease stages.
10-Fold Cross-Validation Methodology	The experimental study assumes the use of a 10-fold cross-validation methodology, which involves dividing the dataset into 10 subsets (folds) for training and testing. This methodology ensures robust model evaluation by exposing it to different training-test splits.
Model Comparison (RF-RFE, GWO, RESRANDSVM)	The study assumes the comparison of the RF-RFE model's performance with that of two contemporary models, GWO [9] and RESRANDSVM [16]. It is assumed that these models are suitable benchmarks for evaluating the effectiveness of RF-RFE in leukemia detection.
Performance Metrics	Multiple performance metrics, including precision, recall, specificity, accuracy, f-measure, and ROC (Receiver Operating Characteristic), are assumed to be used for model evaluation. These metrics provide a comprehensive understanding of the models' capabilities.
Goal of Comprehensive Evaluation	The ultimate goal of the experimental study is to assess the true predictive power of the models and their ability to balance false positives and false negatives. This assessment aims to determine the potential real-world applicability of the models in leukemia detection.

1. Adjust p within a predefined range or set.
2. Retrain XGBoost on D_{train} using the updated hyperparameters.
3. Evaluate the model on $D_{al}$to Obtain performance matricM_p .
4. If M_{pis} better (e.g. higher accuracy or AUC, lower loss) than the best metric so far, update P_{best} with the current set of hyperparameters.

Model Re-evaluation with Updated Hyper-parameters. Train the XGBoost model on D_{train}using features selected by RF-RFE and P_{best}.

Evaluate the model on $D_{al}$to confirm performance improvement.

The flow diagram shown in figure 3.2 illustrates the process of Random Forest - Recursive Feature Elimination (RF-RFE), a feature selection technique used in machine learning. The diagram starts with the initialization of the Random Forest. It then enters a loop where it evaluates the importance of features in the current feature set. If a feature meets the importance threshold, it is selected as significant, and the feature set is updated. If not, the feature is eliminated from the set, and the feature set is also updated. This loop continues until there are no more features to eliminate or until a predefined stopping condition is met. Once the loop ends, the RF-RFE process concludes, and the diagram depicts the end of the process. RF-RFE is a systematic approach to select the most relevant features for model training, reducing complexity and improving model performance.

4. Experimental Study. In order to assess the efficacy of the RF-RFE model in the diagnosis of Acute Lymphoblastic Leukemia (ALL) [42] using the C_NMC_2019 dataset [43], an experimental study was carefully planned. The study applied a rigorous 10-fold cross-validation methodology while utilizing the dataset's richness, which offers a wide variety of blood smear images indicative of different stages of the disease. This improved the reliability of the performance evaluation by ensuring that the model was exposed to a variety of training-test splits. The assumptions have been listed in table 4.1 The performance of the RF-RFE model was then compared to that of the contemporary models GWO [9] and RESRANDSVM [16]. Precision, recall, specificity, accuracy, f-measure, and ROC were just a few of the metrics used to provide a thorough understanding of the models' capabilities. The goal of this comprehensive evaluation strategy was to reveal the models' true predictive power as well as their capacity to balance false positives and false negatives, thereby capturing their potential real-world applicability in the crucial field of leukemia detection.

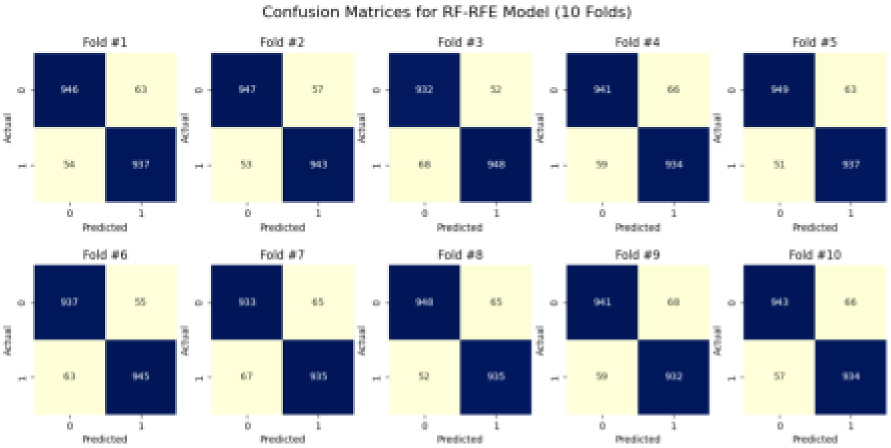


Fig. 4.1: Confusion matrices of 10-fold cross validation performed on proposed model RF-RFE

4.1. The Data. The C_NMC_2019 (Children's Leukemia Data Challenge 2019) dataset is a valuable resource for the development of machine learning models aimed at pediatric leukemia diagnosis, encompassing a total of 12,528 cell images. Among these images, 8,491 represent cases of Acute Lymphoblastic Leukemia (ALL), a critical cancer subtype, while 4,037 images depict normal cell samples. This dataset's significant size and the balanced distribution of cancer and normal cell images make it an ideal choice for robust and comprehensive model training and evaluation in the domain of pediatric leukemia diagnosis.

For a precise diagnosis, the integrity of medical images, particularly blood smear images, is crucial. However, in actual situations, a number of factors may add noise to these images. The C_NMC_2019 dataset intentionally include noise to simulate these imperfect conditions, testing the robustness of the diagnostic algorithms. Intentionally reducing the specificity and sensitivity of the features extracted from the blood smear images allows for a more thorough assessment of the algorithms being tested. A total of 20,000 cell segments have been meticulously selected from the source microscopic images of blood smears, comprising an equal distribution of 10,000 cells from leukemia-infected blood smear images and an additional 10,000 cells from the source images of normal blood smears. This balanced and comprehensive dataset ensures a diverse representation of both pathological and healthy cell samples, offering a robust foundation for subsequent analyses and research endeavors.

4.2. Performance analysis. The RF-RFE model, evaluated through 10-fold cross-validation on a dataset with balanced positives and negatives, consistently demonstrated exceptional performance in leukemia detection. It achieved high true positives and true negatives across all folds, indicating its proficiency in accurately classifying both leukemia-infected and normal cells that shown in figure 4.1. With precision values ranging from 0.9326 to 0.9472 and sensitivity values between 0.932 and 0.949, the model showcased its ability to minimize false positives while effectively identifying positive instances. Furthermore, its specificity remained consistently high, varying from 0.932 to 0.948, ensuring reliable negative classifications. The model's overall accuracy ranged from 0.934 to 0.945, highlighting its capacity for accurate predictions. The F-measure, between 0.9323 and 0.9476, struck a balance between precision and recall, while the false alarming rate remained impressively low at 0.055 to 0.066. With Matthews correlation coefficients ranging from 0.868 to 0.890 and a false positive rate varying from 0.052 to 0.068, the RF-RFE model consistently exhibited robust and reliable leukemia detection capabilities across different folds, making it a promising tool for accurate disease diagnosis. According to the confusion matrices visualized in figure 4.2, the Grey Wolf Optimization (GWO) model demonstrated robust performance across the ten-fold cross-validation, showcasing its effectiveness in distinguishing between leukemia-infected and normal blood smear images. With an average accuracy of 92.35%, GWO exhibited a strong ability to correctly classify instances, supported by high precision (93.95%) and sensitivity

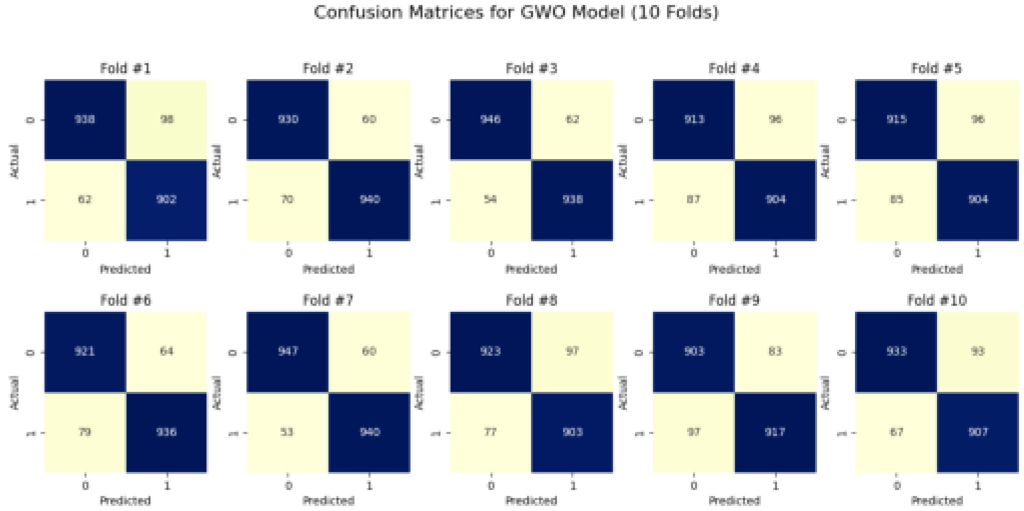


Fig. 4.2: Confusion matrices of 10 fold cross validation performed on contemporary model GWO

(93.05%). The model maintained a well-balanced trade-off between specificity (93.75%) and false positive rates, indicating its competence in avoiding misclassifications. Additionally, the F1-score of 93.82% highlights the model's capability to achieve a harmonious balance between precision and recall. The Matthews Correlation Coefficient (MCC) of 0.8692 further affirmed its performance. Overall, the GWO model exhibited promising potential in the task of leukemia prediction, demonstrating consistent and reliable results across different folds of the dataset. The RESRANDSVM model demonstrates consistent and performance across the 10-fold cross-validation experiments that visualized as confusion matrices of all 10-folds of the cross validation in figure 4.3. It exhibits good precision, specificity, and sensitivity, with values consistently above 0.88, indicating a strong ability to correctly classify both positive and negative cases. The model maintains a high accuracy ranging between 0.88 and 0.91, demonstrating its effectiveness in overall classification. Furthermore, the F-measure, a harmonic mean of precision and sensitivity, consistently exceeds 0.88, indicating a balanced trade-off between precision and recall. The false alarm rate is acceptably low, with values around 0.10, indicating a relatively low rate of misclassification. The Matthews correlation coefficient (MCC) values range between 0.76 and 0.82, signifying a moderate to substantial degree of correlation between predicted and actual classifications. Overall, the RESRANDSVM model showcases a commendable performance in binary classification tasks across various folds, highlighting its reliability and suitability for the given dataset and problem domain.

4.3. Comparative Study. Precision is an imperative metric that gauges the capability of a classification model to identify only the relevant data points accurately. High precision suggests that false positives (incorrectly identified positives) are minimal. According to the results visualized in figure 4.4, RF-RFE emerges as a consistent performer with its precision scores maintaining a tight range around the 0.93 to 0.94 mark across all ten folds. This suggests that its predictions are both accurate and reliable. In contrast, GWO showcases a slightly broader range of fluctuation. Although its precision peaks around 0.940 in a couple of folds, some dips to 0.904 highlight pockets of inconsistency. The RESRANDSVM method exhibits the most variability, with precision scores hovering between 0.87 and 0.90. This indicates a higher propensity to misclassify positive instances compared to the other two methods.

Specificity is a crucial metric, particularly when the cost of false positives is high. It gauges the accuracy of a model in identifying negative outcomes. As visualized in figure 4.4, once again, RF-RFE stands out with its specificity values mirroring its precision scores, ranging mostly between 0.93 and 0.94. Its consistent performance across both metrics emphasizes its balanced and effective classification capabilities. GWO, on the other hand, portrays a pattern akin to its precision values. While it achieves commendable specificity scores upwards of

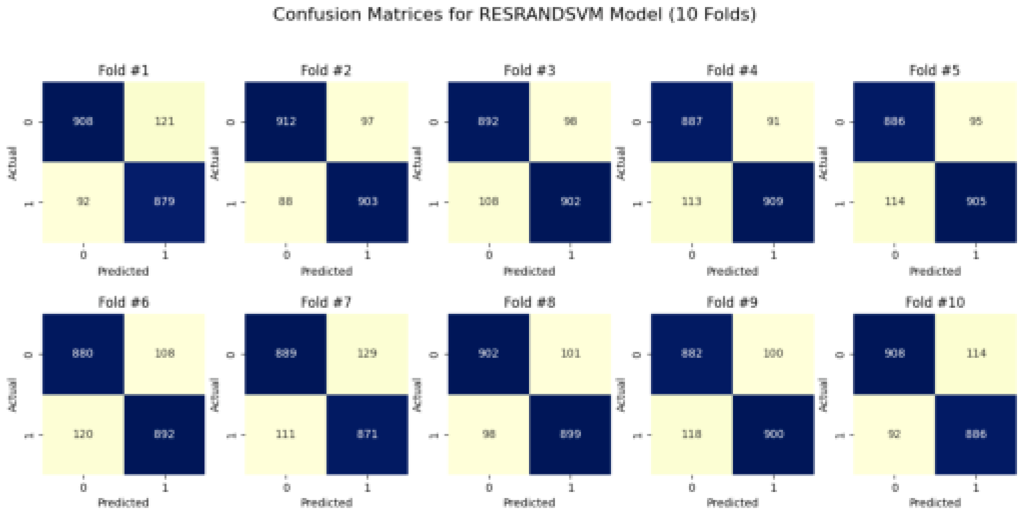


Fig. 4.3: Confusion matrices of 10-fold cross validation performed on contemporary model RESRANDSVM

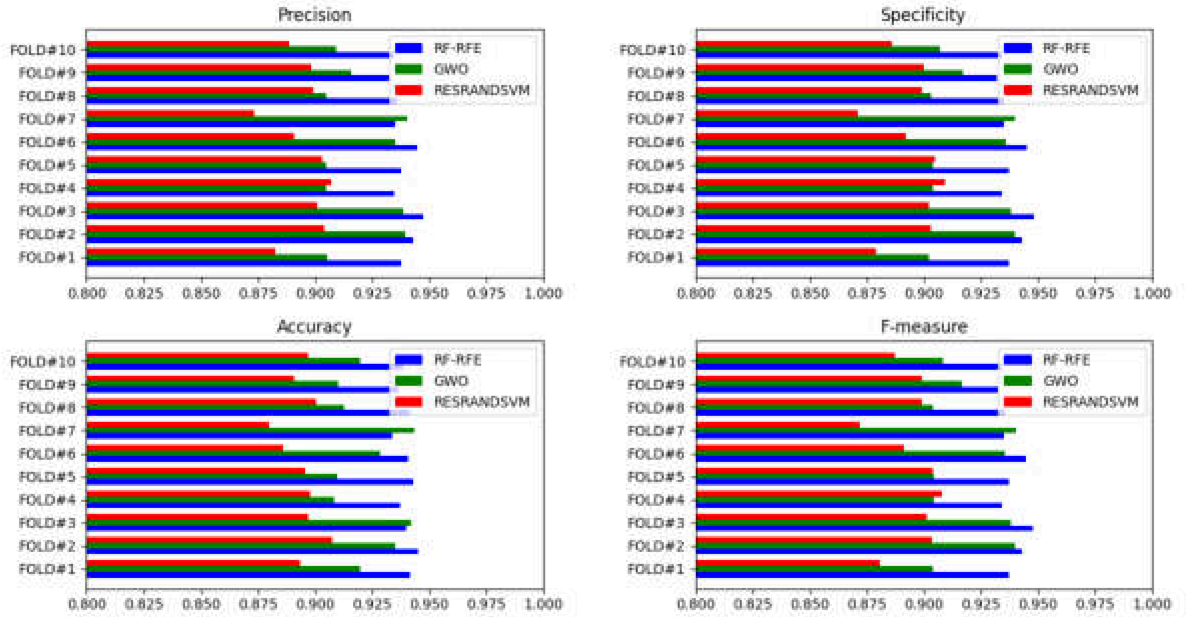


Fig. 4.4: graphs representing the performance metrics precision, specificity, accuracy and f-measure of RF-RFE, GWO, and RESRANDSVM obtained from 10-fold cross validation

0.940 in certain folds, occasional dips towards 0.902 suggest occasional inconsistencies. RESRANDSVM remains the least consistent across the board, with most of its specificity scores settled between 0.87 and 0.90.

Accuracy, perhaps one of the most intuitive performance metrics, offers a comprehensive overview of a model's classification prowess by accounting for both true positives and true negatives. Based in the figure 4.4, RF-RFE consistently achieves the pinnacle of accuracy among the methods, oscillating mainly between

Table 4.2: The presents a Cross Validation of three Methods: RF-RFE, GWO, and RESRANDSVM, across four metrics: Precision, Specificity, Accuracy, and F-measure

Metric	**RF-RFE**	**GWO**	**RESRANDSVM**
Precision	0.9398 ± 0.0048	0.9199 ± 0.0145	0.8948 ± 0.0099
Specificity	0.9390 ± 0.0048	0.9186 ± 0.0146	0.8946 ± 0.0101
Accuracy	0.9401 ± 0.0036	0.9230 ± 0.0133	0.8947 ± 0.0074
F-measure	0.9397 ± 0.0048	0.9195 ± 0.0143	0.8946 ± 0.0096

0.934 and 0.945 across the folds. This affirms its ability to make correct predictions reliably. GWO presents a mixed bag, with accuracy values that diverge notably from one fold to another, spanning from 0.9085 to 0.9435. This variability suggests that its performance might be context-dependent. RESRANDSVM continues its trend of trailing the pack, managing accuracy primarily in the 0.88 to 0.90 range, indicating potential areas for improvement.

The F-measure is a composite metric that strikes a balance between precision and recall, offering a more holistic view of a model's performance, especially when classes are imbalanced. The consistent brilliance of RF-RFE is evident once more from the figure 4.4, as its F-measure scores are closely aligned with its precision, averaging around 0.93 to 0.94. This indicates a harmonious balance between its precision and recall capabilities. GWO displays scores ranging from 0.9037 to 0.9402, reinforcing the narrative of its slightly fluctuating performance. Finally, RESRANDSVM hovers in the lower spectrum with F-measure values mostly between 0.87 and 0.90, reinforcing the notion that it might not be as effective as the other two in the tested scenarios.

Table 4.2 presents a comparative analysis of three methods: RF-RFE, GWO, and RESRANDSVM, across four metrics: Precision, Specificity, Accuracy, and F-measure. Each entry is represented by its average value followed by a deviation. Among the methods, RF-RFE consistently showcases the highest values across all metrics, with Precision at 0.9398 ± 0.0048, Specificity at 0.9390 ± 0.0048, Accuracy at 0.9401 ± 0.0036, and F-measure at 0.9397 ± 0.0048. GWO follows closely, while RESRANDSVM tends to have the lowest values in each category. The deviations also highlight the consistency in the results, with RF-RFE having the smallest variations, indicating its robust performance.

Sensitivity measures the proportion of actual positives that are correctly identified, which is showcased in figure 4.5. RF-RFE shows commendable sensitivity, predominantly fluctuating in the range of 0.932 to 0.949 across the folds. This consistent performance indicates that RF-RFE is adept at identifying true positive cases. On the other hand, GWO exhibits a broader spread ranging from 0.903 to 0.947. While in some folds it manages to rival RF-RFE, in others, it tends to drop notably. RESRANDSVM lingers mostly in the 0.880 to 0.912 bracket, making it the method with the lowest sensitivity on average. It suggests that of the three methods, RESRANDSVM might miss a higher proportion of positive instances.

The false positive rate quantifies the proportion of negatives that are mistakenly classified as positive. Lower FPR values are desirable. As shown in figure 4.5, RF-RFE showcases impressive control over FPR, with values mainly clustered between 0.052 and 0.068. GWO presents a wider range, oscillating between 0.06 and 0.098, signifying a slightly elevated risk of incorrectly classifying negatives. RESRANDSVM consistently registers the highest FPR among the three, with values spanning from 0.091 to 0.129, highlighting its potential vulnerability in misclassifying negative instances.

Similar in essence to FPR, the false alarm rate gauges the frequency of false alarms that presented in figure 4.5. RF-RFE continues its trend of robust performance with values chiefly contained within the 0.055 to 0.066 bracket. This demonstrates its reliability in curbing false alarms. GWO, while respectable in its performance, exhibits a tad more variability, spanning 0.0565 to 0.0915. RESRANDSVM again lags, recording rates from 0.0925 to 0.120, signifying its increased likelihood to raise false alarms compared to the other two techniques.

MCC is a balanced metric that considers all values in the confusion matrix, with 1 indicating perfect prediction, -1 indicating total disagreement, and 0 denoting no better than random prediction. RF-RFE consistently leads in this metric as shown in figure 4.5, with scores ranging from 0.868 to 0.890, underscoring its

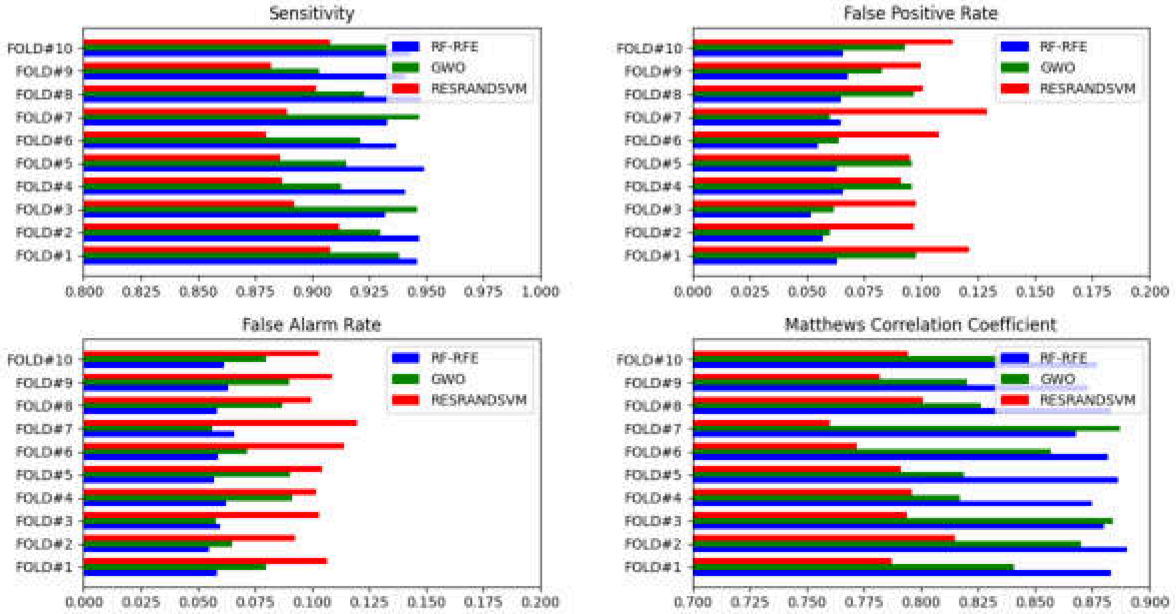


Fig. 4.5: Graphs Representing the Performance Metrics Sensitivity, FPR, FAR, MCC proposed RF-RFE of the compared GWO and RESRANDSVM obtained from 10-fold Cross Validation

Table 4.3: Presents a Cross Validation of three methods: RF-RFE, GWO, and RESRANDSVM, across four metrics: Sensitivity, FPR, FAR, and MCC

Metric	RF-RFE	GWO	RESRANDSVM
Sensitivity	0.9413 ± 0.0061	0.9216 ± 0.0144	0.8961 ± 0.0099
False Positive Rate	0.0615 ± 0.0052	0.0809 ± 0.0151	0.1054 ± 0.0108
False Alarm Rate	0.0602 ± 0.0028	0.0761 ± 0.0133	0.1055 ± 0.0067
Matthews Correlation Coefficient	0.8817 ± 0.0070	0.8461 ± 0.0249	0.7903 ± 0.0158

all-rounded efficacy. GWO follows suit with values mostly between 0.817 and 0.887, suggesting a commendable yet slightly more varied performance. RESRANDSVM, while not too far behind, predominantly hovers in the 0.7601 to 0.815 range. This indicates that, on average, its predictions might be somewhat less correlated with the actual outcomes compared to the other two methods. In the comparative analysis based on the metrics provided in the table 4.3, the RF-RFE method consistently showcased superior performance across all metrics when compared to GWO and RESRANDSVM. Specifically, for Sensitivity, RF-RFE averaged 0.9413, which was higher than GWO's average of 0.9216 and RESRANDSVM's average of 0.8961. Similarly, RF-RFE also exhibited the lowest False Positive Rate and False Alarm Rate among the three methods, indicating a lower likelihood of erroneous classifications. In terms of the Matthews Correlation Coefficient, which measures the quality of binary classifications, RF-RFE again outperformed with an average score of 0.8817. Overall, while all three methods yielded commendable results, RF-RFE stood out as the most effective in this analysis.

4.3.1. Precision-Recall (PR)-Curve. Precision-Recall (PR) curves that presented in figure 4.6, provide an insightful way of examining the performance of classification algorithms, particularly in scenarios where classes are imbalanced. They plot the trade-off between the positive predictive value (precision) and the true positive rate (recall/sensitivity), providing a holistic view of an algorithm's ability to distinguish between classes. In our comparative evaluation of the PR-curves for RF-RFE, GWO, and RESRANDSVM methods, distinct

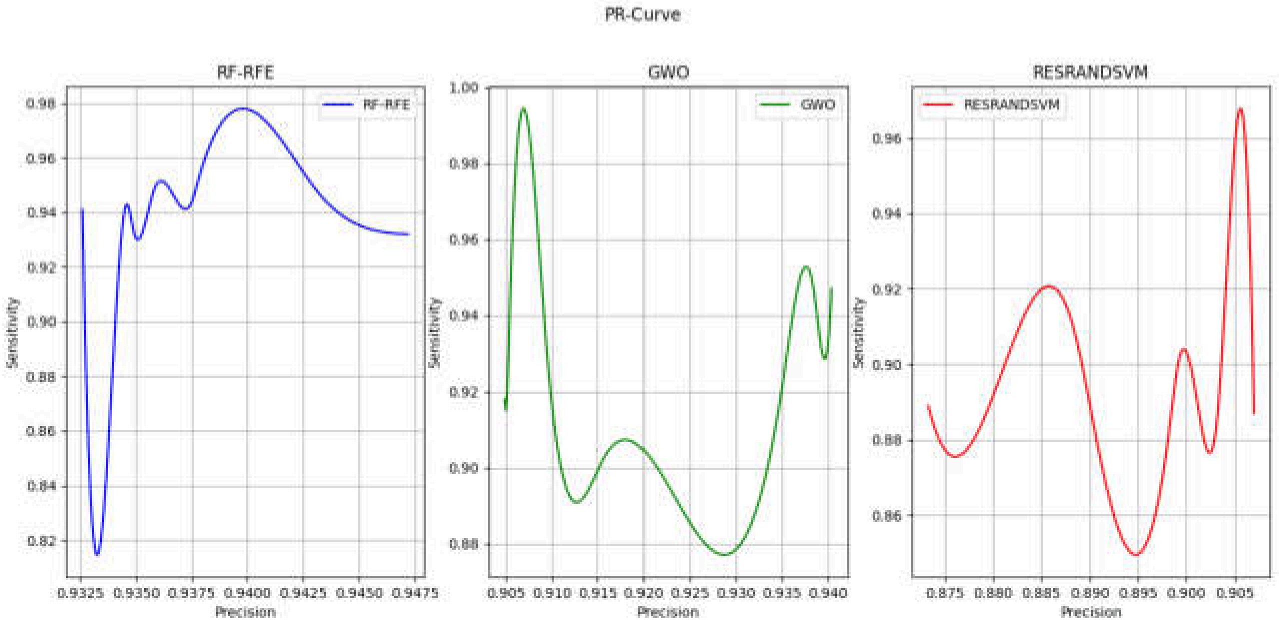


Fig. 4.6: PR-Curve of RF-RFE, GWO, and RESRANDSVM Methods Derived from 10-Fold Cross Validation.

trends are observed. RF-RFE stands out with a consistently superior performance, demonstrated by its steeper ascent in the curve, which implies a robust capability to maintain high precision across diverse sensitivity levels. GWO, though exhibiting some fluctuations suggesting potential variances in precision at different recall intervals, still holds a notable position in the analysis. RESRANDSVM, while displaying a more equilibrated precision-recall trade-off, might not reach the precision peaks of RF-RFE. Analytically, RF-RFE emerges as the top performer in this comparison, suggesting that it's likely to produce fewer false positives for a given recall threshold. However, each method brings its strengths and weaknesses, reinforcing the importance of using PR-curves in understanding the nuances of classifier performance.

4.3.2. Receiver Operating Characteristic (ROC)-Curve. The ROC-Curve that shown in figure 4.7 is a graphical representation of the true positive rate (Sensitivity) against the false positive rate for various threshold values. An ideal method would yield a point in the upper-left corner of the ROC space, representing 100% sensitivity and 0% false positive rate.

From the provided data, RF-RFE demonstrates higher sensitivity across almost all folds compared to the other two methods, especially when false positive rates are low. This means that RF-RFE is potentially better at discriminating between the positive and negative classes. GWO, on the other hand, shows competitive sensitivity values but often at the cost of higher false positive rates. The RESRANDSVM method appears to have the lowest sensitivity values among the three methods in most of the folds, suggesting it might have a lower discriminative ability in this specific context.

5. Conclusion. A sophisticated machine learning and image processing method for blood cancer prediction from blood smear images, the RF-RFE model, was introduced in this study. Blood cancer diagnosis, especially Acute Lymphoblastic Leukemia, has improved with RF-RFE's clarity and precision. Our model greatly improves leukemia detection efficiency and accuracy. Its ability to reduce redundancy in high-dimensional medical imagery data makes RF-RFE unique. The feature set of modern models like GWO [9] and RESRANDSVM [16] is optimised by RF-RFE in detail. Specificity, Accuracy, Precision, and F-measure are a set of performance metrics. It shows the model's dependability. Our extensive analysis showed RF-RFE's precision and low binary classification errors. Further comparisons of Sensitivity, False Positive Rate, False Alarm Rate, and Matthews

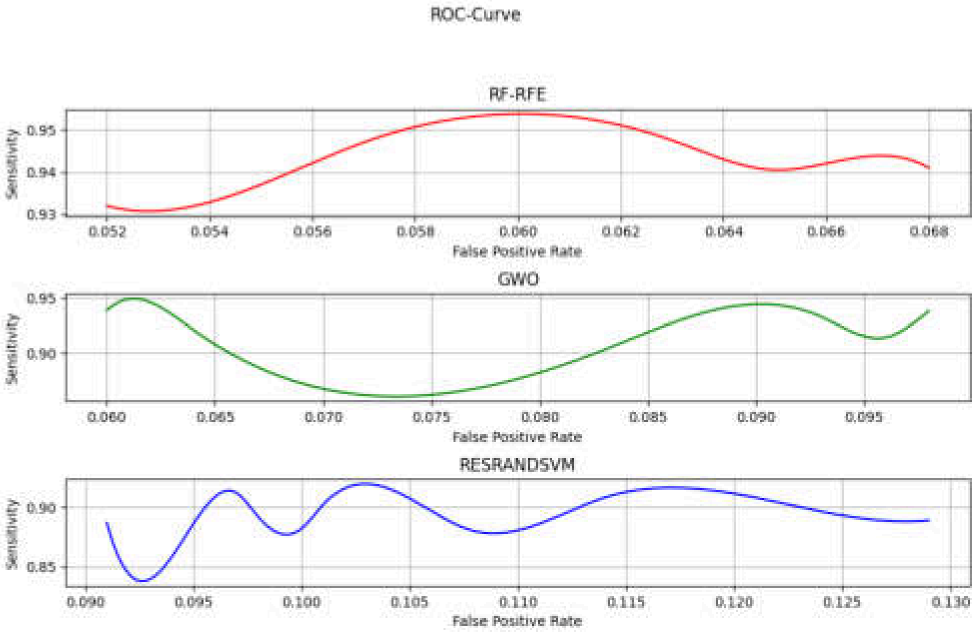


Fig. 4.7: ROC-curve of RF-RFE, GWO, and RESRANDSVM Methods Derived from 10-Fold Cross Validation
.

Correlation Coefficient show the model's precision and ability to balance false positives and negatives. Grey Wolf Optimization, Bayesian-based CNNs [14], and GLCM [6] are impressive, but RF-RFE stands out. Strategically integrating the XGBoost algorithm, RF-RFE sets the standard for ALL diagnosis and biological data analysis. Due to its unique approach and optimized feature set, it is a blood cancer diagnostic breakthrough with consistent performance across diverse datasets. Early blood cancer detection, especially for ALL, is transformed by the model's false positive and negative ability. Many medical diagnostics applications and research are promising with RF-RFE. Adapting RF-RFE for multi-class classification will increase its applicability and help us understand subtype-specific treatment approaches for blood cancer. High-dimensional feature values can be handled without affecting predictive quality with advanced dimensionality reduction. These improvements may enhance RF-RFE's medical diagnostic performance, relevance, and applicability.

REFERENCES

[1] Y. M. Alomari, S. N. H. Sheikh Abdullah, R. Z. Azma, and K. Omar, *Automatic detection and quantification of WBCs and RBCs using iterative structured circle detection algorithm*, Computational and mathematical methods in medicine, 2014.

[2] *Blood Smear Test*, Available at: https://www.testing.com/tests/blood-smear/.

[3] S. Fathima, P. Meenatchi, and A. Purushothaman, *Comparison of manual versus automated data collection method for haematological parameters*, Biomedical Journal of Scientific & Technical Research, 15, no. 3 (2019), pp. 11372-11376.

[4] W. Jiao, X. Hao, and C. Qin, *The image classification method with CNN-XGBoost model based on adaptive particle swarm optimization*, Information, 12, no. 4 (2021), p. 156.

[5] X. Ren, H. Guo, S. Li, S. Wang, and J. Li, *A novel image classification method with CNN-XGBoost model*, In: Digital Forensics and Watermarking: 16th International Workshop, IWDW 2017, Magdeburg, Germany, August 23-25, Proceedings 16, Springer International Publishing, 2017, pp. 378-390.

[6] A. Muntasa, and M. Yusuf, *Multi Distance and Angle Models of the Gray Level Co-occurrence Matrix (GLCM) to Extract the Acute Lymphoblastic Leukemia (ALL) Images*, International Journal of Intelligent Engineering & Systems, 14, no. 6 (2021).

[7] *Classification of Acute Lymphoblastic Leukemia on White Blood Cell Microscopy Images Based on Instance Segmentation Using Mask R-CNN.*

[8] Z. Moshavash, H. Danyali, and M. S. Helfroush, *An automatic and robust decision support system for accurate acute leukemia diagnosis from blood microscopic images*, Journal of digital imaging, 31 (2018), pp. 702-717.

[9] N. M. Sallam, A. I. Saleh, H. A. Ali, and M. M. Abdelsalam, *An efficient strategy for blood diseases detection based on grey wolf optimization as feature selection and machine learning techniques*, Applied Sciences, 12, no. 21 (2022), p. 10760.

[10] G. E. Atteia, *Latent Space Representational Learning of Deep Features for Acute Lymphoblastic Leukemia Diagnosis*, Computer Systems Science & Engineering, 45, no. 1 (2023).

[11] P. Manescu, et al., *Detection of acute promyelocytic leukemia in peripheral blood and bone marrow with annotation-free deep learning*, Scientific Reports, 13, no. 1 (2023), p. 2562.

[12] S. Praveena, and S. P. Singh, *Sparse-FCM and Deep Convolutional Neural Network for the segmentation and classification of acute lymphoblastic leukaemia*, Biomedical Engineering/Biomedizinische Technik, 65, no. 6 (2020), pp. 759-773.

[13] A. T. Sahlol, A. M. Abdeldaim, and A. E. Hassanien, *Automatic acute lymphoblastic leukemia classification model using social spider optimization algorithm*, Soft Computing, 23 (2019), pp. 6345-6360.

[14] G. Atteia, et al., *Bo-allcnn: Bayesian-based optimized cnn for acute lymphoblastic leukemia detection in microscopic blood smear images*, Sensors, 22, no. 15 (2022), p. 5520.

[15] A. M. Abdeldaim, A. T. Sahlol, M. Elhoseny, and A. E. Hassanien, *Computer-aided acute lymphoblastic leukemia diagnosis system based on image analysis*, Advances in Soft Computing and Machine Learning in Image Processing, 2018, pp. 131-147.

[16] A. Sulaiman, et al., *ResRandSVM: Hybrid Approach for Acute Lymphocytic Leukemia Classification in Blood Smear Images*, Diagnostics, 13, no. 12 (2023), p. 2121.

[17] T. Chen, et al., *Xgboost: extreme gradient boosting*, R package version 0.4-2, 1, no. 4 (2015), pp. 1-4.

[18] A. Herod, *Scale*, Routledge, 2010.

[19] B. K. P. Horn, *Relative orientation*, International Journal of Computer Vision, 4, no. 1 (1990), pp. 59-78.

[20] J. J. Gabszewicz, and J.-F. Thisse, *Location*, Handbook of game theory with economic applications, 1 (1992), pp. 281-304.

[21] C. Owsley, *Contrast sensitivity*, Ophthalmology Clinics of North America, 16, no. 2 (2003), pp. 171-177.

[22] J. Thewlis, S. Albanie, H. Bilen, and A. Vedaldi, *Unsupervised learning of landmarks by descriptor vector exchange*, In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6361-6371.

[23] Y. Hou, J. Li, and Y. Pan, *On the Laplacian eigenvalues of signed graphs*, Linear and Multilinear Algebra, 51, no. 1 (2003), pp. 21-30.

[24] R. E. Kirk, *The importance of effect magnitude*, Handbook of research methods in experimental psychology, 2003, pp. 83-105.

[25] P. H. Pelham, and J. G. Dickson, *Physical characteristics*, The wild turkey: biology and management, Stackpole Books, Mechanicsburg, Pennsylvania, USA, 1992, pp. 32-45.

[26] E. Miyamoto, and T. Merryman, *Fast calculation of Haralick texture features*, Human computer interaction institute, Carnegie Mellon University, Pittsburgh, USA, Japanese restaurant office, 2005.

[27] J. R. Movellan, *Tutorial on Gabor filters*, Open source document, 40 (2002), pp. 1-23.

[28] B. Bein, *Entropy*, Best Practice & Research Clinical Anaesthesiology, 20, no. 1 (2006), pp. 101-109.

[29] J. Theiler, *Estimating fractal dimension*, JOSA A, 7, no. 6 (1990), pp. 1055-1073.

[30] T. Lindeberg, *Scale invariant feature transform*, 2012, p. 10491.

[31] A. Bundy and L. Wallen, *Difference of gaussians*, Catalogue of Artificial Intelligence Tools, (1984), p. 30.

[32] U. Ziegler and P. Groscurth, *Morphological features of cell death*, Physiology, 19, no. 3 (2004), pp. 124-128.

[33] S. V. Bino, A. Unnikrishnan, and K. Balakrishnan, *Gray level co-occurrence matrices: generalisation and some new features*, arXiv preprint arXiv:1205.4831, (2012).

[34] J. A. Sebastian, M. J. Moore, E. S. L. Berndl, and M. C. Kolios, *An image-based flow cytometric approach to the assessment of the nucleus-to-cytoplasm ratio*, PLoS One, 16, no. 6 (2021), e0253439.

[35] S. J. Rigatti, *Random forest*, Journal of Insurance Medicine, 47, no. 1 (2017), pp. 31-39.

[36] X.-w. Chen and J. C. Jeong, *Enhanced recursive feature elimination*, In Sixth international conference on machine learning and applications (ICMLA 2007), IEEE, 2007, pp. 429-435.

[37] J. Kim, H. Lee, M. Imani, and Y. Kim, *Efficient Hyperdimensional Learning with Trainable, Quantizable, and Holistic Data Representation*, In 2023 Design, Automation & Test in Europe Conference & Exhibition (DATE), IEEE, 2023, pp. 1-6.

[38] H. Barlow, *Redundancy reduction revisited*, Network: computation in neural systems, 12, no. 3 (2001), p. 241.

[39] L. Breiman, *Out-of-bag estimation*, (1996).

[40] M. Onciu, *Acute lymphoblastic leukemia*, Hematology/oncology clinics of North America, 23, no. 4 (2009), pp. 655-674.

[41] *The Cancer Imaging Archive*, Available at: https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=52758223.

Edited by: S. B. Goyal
Special issue on: Soft Computing and Artificial Intelligence for wire/wireless Human-Machine Interface
Received: Oct 13, 2023
Accepted: Jan 23, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 2001–2012, DOI 10.12694/scpe.v25i3.2739

MASTER DATA QUALITY MANAGEMENT FRAMEWORK: CONTENT VALIDITY

AZIRA IBRAHIM[*], IBRAHIM MOHAMED[†], AND MOHAMMAD KAMRUL HASAN[‡]

Abstract. Organizations rely on high quality master data as a critical component in achieving their operational and strategic performance. To accomplish high quality master data, they need to be managed properly through a systematic and holistic framework. However, prevalent master data quality management frameworks lack in providing comprehensive management practice in assuring the quality of master data. Hence, stimulates the need to develop an improved master data quality framework. Prior to the development of the framework, the identification and validation of factors that contribute to the management of master data quality must be performed. Thus, this paper underlined four elements and seven factors affecting master data quality management. Further, the identified factors were validated using a questionnaire as the validation instrument. The questionnaire consists of 95 items representing the identified seven factors that were derived from previous studies in the domain of total quality management, data quality management, and master data. Since the items are derived from the different contexts of study, content validation is a need. Previous research has suggested several techniques for performing content validation, covering both quantitative and qualitative approaches. The quantitative approach employed objective assessment and the result was statistically analysed. While the qualitative approach adopted subjective assessment such as comments, ideas, or respond. In this paper, the quantitative approach is selected over the qualitative approach, considering the effort in analyzing several items (95 items) is less complex compared to the qualitative approach which is more difficult to interpret and account for biased results. The selected panel of experts validate the instrument using a three-point scale namely "1 = not relevant", "2 = important (but not essential)", and "3 = essential". Later, using the technique proposed by Lawshe, the value of the content validity ratio (CVR) is calculated. As a result, 92 items are accepted, and 3 items are rejected. The elimination of the 3 items is due to the unsuitableness to be used in the context of the public sector. The validated items can be used as an instrument to validate the factors affecting master data quality management. The proposed factor would support the organization in managing master data quality more effectively.

Key words: content validity; total quality management; data quality management, master data.

1. Introduction. Master data represents the company's core business entities which is the main component in executing business processes, reporting, and decision making [13, 21, 33]. In the setting of the public sector, data about customers, services, products, and service providers are categorized as master data [13].

Master data retain worthy information about an organization, accounting for the preeminence to be managed [31, 14]. The consequence of master data on the organization's performance either operationally or strategically is highly evidenced. Hence, managing master data to assuring and maintaining its quality must be the focal point for the organization [33, 14].

The quality of master data is managed by a master data quality management framework covering management tasks on strategic, governance, and technical aspect [35, 32, 40, 42]. Most of these frameworks highlighted the most imperative objective in master data quality management is the achievement of high quality data [6].

However, previous researchers emphasized the need to improve the existing framework since data quality issues are context specifics where one size does not fit all[48]. Furthermore, correlative with the digital transformation, the roles of data have evolved from just fulfilling the business process requirement and achieving the strategic objective of the organization to the strategic and valuable resource for the organization, hence creating the need to continuously update the current model or framework [35, 24].

As highlighted by [6], the fundamental reason why different frameworks or models are needed is due to the evolution in technology and the consequent growing complexity in data quality. As debated by [39], most

[*] Centre for Software Technology and Management, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia

[†] Centre for Software Technology and Management, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia Nurhizam Safie Mohd Satar[§]

[‡] Centre for Cyber Security, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia

Table 1.1: Factors for managing data quality from the total quality managament (TQM) appraoch

Factor	**Description**	**Reference**
Leadership	The top management role in driving and strengthening the management of master data quality including establishing an effective master data quality governance	[14, 8, 29, 3, 26, 18, 15, 34, 41, 45, 50]
Strategic planning	The development and implementation of a strategy to manage master data quality to achieve the outlined vision, missions, and goals	[14, 8, 3, 26, 34, 41, 45, 50]
Customer focus	The prioritization of customers' master data quality requirements and ensuring master data quality fulfills customer's requirement	[8, 3, 18, 34]
Human resource focus	The provision of an adequate and capable workforce, a conducive working environment, a culture that promotes workforce engagement, and a training and development program	[8, 3, 26, 18, 34]
Operation focus	The design, production, and quality control of master data, data supplier management, safety and security of data operational environment, and innovation management	[32, 3, 26, 18, 15, 34, 41, 45, 50]
Master data quality	The product data quality in terms of conformance to specification, is usable in a specific context, concisely represented, and available and accessible	[14, 40, 46, 50, 4]
Result	The effect of a high quality master on an organization's strategic and operational performance	[8, 3, 18, 15, 34, 41]

data quality frameworks are provisional, intuitive, and fragmentary, and consequently produced measurement models that are not robust and systematic. Significantly, current master data quality management frameworks do not accentuate the requirement to adequately manage the quality of master data such as: (i) overlooking the effect of master data quality on an organization's performance [40, 42], (ii) incompletely defining master data quality dimension [35, 32, 40, 42], (iii) partially underline a holistic and systematic management practice in managing master data quality [42], and (iv) most importantly developed and validated in context other than public sector [35, 32, 40, 42].

Thus, to deal with the insufficiency of the present frameworks and models, the total quality management (TQM) concept is proposed as a pillar to determine the factors that affect the management of master data quality.

In accordance, this study proposed a Master Data Quality (MDQM) Framework that comprises elements and factors that affect master data quality management in the context of the public sector. The MDQM Framework is developed from the perspective of TQM by adapting two influential models which are Malcolm Baldrige National Quality Awards Model (MNBQA Model) [8], and Wang and Strong Model [46]. These two models were chosen to highlight the synergy of leadership, strategic planning, customer-oriented, human resource management, and operation in guaranteeing the high quality of master data, further supporting the organization in achieving its strategic and operational performance.

Moreover, the selected models are sufficiently able to accommodate the need to establish a more methodic framework emphasizing the systematic master data quality management practice in the organization in line with the important roles of high-quality master data in digital transformation specifically in supporting organizations to make informed decisions, increase operation efficiency, and improve service delivery. The proposed framework ensures the effective master data management practice in place by coherently integrating all aspects of the management practice revolving from top management commitment, data governance, strategic planning, capability and capacity of human resources and also process effectiveness to ensure the availability of high quality master data.

The description and reference for factors affecting master data quality management are explained in Table 1.1.

Further, to ensure the applicability of the above-proposed factors in the context of the study, a validation process needs to be performed. Later, the validated factors can be synergized to establish a master data

Table 1.2: Description of items for the quastionnaire

Factor	**Item Code**	**Description**
Leadership	1.A1-1 : 1.A1-6 2.A1-7 : 2.A1-12	• Top management commitment • Top management knowledge and experience • Top management communication skills • Establish a master data quality policy • Define master data quality roles and responsibilities • Involve technical and business people at all levels.
Strategic planning	4.A2-1 : 4.A2-5 5.A2-6 : 5.A2-11	• Establish a strategy development process • Involve customer in strategy development • Implement innovation • Analyse strengths, weaknesses, opportunities, and threats • Align strategic objectives with the vision, mission, and master data quality management • Establish a short-term and long-term action plan • Implement action plan • Allocation adequate resources • Develop human resource planning • Implement change management • Assess and monitor action plan performance
Customer focus	7.A3-1 : 7.A3-5 8.A3-6 : 8.A3-11	• Interact with customers using various medium • Provide customer support services • Identify and categorize customer • Build and manage customer relationship • Evaluation complaint and suggestion • Identify master data quality requirements • Assess the quality of master data • Analyse the quality level of master data • Appraise and update periodically master data quality requirements
Human resource focus	10.B1-1 : 10.B1-5 11.B1-6 : 11.B1-10	• Develop a human resource management plan • Define the skills and expertise needed for each task • Appoint personnel with the right knowledge, experience, and qualification • Provide a conducive working environment • Provide appropriate benefits and policies for the workforce • Establish the right organizational culture • Manage human resource performance • Develop a training and development plan • Provide technical and business training related to data quality • Assess periodically the effectiveness of the training and development plan
Operation focus	13.B2-1 : 13.B2-7 14.B2-8 : 14.B2-11	• Establish a systematic methodology for designing and production of master data • Design and produce master data based on business and technical requirements. • Define and document data flow • Establish a systematic methodology for designing and production of master data • Design and produce master data based on business and technical requirements. • Define and document data flow • Integrate master data from multiple sources • Control quality of master data throughout the life cycle • Monitor master data quality and production process • Improve continuously the quality of master data quality and production process • Control master data production cost • Manage data supplier • Provide a secure operating environment • Manage innovation for master data and production process
Master data quality	16.C1-1 : 16.C1-4 17.C1-5 : 17.C1-9 18.C1-10 : 18.C1-13 19.C1-14 : 19.C1-15	• Master data comply with the specification • Master data can be used to perform a specific task • Master data concisely represented • Master data available, and accessible
Result	21.D1-1 : 21.D1-8 22.D1-9 : 22.D1-18	• Master data quality impact on the strategic performance of the organization • Master data quality impact on the operational performance of the organization

quality management framework. The validation of the factor is performed using an instrument that consists of a developed questionnaire. The questionnaire's items are obtained from a comprehensive study of previous research in the domain of TQM, data quality management, and master data. Then, the identified items were reworded to befitting the study context. The developed questionnaire consists of items that represent all the factors that affect the master data quality management. The description of the items for the questionnaire is elucidated in Table 1.2.

Table 1.3: Summary of iteams for the questionnaire

Factor	**Element**	**No. of Item**	**Item Code**
A:Leadership	A1: Leadership	13	1.A1-1 : 2.A1-12 3. General item
	A2: Strategic planning	12	4.A2-1 : 5.A2-11 6. General item
	A3: Customer focus	12	7.A3-1 : 8.A3-11 9. General item
B: System	B1: Human resource focus	11	10.B1-1 : 11.B1-10 12. General item
	B2: Operation focus	12	13.B2-1 : 14.B2-11 15. General item
C: Master Data Quality	C1: Master data quality	16	16.C1-1 : 19.C1-15 20. General item
D: Result	D1: Result	19	21.D1-1 : 22.D1-18 23. General item
Total item		95	

Overall, the questionnaire consists of 95 items, representing 4 elements and 7 factors that form the master data quality management framework. The summary of the items for the questionnaire is explained in Table 1.3

Based on the table above, the questionnaire consists of 95 items covering 88 individual items and 7 general items. Referring to [5, 1], one general item is needed to enable the experts to validate the importance of each factor. For that reason, for each factor, one general item is added to better validate the MDQM framework. All the items are quantitatively evaluated by applying a three-point scale: "1 = not relevant", "2 = important (but not essential)", and "3 = essential". In addition to that, one column section is also provided next to each item as a provision for the experts to state any comments.

Taking into account that the establishment of an instrument grounded on the theoretical theory and items are generated based on the existing instrument, later reworded based on the researcher's comprehension of the theory concept and the study context, the validity of the items is uncertain. Hence, content validity must be performed after the items have been developed [37]. The validation of the questionnaire is needed to assure it is relevant and suitable for representing the research concept [38]. An instrument that is valid is necessary to measure what it is supposed to measure [11]. According to [43], content validity is described as "the degree to which items in an instrument reflect the content universe to which the instrument is to be generalized". Instruments that lack content validity will negatively impact the final result of the study [30]. Content validity could be performed using a few methods as detailed in the following sub-sections.

1.1. Intensive Literature Review. Contents are validated by relying only on an intensive literature review [7, 47]. Most of the items are derived from a comprehensive literature review and existing instruments and do not involve any expert assessment [7, 47]. This method relies solely on the researcher's subjective judgment.

1.2. Intensive Literature Review. The experts' engagement is important to achieve content validity [23, 27]. Experts are individuals that have experience with the capability to communicate their opinion on the subject [5]. Expert assessment can be performed using a qualitative or quantitative approach [19]. Through a qualitative approach, no statistical calculation is involved and purely depends on the subjective review of the item by the selected experts [5]. However, through a quantitative approach, experts will validate the contents in terms of the degree to of each item is relevant and suitable to the construct, and involve statistical calculation and analysis which informs either the item should be retained or rejected [5]. The quantitative approach can be performed via a few techniques as detailed below.

- Content Validity Ratio (CVR) by [23]: Experts will evaluate the degree of relevancy and suitability of each item on a three-point scale: "1 = not relevant", "2 = important (but not essential)", and "3 = essential". CVR value was calculated for each item by applying a formula developed by [23]. Then, later items are removed or retained based on their rating.
- Content Validity Index (CVI) by[27]: Experts will rate the degree of relevancy and suitability of each item on a four-point scale: "1= irrelevant", "2 = somewhat relevant", "3 = quite relevant", and "4 = highly relevant".

Based on[5], content validity assessment via subjective judgment either through intensive literature review or expert assessment method can result in biased outcomes due to its unstructured nature and the process involved may be difficult to reproduce. Furthermore, the outcome of the qualitative approach is usually hard to explicate due to the numerous items in the questionnaire [5].

Hence, [5, 28] proposed that quantitative judgment has a better outlook due to the more systematic process and relies on statistical calculation rather than subjective judgment. [5] also highlighted that the quantitative method suggested by [23] is better than the method suggested by [27]. This is due to the reason that Lawshe's technique does not involve too many panels of experts, and provides a clear and easily understood table in deciding either to accept or reject the item. Furthermore, the calculated CVR value using Lawshe's technique is pragmatic and can be performed in a reasonable time frame, especially during the evaluation process [44]. Additionally, only small number of experts are required to implement Lawshe's technique.

In contrast, this study did not apply the CVI technique since a four-point scale is not common and could be increased by coincidence [44]. Furthermore,[5] highlighted that CVI is not appropriate for a small number of experts and could produce inconsistency due to the uses of normal distribution. Thereby, the quantitative approach as introduced by [23] was chosen to validate the content due to its practicality.

Thus, the content validity process involved four steps which are the selection of a panel of experts, invitation and distribution of the instruments to the appointed panel of experts, calculation and analysis of CVR value and lastly finalising the instruments. The selected panel of experts validated the instrument using a three-point scale namely "1 = not relevant", "2 = important (but not essential)", and "3 = essential" as suggested by [23]. Based on the response, the CVR value is calculated and analysed accordingly. Later, the instrument is revised based on the analysis. The validated instrument mainly contributes to improving master data quality in the public sector domain by the systematization of more rigorous management practices.

The remainder of this article has been organized as follows: Section II explains the materials and method, Section III details the result and discussion, and Section IV describes the conclusion.

2. The material and method. The process for conducting content validity was adapted from [5, 1, 2] which consist of four steps as explained below.

2.1. Step 1- Selection of a Panel of Experts. The chosen panel of experts should possess adequate technical knowledge and experience in the domain of study, be inclined to participate, be able to spend reasonable time, and have satisfying communication skills [9, 36]. Apart from that,[5, 9, 36, 17, 16, 20, 22] emphasized that the panel of experts should also consist of individuals from an academic and practical field that have expertise in the domain of study and also instrument development. Hence, in this study, the panel of experts should present the characteristics below:

- pose knowledge and experience in quality management and/or, data quality management, and/or master data, in either academic or industry area and/or
- have a publication in quality management and/or, data quality management, and/or master data, in either an academic or industry area.

In deciding the number of experts that should be involved, the suggestions by previous researchers vary. The number of appointed experts usually depends on the scope of the research, the availability of the resources available, and the objective of the research [10]. According to [36], at least three panels of experts should be selected. Other than that, [12] suggested that the total number of individuals should be in the range of two to 20. [14] suggesting 11 experts in the field of academics, industry, and statistics. [1] proposed 8 experts for content validation.

Nevertheless, there is no specific procedure for determining the total number of experts that need to be involved in the process of content validation [49]. Nonetheless, the number of a selected panel of experts should consider the criteria for agreeing or denying the items as regards the number of experts as proposed by [23].

2.2. Step 2- Issuance of Invitations and Distribution of Instruments. The invitation was done in two stages. The first stage involves informal communication to get an early agreement. Once the agreement was received, an official invitation was done through e-mail which include detailed instructions on performing content validation.

Table 2.1: CVR value interpretation

CVR Value	**Interpretation**
1.00	all experts answer “3=essential”, which indicates 100 percent agreement, and the item is valid
0 - 0.99	more than 50 percent, but less than 100 percent of the total number of experts responded “3=essential”, which indicates a positive value
<0 (negative value)	less than 50 percent of the total number of experts responded “3=essential”, which indicates a negative value

Table 2.2: Minimum CVR value based on the number of experts

No. of Experts	**Minimum CVR Value**
5	0.99
6	0.99
7	0.99
8	0.75
9	0.78
10	0.62
11	0.59
12	0.56
13	0.54
14	0.51
15	0.49
20	0.42
25	0.37
30	0.33
35	0.31
40	0.29

Source: [23]

2.3. Step 3- Computation and Analysis . CVR value was computed for each individual and general item by applying the equation suggested by [23] as below:

$$CVRValue = (2Ne/N) - 1$$

Note:
Ne = number of experts who answer “3=essential”
N = total number of experts

The computed CVR value is interpreted in Table 2.1

Referring to [23], only the item with the response “3=essential” is considered valid and should be included in the CVR computation. However [5, 25] suggested that the items with the response “2=important (but not essential)” was considered relevant as regards the positive value result. Consequently, this study considers all items with the answer “3=essential” or “2=important (but not essential)” in the CVR calculation.

In [23] also highlighted the issue of the probability the items get positive CVR value purely based on chance. Therefore, [23] suggested the setting of acceptance criteria for each item based on a minimum CVR value that was settled at 5 percent probability (p = 0.05) concerning the total number of experts involved as detailed in Table 2.2

2.4. Revision and Finalising the Item. Regards to the analysis of the CVR value, the items are revised and finalized.

Table 3.1: Summary of the panel of experts

Expert Code	**Title**	**Experience**
Academic field		
Expert 01	Lecturer in the Faculty of Technology and Informatics	Data management specializing in the data visualization
Expert 02	Lecturer and Head of Department in the Faculty of Information Technology	Data management specializing in the data security and database management
Expert 03	Lecturer in the Faculty of Technology and Information Science	Data management
Expert 04	Lecturer in the Faculty of Technology and Information Science	Data management, information system, impact study, strategic development, and quality model
Public sector		
Expert 05	ICT Officer and Ph.D candidate in the information management field	Data management specializing in the data visualization
Expert 06	ICT Officer in the field of strategic development	Data management and strategic planning
Expert 07	Head of ICT department	Data management and ICT management
Expert 08	ICT Officer in the field of ICT architecture	Data management, system development, and ICT strategic and architecture management

3. Result and discussion. The result of the content validity process is explained in sequence based on the steps adapted from [5, 1].

3.1. Step 1- Selection of a Panel of Experts. A total of eight experts from the public sector and academic fields had been involved in the content validation process. The list of participating panel of experts is shown in Table 3.1.

3.2. Step 2- Issuance of Invitations and Distribution of Instruments. All selected panels of experts were unofficially approached through phone calls and online medium communication to get prior consent before the issuance of an official invitation. Then, an official email was sent to the panel of experts with details on how to perform the content validity. The email was attached with an official letter endorsed by the institutions and the instrument to be validated. The panel of experts was given 14 days to return the completed instrument through e-mail. The panel of experts is also allowed to contact the researcher if needs further explanation on the content of the instrument.

3.3. Step 3- Computation and Analysis. The instrument consists of 95 items covering 88 individual items and 7 general items. The formula proposed by [23] was applied to calculate the CVR value for individual and general items. Referring to the explanation in Section III for step 3, the minimum CVR value to be accepted is depending on the total number of experts selected. Since this study appointed eight experts, hence the minimum CVR value for the item to be accepted is 0.75. An example of the CVR value calculation for the strategic planning factor and operation focus factor is depicted in Table 3.2.

Based on Table 3.2 all items in the strategic planning factor are accepted due to all items receiving a CVR value of 0.75 and above (refer to Table 2.2). Meanwhile, only 11 out of 12 items in the operation focus factor are accepted for getting the CVR value of 0.75 and above (refer to Table 2.2), whilst one item (14.B2-8) was rejected for getting the CVR value of 0.50 (refer to Table 2.2). The result for the whole instrument is detailed in Table 3.3.

Referring to Table 3.3 above, the analysis of the content validity for individual and general items is explained below:

- For individual items, out of 88 items, 66 items had a CVR value of 1.00, 19 items had a CVR value of 0.75, and 3 items had a CVR value of 0.50. Referring to Table V, items with a CVR value of 0.75 and above are accepted, making 85 items accepted and 3 items rejected.

Table 3.2: Example of CVR calculation for strategic planning and operation focus factor

Item No	**Expert No.**	**Answer = 2 or 3**	**CVR Value**
	1 2 3 4 5 6 7 8		
A2 – Strategic planning			
4.A2-1	3 2 3 3 3 3 2 3	8	1.00
4.A2-2	3 3 3 3 3 3 1 2	7	0.75
4.A2-3	3 2 3 3 3 3 2 2	8	1.00
4.A2-4	3 2 3 3 1 3 2 3	7	0.75
4.A2-5	3 3 3 3 3 3 3 3	8	1.00
5.A2-6	3 2 3 3 3 3 3 3	8	1.00
5.A2-7	3 3 3 3 3 3 3 2	8	1.00
5.A2-8	3 3 3 3 3 3 2 2	8	1.00
5.A2-9	3 2 3 3 3 3 2 2	8	1.00
5.A2-10	3 3 3 3 3 3 3 2	8	1.00
5.A2-11	3 3 3 3 3 3 2 2	8	1.00
6. (generic)	3 2 3 3 3 3 2 2	8	1.00
B2 – Operation focus			
13.B2-1	3 3 3 3 3 3 2 2	8	1.00
13.B2-2	3 3 3 3 3 3 2 3	8	1.00
13.B2-3	3 2 3 3 3 3 2 2	8	1.00
13.B2-4	3 2 3 3 3 3 2 3	8	1.00
13.B2-5	3 3 3 3 3 3 2 3	8	1.00
13.B2-6	3 3 3 3 3 3 2 2	8	1.00
13.B2-7	3 2 3 3 3 3 3 2	8	1.00
14.B2-8	3 2 3 3 3 3 1 1	6	0.50
14.B2-9	3 2 3 2 3 3 2 1	7	0.75
14.B2-10	3 3 3 3 2 3 3 2	8	1.00
14.B2-11	3 2 3 3 3 3 2 1	7	0.75
15. (generic)	3 3 3 3 3 3 2 2	8	1.00

- Items that get a CVR value equal to 1.00 indicate that all experts evaluate the respected item with the combination of either “3 = essential”, or “2=important (but not essential)” only.
- Items that have a CVR value of 0.75 indicate that at least one panel of experts evaluate the item as “1 = not relevant”.
- Items that have a CVR value of 0.50 indicate more than two panels of experts evaluate the item as “1 = not relevant”. For all three items, a total of two panels of experts give responses “1 – not relevant”, one panel of an expert gives responses “2 = relevant (not essential)”, and five panels of experts give responses “3 = essential”. However, no comment was provided by any panel of experts for those three items.
- As for the rejected items namely 11.B1-7, 14.B2-8, and 22.D1-15 indicate that all three items are not suitable and not representing the study concept and should be rejected. The details of the items are 11.B1-7 (managing staff performance related to the achievement of data quality through the practice of giving rewards, recognition, and also penalties) from the human resource factor, 14.B2-8 (control the costs involved in the production of data products through increase productivity, reduce errors and perform corrections) from the operation focus factor, and 22.D1-15 (cost saving because additional staff is required) from the result factor. This study was contextualized within the public sector domain. Hence, the rejection of all these three items highlighted that the practice of giving acknowledgment or imposing punishment to manage staff’s performance, controlling costs by improving the production process and improving the organization’s operation performance by hiring additional staff do not reflect the government’s convention. Thus, the elimination of these three factors further improves the validity

Table 3.3: CVR value of individial and general iteams for managing master data quality

Element and Factor	**CVR Value for Individual Item**	**CVR Value for Generic Item**
Element A: Leadership		
Factor A1: Leadership	**7 items = 1.00** 1.A-1, 1.A1-3, 1.A1-5, 2.A1-8, 2.A1-9, 2.A1-11, 2.A1-12 **5 items = 0.75** 1.A1-2, 1.A1-4, 1.A1-6, 2.A1-7, 2.A1-10	**1 item = 1.00** Item 3
Factor A2: Strategic planning	**9 items = 1.00** 4.A2-1, 4.A2-3, 4.A2-5, 5.A2-6, 5.A2-7, 5.A2-8, 5.A2-9, 5.A2-10, 5.A2-11 **2 items = 0.75** 4.A2-2, 4.A2-4	**1 item = 1.00** Item 6
Factor A3: Customer focus	**10 items = 1.00** 7.A3-1, 7.A3-2, 7.A3-3, 7.A3-4, 7.A3-5, 8.A3-7, 8.A3-8, 8.A3-9, 8.A3-10, 8.A3-11 **1 items = 0.75** 7.A3-6	**1 item = 1.00** Item 9
Element B: System		
Factor B1: Human resource focus	**7 items = 1.00** 10.B1-1, 10.B1-2, 10.B1-3, 10.B1-4, 11.B1-8, 11.B1-9, 11.B1-10 **2 items = 0.75** 10.B1-5, 11.B1-6 **1 item = 0.05** 11.B1-7	**1 item = 1.00** Item 12
Factor B2: Operation focus	**8 items = 1.00** 13.B2-1, 13.B2-2, 13.B2-3, 13.B2-4, 13.B2-5, 13.B2-6, 13.B2-7, 14.B2-10 **2 items = 0.75** 14.B2-9, 14.B2-11 **1 item = 0.05** 14.B2-8	**1 item = 1.00** Item 15
Element C: Master Data Quality		
Factor C1: Master data quality	**12 items = 1.00** 16.C1-1, 16.C1-2, 16.C1-3, 16.C1-4, 17.C1-5, 17.C1-6, 17.C1-8, 17.C1-9, 18.C1-10, 18.C1-11, 18.C1-13, 19.C1-15 **3 items = 0.75** 17.C1-7, 18.C1-12, 18.C1-14	**1 item = 1.00** Item 20
Element D: Result		
Factor D1: Result	**13 items = 1.00** 21.D1-1, 21.D1-3, 21.D1-4, 21.D1-5, 21.D1-6, 21.D1-7, 21.D1-8, 22.D1-10, 22.D1-11, 22.D1-12, 22.D1-16, 22.D1-17, 22.D1-18 **4 items = 0.75** 21.D1-2, 22.D1-9, 22.D1-13, 22.D1-14 **1 item = 0.05** 22.D1-15	**1 item = 1.00** Item 23

of the framework in the context of the public sector.

- For the general item, all seven items have a CVR value of 1.00. Referring to [26], the result indicates that the measured factors are relevant.

3.4. Revision and Finalising the Item. According to the calculation and analysis performed in Step 3, 92 items are accepted and ready to be used in the subsequent phase. The summary of the finalised item is shown in Table 3.4.

Table 3.4: Summary of the finalized item

Element/ Factor	**Total Initial Item**	**Total Rejected Item**	**Total Accepted Item**
A: Leadership			
A1: Leadership	13	0	13
A2: Strategic planning	12	0	12
A3: Customer focus	12	0	12
B: System			
B1: Human resource focus	11	1 (11.B1-7)	10
B2: Operation focus	12	1 (14.B2-8)	11
C: Master Data Quality			
C1: Master data quality	16	0	16
D: Result			
D1: Result	19	1 (22.D1-15)	18
Total item	**95**	**3**	**92**

4. Conclusion. Based on the CVR value analysis, altogether 92 items are accepted that comprising 85 specific items and seven general items. However, 3 items were rejected which are one item each for the human resource factor, operation focus factor, and result factor. In conclusion, the content validation process involved eight experts successfully validating the questionnaire, and acceptable to be applied as an instrument to validate the MDQM framework. The research findings contribute theoretically to the TQM body of knowledge by extending the concept of master data quality into the TQM thrust and also practically improving the master data quality management in the domain of the public sector. Consequently, the validated MDQM framework can be used by the organization to manage master data quality more systematically.

5. Acknowledgment. This study was financially supported by the Research Grants ETP-2013-060, TAP & FTM of Universiti Kebangsaan Malaysia, and the Public Service Department (PSD) of the Prime Minister's Department.

REFERENCES

[1] W. A. Z. W. Ahmad, M. Mukhtar, and Y. Yahya, *Validating the contents of a social content management framework*, in 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI), IEEE, 2017, pp. 1–6.

[2] ———, *Developing and validating an instrument for social content management*, innovations, 12 (2020), p. 13.

[3] M. H. Alanazi, *The mediating role of primary tqm factors and strategy in the relationship between supportive tqm factors and organisational results: An empirical assessment using the mbnqa model*, Cogent Business & Management, 7 (2020), p. 1771074.

[4] H. Alenezi, A. Tarhini, A. Alalwan, N. Al-Qirim, et al., *Factors affecting the adoption of e-government in kuwait: A qualitative study*, Electronic Journal of e-Government, 15 (2017), pp. pp84–102.

[5] N. Ali, A. Tretiakov, and D. Whiddett, *A content validity study for a knowledge management systems success model in healthcare*, JITTA: Journal of Information Technology Theory and Application, 15 (2014), p. 21.

[6] C. Batini, C. Cappiello, C. Francalanci, and A. Maurino, *Methodologies for data quality assessment and improvement*, ACM computing surveys (CSUR), 41 (2009), pp. 1–52.

[7] A. Bhattacherjee, *Individual trust in online firms: Scale development and initial test*, Journal of management information systems, 19 (2002), pp. 211–241.

[8] M. G. Brown, *Baldrige award winning quality: How to interpret the Baldrige criteria for performance excellence*, CRC press, 2017.

[9] L. L. Davis, *Instrument review: Getting the most from a panel of experts*, Applied nursing research, 5 (1992), pp. 194–197.

[10] A. L. Delbecq, A. H. Van de Ven, and D. H. Gustafson, *Group techniques for program planning: A guide to nominal group and Delphi processes*, Scott, Foresman,, 1975.

[11] R. F. DeVellis and C. T. Thorpe, *Scale development: Theory and applications*, Sage publications, 2021.

[12] R. K. Gable and M. B. Wolf, *Instrument development in the affective domain: Measuring attitudes and values in corporate and school settings*, vol. 36, Springer Science & Business Media, 2012.

[13] F. Haneem, N. Kama, A. Azmi, A. Azizan, S. M. Sam, O. Yusop, and H. Abas, *Master data definition and the privacy classification in government agencies: Case studies of local government*, Advanced Science Letters, 23 (2017), pp. 5094–5097.

[14] F. Haneem, N. Kama, N. Taskin, D. Pauleen, and N. A. A. Bakar, *Determinants of master data management adoption*

by local government organizations: An empirical study, International Journal of Information Management, 45 (2019), pp. 25–43.

[15] H. Hannila, R. Silvola, J. Harkonen, and H. Haapasalo, *Data-driven begins with data; potential of data assets*, Journal of Computer Information Systems, 62 (2022), pp. 29–38.

[16] M. K. Hasan, M. Akhtaruzzaman, S. R. Kabir, T. R. Gadekallu, S. Islam, P. Magalingam, R. Hassan, M. Alazab, and M. A. Alazab, *Evolution of industry and blockchain era: monitoring price hike and corruption using biot for smart government and industry 4.0*, IEEE Transactions on Industrial Informatics, 18 (2022), pp. 9153–9161.

[17] M. K. Hasan, T. M. Ghazal, A. Alkhalifah, K. A. Abu Bakar, A. Omidvar, N. S. Nafi, and J. I. Agbinya, *Fischer linear discrimination and quadratic discrimination analysis–based data mining technique for internet of things framework for healthcare*, Frontiers in Public Health, 9 (2021), p. 737149.

[18] E. Hassan, *Model pengurusan kualiti maklumat sektor awam Malaysia*, PhD thesis, UKM, Bangi, 2019.

[19] S. N. Haynes, D. Richard, and E. S. Kubany, *Content validity in psychological assessment: A functional approach to concepts and methods.*, Psychological assessment, 7 (1995), p. 238.

[20] H. M. Judi, H. Hashim, and T. S. M. T. Wook, *Knowledge sharing driving factors in technical vocational education and training institute using content analysis*, Asia-Pacific Journal of Information Technology and Multimedia, 7 (2018), pp. 11–28.

[21] D. Kaur and D. Singh, *Master data management maturity evaluation: A case study in educational institute*, in ICT with Intelligent Applications: Proceedings of ICTIS 2022, Volume 1, Springer, 2022, pp. 211–220.

[22] A. Khan, M. K. Hasana, T. M. Ghazal, S. Islam, H. M. Alzoubi, U. Asma'Mokhtar, R. Alam, and M. Ahmad, *Collaborative learning assessment via information and communication technology*, in 2022 RIVF International Conference on Computing and Communication Technologies (RIVF), IEEE, 2022, pp. 311–316.

[23] C. H. Lawshe et al., *A quantitative approach to content validity*, Personnel psychology, 28 (1975), pp. 563–575.

[24] C. Legner, T. Pentek, and B. Otto, *Accumulating design knowledge with reference models: insights from 12 years' research into data management*, Journal of the Association for Information Systems, 21 (2020), p. 2.

[25] B. R. Lewis, G. F. Templeton, and T. A. Byrd, *A methodology for construct development in mis research*, European Journal of Information Systems, 14 (2005), pp. 388–400.

[26] C. Liu, D. Zowghi, and A. Talaei-Khoei, *An empirical study of the antecedents of data completeness in electronic medical records*, International Journal of Information Management, 50 (2020), pp. 155–170.

[27] M. R. Lynn, *Determination and quantification of content validity*, Nursing research, 35 (1986), pp. 382–386.

[28] S. B. MacKenzie, P. M. Podsakoff, and N. P. Podsakoff, *Construct measurement and validation procedures in mis and behavioral research: Integrating new and existing techniques*, MIS quarterly, (2011), pp. 293–334.

[29] R. Mahanti and R. Mahanti, *Data and its governance*, Data Governance and Data Management: Contextualizing Data Governance Drivers, Technologies, and Tools, (2021), pp. 5–82.

[30] J. F. McKenzie, M. L. Wood, J. E. Kotecki, J. K. Clark, and R. A. Brey, *Establishing content validity: Using qualitative and quantitative steps.*, American Journal of Health Behavior, (1999).

[31] S. Nelke, M. Oberhofer, Y. Saillet, and J. Seifert, *Method and system for accessing a set of data tables in a source database*, June 12 2018. US Patent 9,996,558.

[32] B. Otto, *Quality management of corporate data assets*, in Quality management for IT services: Perspectives on business and process performance, IGI Global, 2011, pp. 193–209.

[33] ———, *How to design the master data architecture: Findings from a case study at bosch*, International journal of information management, 32 (2012), pp. 337–346.

[34] X. Peng, V. Prybutok, and H. Xie, *Integration of supply chain management and quality management within a quality focused organizational framework*, International Journal of Production Research, 58 (2020), pp. 448–466.

[35] T. Pentek, C. Legner, and B. Otto, *Towards a reference model for data management in the digital economy*, in Designing the Digital Transformation: DESRIST 2017 Research in Progress Proceedings of the 12th International Conference on Design Science Research in Information Systems and Technology. Karlsruhe, Germany. 30 May-1 Jun., Karlsruher Institut für Technologie (KIT), 2017, pp. 73–82.

[36] D. M. Rubio, M. Berg-Weger, S. S. Tebb, E. S. Lee, and S. Rauch, *Objectifying content validity: Conducting a content validity study in social work research*, Social work research, 27 (2003), pp. 94–104.

[37] C. A. Schriesheim, K. J. Powers, T. A. Scandura, C. C. Gardiner, and M. J. Lankau, *Improving construct measurement in management research: Comments and a quantitative approach for assessing the theoretical content adequacy of paper-and-pencil survey-type instruments*, Journal of Management, 19 (1993), pp. 385–417.

[38] U. Sekaran and R. Bougie, *Research methods for business: A skill building approach*, john wiley & sons, 2016.

[39] P. Shamala, R. Ahmad, A. Zolait, and M. Sedek, *Integrating information quality dimensions into information security risk management (isrm)*, Journal of Information Security and Applications, 36 (2017), pp. 1–10.

[40] R. Silvola, J. Harkonen, O. Vilppola, H. Kropsu-Vehkapera, and H. Haapasalo, *Data quality assessment and improvement*, International Journal of Business Information Systems, 22 (2016), pp. 62–81.

[41] R. Silvola, A. Tolonen, J. Harkonen, H. Haapasalo, and T. Mannisto, *Defining one product data for a product*, International Journal of Business Information Systems, 30 (2019), pp. 489–520.

[42] M. Spruit and K. Pietzka, *Md3m: The master data management maturity model*, Computers in Human Behavior, 51 (2015), pp. 1068–1076.

[43] D. Straub, M.-C. Boudreau, and D. Gefen, *Validation guidelines for is positivist research*, Communications of the Association for Information systems, 13 (2004), p. 24.

[44] D. R. Tojib and L.-F. Sugianto, *Content validity of instruments in is research*, Journal of Information Technology Theory and Application (JITTA), 8 (2006), p. 5.

[45] R. Vilminko-Heikkinen and S. Pekkola, *Changes in roles, responsibilities and ownership in organizing master data management*, International Journal of Information Management, 47 (2019), pp. 76–87.

[46] R. Y. Wang and D. M. Strong, *Beyond accuracy: What data quality means to data consumers*, Journal of management information systems, 12 (1996), pp. 5–33.

[47] Y.-S. Wang, *Assessment of learner satisfaction with asynchronous electronic learning systems*, Information & Management, 41 (2003), pp. 75–86.

[48] K. Weber, B. Otto, and H. Österle, *One size does not fit all—a contingency approach to data governance*, Journal of Data and Information Quality (JDIQ), 1 (2009), pp. 1–27.

[49] P. L. Williams and C. Webb, *The delphi technique: a methodological discussion*, Journal of advanced nursing, 19 (1994), pp. 180–186.

[50] D. V. Zúñiga, R. K. Cruz, C. R. Ibañez, F. Dominguez, and J. M. Moguerza, *Master data management maturity model for the microfinance sector in peru*, in Proceedings of the 2nd international conference on information system and data mining, 2018, pp. 49–53.

Edited by: S. B. Goyal
Special issue on: Soft Computing and Artificial Intelligence for wire/wireless Human-Machine Interface
Received: Oct 27, 2023
Accepted: Feb 5, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 2013–2020, DOI 10.12694/scpe.v25i3.2688

ENTERPRISE AUDIT RISK ASSESSMENT AND PREVENTION BASED ON AHP ANALYSIS

GUOLIANG SUN[*]AND BAYI GUAN[†]

Abstract. If there are no auditing standards or auditing processes of big data, the audit risks of enterprises are increased. This paper first introduces the topic through the research background and literature review in order to ensure the integrity and accuracy of audit evidence to the maximum extent, and then analyses the causes of enterprise audit risk. When analysing the risk level of material misstatement, it is mainly the audit risk generated by the enterprise's unique business model, information system and financial management. Audit risk is mainly caused by the ability of auditors and audit process management. After quantitative analysis of the correctness of enterprise audit risk assessment indicators, this paper builds a multi-level comprehensive assessment model of enterprise audit risk on the basis of AHP analysis. At the same time, this paper puts forward specific measures to improve audit methods and audit processes and prevent audit risks in view of the actual problems encountered in the current audit risk of enterprises, so as to provide certain references for enterprise risk management and control.

Key words: AHP analytical method; risk audit; risk prevention

1. Introduction. Audit risk management is the core of risk-oriented audit, and audit risk assessment is an important part of enterprise risk management. On the basis of the qualitative identification of audit risk factors, these kinds of research make quantitative research and related evaluation of risk factors [1], which are the premise of enterprise audit risk disposal. With the deepening of the study of audit risk by scholars at home and abroad, most researchers currently have two mainstream understandings of the connotation of audit risk assessment for enterprises. One is that audit risk is a strategic risk for enterprises based on the overall, comprehensive and strategic characteristics of enterprise risk management, which will affect the development direction, process and even survival of enterprises [2]. Another view is that audit risk is an effective measure to avoid risks in the course of business activities, and audit risk comes from the strategic management of enterprises. This paper holds that the broad audit risk is the result of the joint action of audit market competition risk from the perspective of audit process, audit project risk and audit expectation gap risk [3]. The essence of audit risk is the risk of the relationship between accounting firms and audit clients, and the risk that accounting firms cannot continue to audit the original clients. Audit project risk includes audit project material misstatement risk and inspection risk. Audit expectation difference is the objective fact that the public and the audit circle have different understanding of audit content and function, and the risk of audit expectation difference is the possibility that this objective fact will cause loss to the audit supplier [4]. The broad definition of audit risk is as follows: audit risk = audit market competition risk × audit project risk × audit expectation difference risk. The narrow sense of audit risk refers to the audit project risk. Considering the general applicability of audit standards, this paper mainly starts from the perspective of narrow audit risk when analyzing enterprise audit risk and its characteristics [5].

2. Theoretical.

2.1. Research needs. In the case of non-standardization of big data auditing standards, new business models have led to changes in the carriers of financial information and business information recorded by enterprises. In addition, many business models have certain uncertainties, which leads to risks in the business activities of enterprises, and then increases the audit difficulty and audit risks. The traditional audit model

[*] School of International Business, Zhengzhou Tourism College, Zhengzhou, 451464, China (sunguoliang@zztrc.edu.cn)
[†] School of International Business, Zhengzhou Tourism College, Zhengzhou, 451464, China (Corresponding author, guanbayi@zztrc.edu.cn)

is not suitable for modern enterprise risk management [6]. This paper establishes the index system of the enterprise audit risk assessment.

2.2. Objectives of the Study. After analyzing the causes of enterprise audit risk, this paper studies the enterprise audit risk evaluation index system quantitatively. The purpose of this study is to remind auditors to pay attention to audit risks different from the traditional audit model, and enterprises need to improve audit methods and audit processes to prevent audit risks. The research results of this paper can provide some reference for improving the audit quality of intelligent information.

2.3. Risk of material misstatement (B1). The risk of material misstatement refers to the possibility of material misstatement of financial statements prior to audit. The risk of material misstatement is a risk that auditors can detect, evaluate but cannot control [7]. It includes :(1) the degree of soundness of relevant laws, regulations, systems and standards (B11). (2) Economic environment (B12). The macroeconomic situation is the most important external environment for the production and operation of enterprises, and the risks faced by enterprises in different economic cycles are different. (3) Policy orientation (B13). The intervention of government policies and the stability of society will affect the normal operation of enterprises, which may lead to the risk of major misstatement. (4) Partner (B14). The enterprise is a number of relevant fit relationships, the selection of partners will determine the operation of the entire enterprise and the final profit. (5) Management level (B15). The partner enterprises in the enterprise usually face different corporate culture and management mode, different technical standards and hardware environment, which greatly increases the risk of management operation, and may directly lead to management loss of control. (6) Information Security (B16) In the highly competitive economic society, the magic weapon is to have the core technology for an enterprise to survive and succeed in competition that other enterprises do not have, and the leakage of the core technology is a fatal blow to the survival and development of enterprises. (7) Core technology (B17). Although enterprises emphasize mutual trust among partners, and the rapid development of information technology makes information sharing a trend, information asymmetry is still a key attribute in the reality of enterprises, and directly leads to the emergence of various unethical behaviors such as false information and cheating in enterprises. (8) Business ethics (B18). Business ethics is the sum of behavioral norms used to regulate the relationship between enterprises and society, enterprises and enterprises, enterprises and workers. It is not only an important part of the social moral system, but also it is the social moral principles and norms of business behavior.

2.4. Check Risks (B2). The risk of inspection is the possibility that a determination has a misstatement that, alone or in conjunction with other misstatements, which would be material but that the CPA has failed to detect such a misstatement. Inspection risk is the risk that auditors can control [8]. It includes:

1. Audit services (B21). The auditor's understanding of virtual enterprise, the network audit, the audit technology of E-commerce, and related audit software will affect the size of the inspection risk.
2. Audit process (B22). An enterprise is a complex system, the design of its audit program is a complex project, and it is also an important concern of auditors when designing audit program.
3. Audit Principles (B23). This also directly increases the risk of inspection.

2.5. Evaluation index system. Analytic Hierarchy Process (AHP) decomposes the decision problem into different hierarchical structures according to the order of the overall goal, sub-goals of each level, evaluation criteria and specific backup plan. Then, by solving the eigenvector of the judgment matrix, the priority weight of each element at each level on an element at the upper level is obtained. Finally, the final weight of each alternative plan on the overall goal is recurred by the method of weighting sum. The one with the greatest final weight is the optimal scheme. AHP can optimize the connection relationship of each layer and its sub-evaluation indicators, and reduce the uncertain factors in the evaluation process to a great extent. This paper establishes an enterprise audit risk assessment index system based on AHP, as shown in Table 2.1.

The identified risk variables still have some subjectivity. Therefore, the obtained risk factors can be further verified by questionnaire survey. This paper collects a large amount of data about audit risk assessment indicators with means of consulting experts and investigation. In this paper, SPSS software is used to analyze the risk factors. Reliability analysis is a common testing method to verify whether the scale questionnaire is scientific and reasonable. Reliability analysis was carried out by SPSS mainly to see the value of Cronbach's alpha reliability coefficient after analysis. In general, most researchers believe that Cronbach's alpha reliability

Table 2.1: Evaluation index system of enterprise audit

Target layer B	Criterion layer Bn	Index layer Bnm
Risk assessment of enterprise audit	Risk of material misstatement of index system B1	Soundness of institutional guidelines B11
		Economic environment B12
		Policy orientation B13
		Partner B14
		Management level B15
		Information Security B16
		Core technology B17
		Business ethics B18
	Check the risk B2	Audit Service B21
		Audit process B22
		Audit Principle B23

Table 2.2: Reliability test of the risk factors

Indexes	Risk factors	Cronbach's Alpha if Item Deleted	Cronbach's AlPha
B1			0.75
	B11	0.80	
	B12	0.82	
	B13	0.71	
	B14	0.82	
	B15	0.88	
	B16	0.70	
	B17	0.86	
	B18	0.61	
B2			0.83
	B21	0.82	
	B22	0.80	
	B23	0.84	

coefficient is greater than 0.6, and indicating that the data is reliable to a certain extent. The reliability coefficient of the questionnaire survey calculated by SPSS is shown in Table 2.2.

As can be seen from the table, the reliability coefficient of the overall data is greater than 0.6. This indicates that the evaluation indicators selected are of high credibility in this paper. The Analytic Hierarchy Process (AHP) is the research basis of comprehensive evaluation. The basic steps of using this analysis method:

The first step is to determine the index system of hierarchical analysis based on comprehensive analysis, and determine the evaluation factors of the target layer (B), criterion layer (Bn) and index layer (Bnm).

The second step, the indicators at the same level are compared in pairwise in the index system. Generally, the importance degree is scored and assigned by one's own experience or organized experts, and a judgment matrix is constructed: B=Bnm, whose elements are as follows:

$$B_{ij} > 0, B_{ij} = \frac{1}{B_{ij}}, \sum_{i=1}^{n}\sum_{j=1}^{m} B_{ij} = 1, \quad (i = 1, 2, \ldots, n;\ j = 1, 2, \ldots, m) \tag{2.1}$$

where, the ratio of importance of element i to element j is B_{ij}. The third step is to calculate the weights. For each line of elements of the judgment matrix, the product Ai of each line of elements is calculated with the behavior vector, and the calculation formula is shown as follows:

$$A_i = \prod_{i=1}^{n} Y_i, \quad (i = 1, 2, \ldots, n) \tag{2.2}$$

For the n-order judgment matrix, the result of m is calculated according to the above formula, and the normalized eigenvector value of each row is calculated to the judgment matrix with the eigenvector $W = (w_1, w_2, ...w_n)T$. The calculation formula is as follows:

$$w_i = \frac{\sqrt[n]{A_i}}{\sum_{i=1}^{n} \sqrt[n]{A_i}} \tag{2.3}$$

The maximum characteristic root max is calculated to the judgment matrix C according to the n-order judgment matrix C and the corresponding eigenvector W.

$$\lambda_{\max} = \sum_{i=1}^{n} \frac{Cw_i}{nw_i} \tag{2.4}$$

The fourth step is hierarchical total sorting and consistency checking. Because of the complexity of objective things and the fuzziness and diversity of people's understanding of things, the judgment matrix given may not be completely consistent. So that it is necessary to carry out consistency test. When the order of the judgment matrix is less than or equal to 2 orders, there is no possibility of the above inconsistency. Then it can be directly judged that it meets the condition of complete consistency. When the order is greater than 2, consistency judgment is required. The consistency index CI value of the judgment matrix at each layer can be expressed as:

$$CI = \frac{\lambda_{\max} - n}{n - 1} \tag{2.5}$$

After calculating the relative importance of factors at all levels, the overall weight of factors can be calculated at each level on the whole evaluation target according to the principle from high level to low level. So that it is necessary to carry out the overall hierarchical ranking. Then the random consistency ratio of the next layer is:

$$CB = \frac{CI \sum_{j=1}^{m} C_j}{GC_j} \tag{2.6}$$

where G is the randomness index.

3. Research method (Multi-level comprehensive evaluation method). The multilevel comprehensive evaluation method is is used to effectively solve the large errors caused by subjective factors and the comprehensive evaluation method based on the AHP, which has greater reliability and practicability. When the problem has uncertainty and fuzziness, the comprehensive evaluation model can be used to deal with it. After comprehensive consideration of various influencing factors, this paper chooses the multi-level comprehensive evaluation method, constantly optimizes the evaluation index system, and makes clear the weight of each evaluation index.

1. Subjection matrix. With the problem evaluation index determined by the above analytic hierarchy process, two finite sets are assumed according to the comprehensive evaluation method: D=d1,d2,... ,dn, set B=B1,B2,... Bn. D represents the set of evaluation factors. The best evaluation result is obtained from alternative concentration after considering various influencing factors. These evaluation index factors are fuzzy and uncertain to a certain extent, and some can be considered as definite values. The establishment of the membership function must consider the variation law of each single index. When there are a large number of indicators, it is necessary to classify the indicators and calculate

the membership degree B for each evaluation level according to the actual value of each evaluation indicator. The form can be expressed as:

$$B = \begin{bmatrix} B_{11} & \cdots & B_{1n} \\ \vdots & \ddots & \vdots \\ B_{m1} & \cdots & B_{mn} \end{bmatrix} \tag{3.1}$$

In the formula, Bij represents the membership degree of d_i evaluation to grade, and $\sum B_{ij} = 1$ after normalization treatment. According to the value domain Vi divided by evaluation grade, the calculation of positive effect index Bi can be expressed as:

$$B_i = \begin{cases} 0 & \text{if } d_i \in (0, x_i) \\ \left(\frac{B_i - x_i}{x_{i+1} - x_i}\right)^{0.5} & \text{if } d_i \in [x_i, x_{i+1}) \\ 1 & \text{if } d_i \in [x_{i+1}, x_{i+2}) \\ 1 - \left(\frac{d_i - x_{i+2}}{x_{i+3} - x_{i+2}}\right)^{0.5} & \text{if } d_i \in [x_{i+2}, x_{i+3}) \\ 0 & \text{if } d_i \in [x_{i+3}, +\infty) \end{cases} \tag{3.2}$$

According to the value domain xi divided by comparison relationship and evaluation grade, the calculation formula of negative effect index Bi can be expressed as follows as long as the conditional interval of the above equation is reversed:

$$B_i = \begin{cases} 0 & \text{if } d_i \in (x_i, +\infty) \\ \left(\frac{d_i - x_i}{x_{i+1} - x_i}\right)^{0.5} & \text{if } d_i \in (x_{i+1}, x_i] \\ 1 & \text{if } d_i \in (x_{i+2}, x_{i+1}] \\ 1 - \left(\frac{d_i - x_{i+2}}{x_{i+3} - x_{i+2}}\right)^{0.5} & \text{if } d_i \in (x_{i+3}, x_{i+2}] \\ 0 & \text{if } d_i \in (-\infty, x_{i+3}] \end{cases} \tag{3.3}$$

2. Comprehensive evaluation result model. With the weight matrix and membership matrix obtained above, this paper establishes the measurement model of the comprehensive evaluation result Y:

$$Y = CB = \begin{bmatrix} B_1 & B_2 & \ldots & B_m \end{bmatrix} \begin{bmatrix} B_{11} & \cdots & B_{1n} \\ \vdots & \ddots & \vdots \\ B_{m1} & \cdots & B_{mn} \end{bmatrix} \tag{3.4}$$

Above the formula, Yij represents the comprehensive subordination degree of the evaluation index to the evaluation level. The calculation model of Yij in this paper is selected as follows:

$$Y_{ij} = \min\{1, \sum_{i=1}^{n} \sum_{j=1}^{m} \min(B_{ij}, B_{ij})\}, \quad i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, m \tag{3.5}$$

For the evaluation and analysis of the comprehensive evaluation result Y, the maximum subjection degree method and the comprehensive score value method are usually used. The maximum subjection degree method is to select the one with the greatest subjection degree from each evaluation result vector in Y. And it is believed that the evaluation index belongs to this evaluation level. The critical value of each evaluation grade can be calculated by using the comprehensive score value method. Then the corresponding vector in $Y$is used to calculate the comprehensive score value V.

$$V = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} (Y_{ij}^{e} \times B_{ij})}{\sum_{i=1}^{n} \sum_{j=1}^{m} Y_{ij}^{e}} \tag{3.6}$$

Where e is the reduction coefficient, the purpose is to weaken the weight position of larger Y_i. When e tends to infinity, the comprehensive score value method is essentially the maximum subjection method.

According to the comprehensive score value, this paper can evaluate the audit risk of enterprises. In general, the higher the comprehensive score, the higher the audit risk. On the contrary, it indicates that the audit risk of enterprises is lower.

In this paper, according to the nature of the problem and the overall goal to be achieved, the problem is decomposed into different components, and the factors are aggregated and combined according to different levels according to the interrelated influence and membership relationship among the factors, forming a multi-level analysis structure model, so that the problem is finally reduced to the lowest level relative to the highest level of the relative merits and demerits of the arrangement. With the evaluation process of each factor of objective layer (B), criterion layer (Bn) and index layer (Bnm), this paper establishes a comprehensive evaluation model to achieve the purpose of audit risk assessment of enterprises.

4. Results and discussion.

4.1. Conduct full sample audit. An enterprise may have diversified sources of revenue, more complex types of business, and more preferential policies, and auditors must understand and evaluate whether the accounting policies and accounting estimates are consistent with the new revenue standards. Enterprises have a large number of small amount but high volume of transactions, and small amount of online transactions are easy to be faked and not easy to be discovered [10]. When auditors plan to implement audit procedures, they should appropriately tilt audit resources to income based on the principle of cost-effectiveness, expand the scope of audit objects, and take full sample audit to examine all transactions and transaction volumes of enterprises in detail [11].

4.2. Focus on business model audits. During auditing, auditors should first analyze the business model of the enterprise, focusing on whether it conforms to the development of the market, and whether it conforms to the scale of the enterprise, and whether it corresponds to the characteristics of the enterprise [9]. The business model of online transaction, offline experience and platform logistics delivery of enterprises makes them more complex than the related party transactions of traditional enterprises [12]. Auditors can conduct data analysis through audit software and pay more attention to whether there is collusion and fraud in related party transactions and management. This can avoid the risk of material misstatement associated with continuing operations [13].

4.3. Focus on audits of information systems. First of all, by establishing a set of audit risk identification and assessment information system, the information system of the audited entity is to identify risks, and review whether it can reflect the complex business process of the enterprise, effective internal control, and whether it can ensure the truth and integrity of the data. Secondly, the risk assessment of the information system is carried out to assess whether the stored data will be lost or damaged and to what extent when the information system is attacked by the outside world. In addition, auditors must use the knowledge, experience and technology of professionals to use computer technology to audit information systems, ensure the security and stability of information systems, and ensure the authenticity of financial data [14].

4.4. Improve relevant audit laws and regulations. The construction of audit regulations that adapt to the era of big data can not only provide legal support for audit work, but also it can provide a legal basis for audit work, and safeguard the legitimate rights and interests of audited institutions [15]. In the form of laws and regulations, it shall be stipulated with the scope, authority and implementation methods of data acquisition. In terms of data acquisition, the first thing is to ensure that the audited entity shall provide the business system and electronic data related to audit evidence as required by the audit entity [16]. The second is to establish an electronic data collection and submission system to standardize the types of data acquisition and guarantee the authenticity and accuracy of data. In terms of data storage and use, the audit unit is required to ensure the security and confidentiality of data, which is conducive to protecting the security of electronic data and auditors, and creating a good data environment [17].

4.5. Standardize the audit process. It is necessary to create a new technical method to audit enterprises from the perspective of big data audit. The audit object has been expanded from basic financial data to semi-structured and unstructured financial data and non-financial data [18]. The audit environment has been expanded from offline audit to off-site audit including online audit, and the audit method has been changed from sampling audit to full-sample data processing and analysis, and data analysis has been added in the audit process from beginning to end. The changes of many factors lead to the increase of audit risk, and it is urgent to improve and enrich the audit theory to standardize and guide the audit process and reduce the audit risk caused by the non-standard audit process [19].

4.6. Use big data and other technologies to obtain audit evidence. When auditing enterprises, auditors can make use of the advantages of big data auditing to analyze financial data and non-financial data such as images, audio and video through audit software. The development of network technology provides a new way for the acquisition of information. Auditors can obtain the original data through Internet technology, that is, extract useful information from a large number of network information and save it. At the same time, the Internet also provides many third-party data acquisition platforms, and auditors can use some reliable and authoritative third-party platforms to obtain data during auditing [20].

4.7. Cultivate high-end composite audit talents. Firstly, the entry threshold of auditors should be raised to ensure the quality of auditors in terms of professional competence. Secondly, strengthen the training of auditors, minimize audit risks caused by auditors' professionalism, ability, professional ethics and other problems in the audit process [21, 22, 23, 24]. And vigorously cultivate audit talents of computer type and audit computer talents. Because of adding computer talents to the audit team, it can improve the quality of auditors, and cultivate team coordination and efficient cooperation.

5. Scope for future research. The future research direction mainly analyzes the need to develop matching financial software for audit work, broaden the channels for collecting audit evidence from third-party platforms or using blockchain technology, establish audit analysis models to promote audit work, and ensure the integrity and accuracy of audit evidence to the maximum extent.

6. Limitations of the Study. Because of the limited data collection channels and the need to continuously improve study level, this paper has conducted a preliminary study on enterprise audit risk. The arguments and suggestions are still at a very superficial stage, and whether they are operable needs to be further tested. In addition, the enterprise application field is very wide, and there are still some differences between individual cases and the overall situation. In the future, the causes of audit risk and how to deal with audit risk factors are worth further exploration and research, so as to improve audit efficiency and audit quality.

7. Conclusion. In the study of enterprise audit risk assessment and preventive measures, this paper first analyzes the influencing factors of enterprise audit risk assessment, so as to clarify the evaluation indicators. Based on the AHP evaluation method, this paper establishes an enterprise audit risk assessment system, and collects a large number of enterprise audit risk assessment index data by consulting experts and investigation methods. After analyzing the risk factors with SPSS software, this paper verifies the reliability of the evaluation indicators. In order to effectively solve the large errors caused by subjective factors, this paper puts forward the comprehensive multi-level evaluation method, which further improves the reliability and practicability of enterprise audit risk assessment system. After putting forward the enterprise audit risk assessment method, this paper puts forward the concrete preventive measures according to the actual behavior and problems encountered by the enterprise audit risk. Through this research, its purpose is to help more enterprises can continue to forge ahead and avoid risks.

REFERENCES

[1] Lili, T. & Wansheng, W. Research on risk prevention of Internal Audit in Small and medium-sized enterprises [J]. *Journal Of Enterprise Accounting.* **2023**, 6-10 (0)

[2] Anjian, W. Risks and Control Strategies faced by Internal Audit of Small and Medium-sized Enterprises [J]. *Finance And Economics.* **2023**, 18-22 (0)

[3] Yanlin, C. Strengthening Internal Audit of Small and medium-sized Enterprises to improve anti-risk ability []. *China Commerce And Trade.* **2023**, 99-102 (0)

[4] Wei.Analysis, S. of Marketing Innovation Under the New Retail Mode-Taking'Luckin coffee' as an Example[J]. *E.* **2021**, 88-92 (0)

[5] Yoon, K. & Hoogduin, L. Big Data as Compilementary Audit Evidence[J]. *Accounting Horizons.* **29**, 66-73 (2015)

[6] Fang, H. Discussion on financial risk management and internal audit of small and medium-sized enterprises [J]. *China Economic And Trade.* **2023**, 66-70 (0)

[7] Dehua, J. Analysis of problems and countermeasures in internal audit of Chinese small and medium-sized enterprises []. *Reform And Management.* **2023**, 45-50 (0)

[8] Dimitris, B. The Impact of Big Data on Accounting and Auditing[J]. *International Journal Of Corporate Finance And Accounting (IJCFA).* **8**, 185-190 (2021)

[9] Jianxin, C. & Xiaoke, Z. Reconstruction of Retail business Model in the context of Big Data: a factor Perspective[J]. *Business Economics Research.* **2021**, 16-19 (0)

[10] Chen, W. & Ju., J. Research on Feature mining methods of Audit Clues based on Big Data Visualization Technology[J]. *Audit Research.* **2018**, 78-83 (0)

[11] Wei, C. Research on Network Audit Risk Control under Big Data Environment[J]. *Chinese Certified Public Accountants.* **2018**, 111-113 (0)

[12] Zuping, D. Audit Risks and Countermeasures in the environment of Big Data Audit[J]. *Chinese And Foreign Entrepreneurs.* **2020**, 41-43 (0)

[13] Zeng, F. Research on Audit Risk Prevention of Enterprises under the New Retail Model[J]. *Guangzhou Quality Supervision Herald.* **2020**, 9-13 (0)

[14] Qiufei, W., Shuang, Q. & Dan, S. Research on Audit risk Identification and Control based on Big Data[J]. *Friends Of Accounting.* **2018**, 91-96 (0)

[15] Wei, W. Reconstruction and Trend Outlook of Traditional Auditing Methods in the Era of Big Data[J]. *The Economist.* **2017**, 161-166 (0)

[16] Liqin, X. Analysis of some Problems in Policy Tracking Audit[J]. *Modern Economic Information.* **2019**, 106-109 (0)

[17] Hongfei, X. A Study on the impact of Business Audit Risk[D]. (Shandong University,2021)

[18] Helmy, A., Jaseemuddin, M. & Multicast-ility, B. a novel architecture for efficient micro mobility[J]. *Selected Areas In Communications.* **22**, 2020 (0)

[19] Weiler, R. How to sharpen virtual business[J]. *Information Week.* **2021**, 132-135 (0)

[20] Markus, L., Manville, B. & Agres, C. What makes a virtual organization work[J]. *Sloan Management Review.* **2021**, 13-26 (0)

[21] Innovation, C. when is virtual virtuous[J]. *Harvard Business Review.* **2022**, 127-134 (0)

[22] Kumar, M., Vimal, S., Jhanjhi, N., Dhanabalan, S. & Alhumyani, H. Blockchain based peer to peer communication in autonomous drone operation. *Energy Reports.* **7** pp. 7925-7939 (2021)

[23] Azeem, M., Ullah, A., Ashraf, H., Jhanjhi, N., Humayun, M., Aljahdali, S. & Tabbakh, T. Fog-oriented secure and lightweight data aggregation in iomt. *IEEE Access.* **9** pp. 111072-111082 (2021)

[24] Lee, S., Abdullah, A. & Jhanjhi, N. A review on honeypot-based botnet detection models for smart factory. *International Journal Of Advanced Computer Science And Applications.* **11** (2020)

Edited by: Zhengyi Chai
Special issue on: Data-Driven Optimization Algorithms for Sustainable and Smart City
Received: Oct 13, 2023
Accepted: Feb 21, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 2021–2033, DOI 10.12694/scpe.v25i3.2698

ECONOMIC DISPATCH OF MULTI REGIONAL POWER SYSTEMS BASED ON CMOPSO ALGORITHM

JINHUA GUO[*]

Abstract. With the progress of society and the aggravation of environmental pollution, the economic dispatch of the power system is developing towards multiple environmental and economic goals. To improve energy utilization efficiency, this study innovatively proposes a multi-objective particle swarm optimization algorithm based on competitive learning, and uses this algorithm to solve multi regional environmental and economic scheduling problems. In addition, the study solves static and dynamic economic (S-DE) scheduling problems in multiple regions through improved competitive group optimization algorithms. The research results show that under different testing systems, the average distribution uniformity indicators of the research algorithm built on competitive learning are 0.8058 and 0.8457, and the average anti generation distance is 67.6316 and 1664.0978. The improved competitive group optimization algorithm solves the maximum, minimum, and average fuel costs for static economic scheduling in multiple regions, which are 656.2243 \$/h, 655.8592 \$/h, and 655.9866 \$/h, respectively. Thus, the designed algorithm can effectively solve economic scheduling problems, which is of great significance for resource integration, saving power generation costs, and reducing pollution emissions.

Key words: Multi region; Economic dispatch; CMOPSO; ImCSO; Constraint condition

1. Introduction.

Economic dispatch, as one of the methods to promote high-quality development of the power system, can ensure the normal operation of the power system. Economic dispatch requires the optimal scheduling of power generation units with the best fuel cost, without violating various operational constraints [1]. The power dispatch system generally involves multiple regions, and it is necessary to dispatch electricity reasonably between regions to fully utilize the system's resources [2]. The early solutions to the problem of multi regional economic dispatch were mathematical methods, such as dynamic programming, gradient algorithms, etc. These mathematical methods have good results in economic scheduling problems with fewer constraints, but they are difficult to solve problems involving complex factors [3]. With the development of technology, more and more researchers are applying heuristic optimization algorithms to economic scheduling problems due to their ability to effectively solve complex problems. However, these studies also have certain shortcomings, such as the scheduling problems involved being relatively single and the constraints considered being incomplete [4]. Based on these issues, this study innovatively proposes the use of Competitive Multiple Objective Particle Swarm Optimization (CMOPSO) based on competitive learning to solve multi regional environmental and economic scheduling problems. The study also adopts an Improved Competitive Swarm Optimizer (ImCSO) algorithm to address S-DE scheduling issues in multiple regions. This study aims to improve energy utilization efficiency and achieve optimal scheduling of power generation units through optimal fuel costs. The study is divided into four parts. The first part is an overview of research related to economic dispatch in the power system. The second part is the design of the algorithm used to solve economic scheduling problems. The third part is the analysis of the results of the research method. The fourth part is the conclusion.

2. Related Works.

Economic dispatch is an important way to promote the high-quality development of the power system and occupies an essential position in the stable operation of the power system. At present, there are many studies on the economic scheduling of power systems. Goudarzi A and other researchers have proposed an intelligent sequence algorithm for synchronous scheduling of electricity and heat. This algorithm includes both an optimization algorithm that combines enthusiasm assistance and mathematics, as well as an improved particle swarm optimization algorithm. In addition, the study also designed constraint management,

[*] College of Economics and Trade, Henan Polytechnic Institute, China, Nan yang, 473000, China (Corresponding author, lwgjh2023@163.com)

indicating that the proposed method is significantly superior in performance to conventional methods [6]. Lyu C and other scholars proposed a new degradation cost model for the cyclic degradation problem in microgrid economic scheduling, and used Wasserstein fuzzy sets to describe uncertainty. In addition, this study also expresses real-time microgrids through distributed robust optimization problems. This method has good performance and can effectively solve the cyclic degradation problem during microgrid economic scheduling [7]. Wang X and other experts designed a sparse polynomial chaotic expansion proxy model based on data-driven to solve the stochastic economic scheduling problem of wind power uncertainty. This model can provide information such as mean and variance in stochastic economic scheduling solutions. This method has high accuracy and efficiency in solving stochastic economic scheduling problems [8]. Marco et al. analyzed existing power energy systems to reduce carbon dioxide emissions and proposed a modeling method for power planning tools. This method utilized computational strategies and linear programming optimization methods. This method had strong analytical ability and can effectively plan the power system [9].

Experts such as Lev K have designed a coordination mechanism between tie line power planning and regional power grid economic dispatch for cross regional power grid economic dispatch under various uncertain power sources and loads. It adopts a quality-of-service approach to analyze the service attributes in power dispatch. In addition, the study also constructed a model free hierarchical optimization method based on learning technology, and used reinforcement learning algorithms to effectively solve the economic dispatch problem of cross regional power grids [10]. Scholars such as Xu D have designed a maximum minimum two-layer optimization model and a two-stage robust optimization model to solve the problem of unpredictability during power grid backup, and used column constraint generation algorithm to solve it. This model can effectively solve the problem of unpredictability during power grid backup, and has obvious advantages in random scenarios [11]. Wang S et al. designed a multi-agent power grid control scheme to meet the requirements of the system and computing platform for autonomous control of the power grid. This scheme is data-driven and adopts deep reinforcement learning, which can be learned from scratch. This method can meet the requirements of power grid autonomous control for systems and computing platforms [12]. Shaheen A M and other experts have designed a multi-objective manta ray foraging algorithm to minimize emissions from DC AC hybrid power grids. This algorithm imitates the feeding process of manta rays and adopts fuzzy decision-making technology. It has been compared and tested on multiple systems, and the results show that the robustness of this method is significantly better than other comparison algorithms [13].

In summary, there is currently a wealth of research on economic dispatch in the power system, and the methods used are also diverse. However, these studies also have certain shortcomings, such as the scheduling problems involved being relatively single and the constraints considered being incomplete. Based on these issues, this study innovatively proposes the use of CMOPSO algorithm to solve multi regional environmental and economic scheduling problems, and uses ImCSO algorithm to solve multi regional S-DE scheduling problems.

3. Design of Economic Dispatching Method for MRPS Based on CMOPSO and ImCSO Algorithms. This study uses an improved competitive group optimization algorithm to solve the S-DE dispatch problems of multi region power systems. It optimized the competitive group optimization algorithm through ranking pairing learning and differential evolution, and designed specific steps for the algorithm in multi region S-DE scheduling problems. A MOPSO based on competitive learning was used to solve the problem of multi regional environmental and economic scheduling, and specific steps were designed for this algorithm to solve the problem.

3.1. Design of S-DE scheduling methods for multiple regions based on ImCSO algorithm. The economic dispatch of multi-regional power systems (MRPS) is mainly divided into Multi Area Static Economic Dispatch (MASED), Multi Area Dynamic Economic Dispatch (MADED), and Multi Area Environment Economic Dispatch (MAEED) [14]. The core goal of MASED is to minimize the combustion cost of the power system. To solve the MASED problem, an ImCSO algorithm was adopted for improvement through ranking pairing learning and differential evolution. The learning process of ranking paired learning is shown in Fig 3.1 [15].

In Fig. 3.1, ranking pairing learning requires sorting all particles first. The sorting is based on the fitness information of the particles, and then the particles are divided into winner and loser groups according to the sorting results. The particles of the loser group need to learn from the particles of the winner group. The

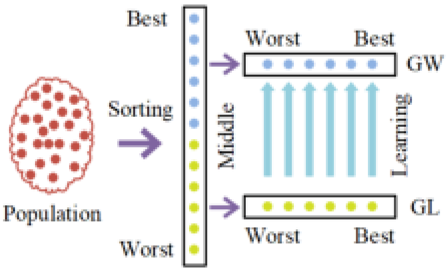


Fig. 3.1: The learning process of ranking paired learning

operation of particle sorting and grouping is equation (3.1).

$$\begin{cases} f\left(X_1^{t,W}\right) \leq f\left(X_2^{t,W}\right) \leq \cdots \leq f\left(X_{ps/2}^{t,W}\right) \leq f\left(X_1^{t,L}\right) \leq f\left(X_2^{t,L}\right) \leq \cdots \leq f\left(X_{ps/2}^{t,L}\right) \\ GW = \left\{X_1^{t,W}, X_2^{t,W}, \cdots, X_{ps/2}^{t,W}\right\} \\ GL = \left\{X_1^{t,L}, X_2^{t,L}, \cdots, X_{ps/2}^{t,L}\right\} \end{cases} \tag{3.1}$$

In equation (3.1), t means the iteration numbers. W are the particles in the winner group population. L is the particle in the loser group population. GW represents the winner group. GL represents the loser group. $X_i^t = \left[X_{i,1}^t, X_{i,1}^t, \cdots, X_{i,D}^t\right]$ represents the position item. Where i is the serial number. D is the dimension of the optimization problem. ps represents the population size. The updates of the velocity and position terms of particles in GL are shown in equation (3.2).

$$\begin{cases} V_i^{t+1,L} = R_1 \times V_i^{t,L} + R_2 \times \left(X_i^{t,W} - X_i^{t,L}\right) + R_3 \times \varphi \times \left(X_{i,center}^t - X_i^{t,L}\right) \\ X_i^{t+1,L} = X_i^{t,L} + V_i^{t+1,L} \end{cases} \tag{3.2}$$

In equation (3.2), $V_i^t = \left[v_{i,1}^t, v_{i,2}^t, \cdots, v_{i,D}^t\right]$ represents the velocity term. R_1, R_2 and R_3 represent a set of random numbers with a distribution range of $[0,1]$. φ is a social factor that can control $\bar{X}^t$. $\bar{X}^t$ represents the average position of the entire population in the t-tph iteration. $X_{i,center}^t$ is the center position, and its calculation is equation (3.3).

$$X_{i,center}^t = a \times \bar{X}_{GW}^t + (1-a) \times \bar{X}_{GL}^t \tag{3.3}$$

In equation (3.3), a represents a random real number within $[0,1]$. $\bar{X}_{GW}^t$ represents the average position of particles in the GW population. $\bar{X}_{GL}^t$ is the average position of particles in the GL. The update of winner particles in the GW is mainly achieved through differential evolution strategy (DES). The specific update steps include mutation, crossover, and selection. The mutation step requires the generation of mutated individuals, and the specific process is equation (3.4).

$$Z_i^{t,W} = X_{r1}^{t,W} + F \times \left(X_{r2}^{t,W} - X_{r3}^{t,W}\right) \tag{3.4}$$

In equation (3.4), $r1$, $r2$, and $r3$ are random positive numbers that are different from each other, with a value range of $\{1, 2, \cdots, ps\}$. $Z_i^{t,W}$ represents the mutant individual. F represents the variation factor. The purpose of the cross step is to generate experimental individuals, and the specific operation is equation (3.5).

$$u_{i,j}^{t,W} = \begin{cases} Z_{i,j}^{t,W}, & if \quad rand_j \leq CR \quad or \quad j = j_{rand} \\ X_{i,j}^{t,W}, & otherwise \end{cases} \tag{3.5}$$

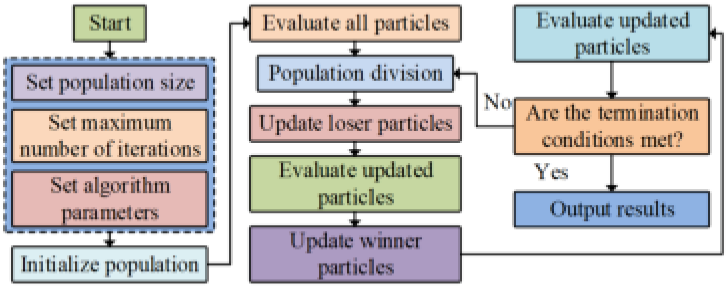


Fig. 3.2: The process of ImCSO algorithm

In equation (3.5), $U_i^{t,W}$ represents the experimental individual, and its value range is $\left[u_{i,1}^{t,W}, u_{i,2}^{t,W}, \cdots, u_{i,D}^{t,W}\right]$. CR represents the crossover factor. $rand_j$ is the dimension, with a value range of $[1, D]$. $rand_j$ represents a random real number within the range of $[0, 1]$. j_{rand} is a random integer in $[1, D]$. The selection step is to choose individuals between $X_i^{t,W}$ and $U_i^{t,W}$ to enter the next generation, grounded on fitness. The specific process of this step is equation (3.6).

$$X_{i,j}^{t+1,W} = \begin{cases} U_i^{t,W}, & if \quad f\left(U_i^{t,W}\right) < f\left(X_i^{t,W}\right) \\ X_i^{t,W}, & otherwise \end{cases} \tag{3.6}$$

The improvement of competitive swarm optimization algorithm can be achieved through ranking pairing learning and differential evolution. Fig 3.2 is the process of the ImCSO.

In Fig. 3.2 the first step of the ImCSO algorithm is to set the population size, maximum number of iterations, and algorithm parameters. The second is to initialize the population. The third is to evaluate all particles and record the globally optimal particle as *Gbest*. The fourth step is to divide the population into GW and GL based on fitness information. The fifth step is to update the particles in GL through a pairing learning strategy. The sixth step is to evaluate the updated particles and update *Gbest*. The seventh step is to use DES to update the particles in GW. The eighth step is to evaluate the updated particles and update *Gbest*. The ninth step is to determine whether the termination condition is met. If so, output the result, otherwise return to the fourth step. When using ImCSO to solve MASED problems, the position of each particle in the population corresponds to an effective solution. The position information X_i of the effective solution of the MASED system composed of M regions is equation (3.7).

$$X\_i = \underbrace{[P_{11}, P_{12}, \cdots, P_{1N_1}}_{power \quad in \quad area \quad 1}, \underbrace{P_{21}, P_{22}, \cdots, P_{2N_2}}_{power \quad in \quad area \quad 2}, \cdots, \underbrace{P_{M1}, P_{M2}, \cdots, P_{MN_M}}_{power \quad in \quad area \quad M}, \underbrace{T_{12}, T_{13}, \cdots, T_{(M-1)M}]}_{power \quad transmission} \tag{3.7}$$

In equation (3.7), P represents the output power (OP) of the generator set. T is the transmission power between regions. N and Mrepresent the number of generators and regions. The most crucial aspect in solving MASED problems is inequality and equality constraints. For the constraint of power generation capacity, the repair of N_{wq} OP is equation (3.8).

$$P_{wq} = \begin{cases} P_{wq}^{\min}, & if \quad P_{wq} \le P_{wq}^{\min} \\ P_{wq}^{\max}, & if \quad P_{wq} \ge P_{wq}^{\max} \\ P_{wq}, & otherwise \end{cases} \tag{3.8}$$

In equation (3.8), N_{wq} represents the q-the generator unit in the w-the region. P_{wq} represents the actual OP of the q-the generator unit in the w-the region. $P_{wq}^{\min}$ represents the minimum OP of the unit. $P_{wq}^{\max}$ is the maximum OP of the unit. Regarding the transmission capacity constraints of the interconnection line, the

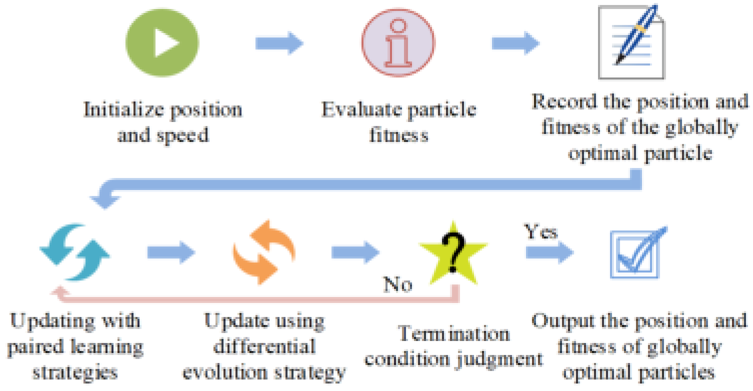

Initialize position and speed
Evaluate particle fitness
Record the position and fitness of the globally optimal particle
Updating with paired learning strategies
Update using differential evolution strategy
No
Yes
Termination condition judgment
Output the position and fitness of globally optimal particles

Fig. 3.3: Specific steps for solving MASED problems using ImCSO

transmission power repair between regions is equation (3.9).

$$T_{wk}=\begin{cases} T_{wk}^{\min}, & if \quad T_{wk}\le T_{wk}^{\min} \\ T_{wk}^{\max}, & if \quad T_{wk}\ge T_{wk}^{\max} \\ T_{wk}, & otherwise \end{cases} \tag{3.9}$$

In equation (3.9), T_{wk} represents the transmission power from region w to region k. $T_{wk}^{\min}$ is the minimum value. $T_{wk}^{\max}$ is the maximum value. For the constraint of prohibited operation zone, if $P_{wq,m}^{l}<P_{wq}<P_{wq,m}^{u}$, its repair is equation (3.10).

$$P_{wq}=\begin{cases} P_{wq,m}^{l}, & if \quad P_{wq,m}^{l}<P_{wq}\le\left(P_{wq,m}^{l}+P_{wq,m}^{u}\right)/2 \\ P_{wq,m}^{u}, & if \quad \left(P_{wq,m}^{l}+P_{wq,m}^{u}\right)/2<P_{wq}<P_{wq,m}^{u} \end{cases} \tag{3.10}$$

In equation (3.10), m represents the m-the prohibited operation zone. $P_{wq,m}^{u}$ represents the upper boundary of the m of the q-the generator unit in the w-the region. $P_{wq,m}^{l}$ represents the lower boundary. For the actual power balance constraints in the region, it is necessary to use a power balance repair operator. The repair of this constraint requires calculating the violation degree dif_w of the w-the region, as shown in equation (3.11).

$$dif_w=PG_w-\left(PD_w+PL_w+\sum_{k=1.k\neq w}^{M}T_{wk}\right) \tag{3.11}$$

In equation (3.10), PG_w represents the total electricity generation of region w. PG_w represents the load demand of region w. PG_w is the network loss of region w. If $dif_w>0$, select a generator set in region w to OP $P_{wq}=\max\left(P_{wq}-dif_w,P_{wq}^{\min}\right)$. If $dif_w<0$, then $P_{wq}=\min\left(P_{wq}+dif_w,P_{wq}^{\max}\right)$. Due to the fact that the equality constraints of the entire MASED system after repair are not fully satisfied, the study further considers the penalty function as the objective function to solve fitness, as shown in equation (3.12).

$$Fit\left(X_w\right)=FC\left(X_w\right)+factor\left(V_1+V_2+\cdots+V_M\right) \tag{3.12}$$

In equation (3.12), $FC\left(X_w\right)$ represents the total fuel cost. $factor$ represents the penalty factor. V_w is the degree of power balance constraint violation in repaired region w. The specific steps for solving the MASED problem using ImCSO are shown in Fig. 3.3.

In Fig. 3.3, the first step in solving the MASED using ImCSO is to initialize the positions and velocities of all particles in the population. The second step is to evaluate the fitness of all particles. The third step is to record the position of the globally optimal particle and its corresponding fitness information. The fourth step is to update GL with a paired learning strategy. The fifth step is to update GW using DES. The sixth step is to determine the termination conditions. If so, proceed to the next step, otherwise go back to step four. The seventh step is to output the position of the globally optimal particle and its corresponding fitness information. The research also adopts the ImCSO algorithm for solving the MADED problem, and the specific solving steps are consistent with the MASED problem.

3.2. Design of Multi regional Environmental and Economic Scheduling Method Based on CMOPSO Algorithm.. MAEED is a multi-objective optimization (MOO) that requires ensuring the comprehensive optimization of power generation costs and pollution emissions in the power system. This study first explains the solution methods for MOO, and then designs the specific steps for solving MAEED problems using the CMOPSO. MOO does not have a unique optimal solution, and can only achieve a relatively optimal overall goal. MOO is mainly solved through the concept of Pareto optimality, involving Pareto dominance, Pareto frontiers, etc. [16, 17]. The effectiveness of different algorithms in solving MOO problems varies. Therefore, this study used distribution uniformity index and comprehensive performance index inverse generation distance to evaluate the performance of the algorithm in solving MOO problems. The expression of the distribution uniformity index is shown in equation (3.13) [18].

$$\Delta\left(B,S\right)=\frac{\sum_{w=1}^{\alpha}d\left(E_w,B\right)+\sum_{X\in B}\left|d\left(X,B\right)-\bar{d}\right|}{\sum_{w=1}^{\alpha}d\left(E_w,B\right)+\left|S\right|\bar{d}} \tag{3.13}$$

In equation (3.13), S represents the set of points uniformly distributed on the leading edge of the real Pareto. B represents the Pareto optimal solution set. E_w is the extreme solution. Where α is the sequence number of the extreme solution, with a value range of $[1-\phi]$. ϕ represents the number of targets. X represents the solution, and d is the calculation of the minimum Euclidean distance. $\bar{d}$ represents the average value of the min Euclidean distance. $|S|$ is the number of concentration points. The calculation of the inverse generation distance of the comprehensive performance index is shown in equation (3.14) [19].

$$IGD\left(B,S\right)=\frac{\sum_{\delta\in S}d\left(\delta,B\right)}{|S|} \tag{3.14}$$

In equation (3.14), δ represents the point on the true frontier. The focus of MOO is to handle multiple constraint conditions, and common constraint processing methods can easily lead to local optimization problems in the algorithm. To avoid this issue, the study adopted the ε constraint criterion and the multi archive set method to handle the multi constraint problem of MOO. The expression of ε in the ε constraint criterion is equation (3.15).

$$\varepsilon\left(Gen\right)=\varepsilon\left(0\right)\times\left(1-\frac{Gen}{\max Gen}\right)^{cp} \tag{3.15}$$

In equation (3.15), ε is the variable value. $\varepsilon\left(0\right)$ represents the initial threshold. Gen represents the current iterations. cp is an index with a value of 2. $\max Gen$ represents the maximum iterations of the population. The core idea of multi archive constraint processing is classification, which categorizes solutions from different categories into different sets. The steps of processing MOO by combining the ε-constraint criterion and the multi archive set method are divided into three steps. The first step is to partition the solution with the support of the ε constraint. The second is to sort the infeasible solutions. The third is to determine whether the particles in the feasible solution meet the size of the population. MOPSO has the advantages of simple structure and fewer parameters that need to be adjusted, and is often used for engineering optimization problems. The process of this algorithm is Fig 3.4.

In Fig. 3.4, the first step of the MOPSO algorithm is to initialize the population, calculate the target vector, and liberate non dominated data into external files. The second step is to update the $Gbest$ and individual optimal value $Pbest$ in the population. The third step is to update the velocity and position information of the particles, calculate the target vector, and then update $Pbest$. The fourth step is to select $Gbest$. The fifth step is to determine the termination condition. If it is determined to be yes, output the external file

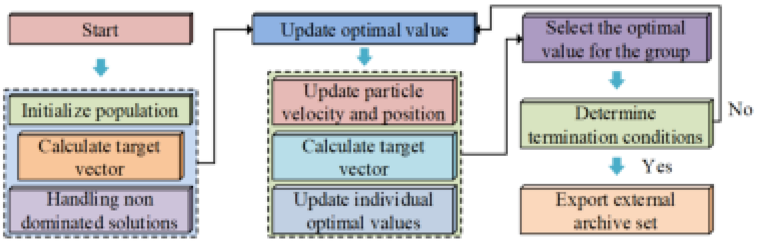

Start
Initialize population
Calculate target vector
Handling non dominated solutions
Update optimal value
Update particle velocity and position
Calculate target vector
Update individual optimal values
Select the optimal value for the group
Determine termination conditions
No
Yes
Export external archive set

Fig. 3.4: Process of MOPSO

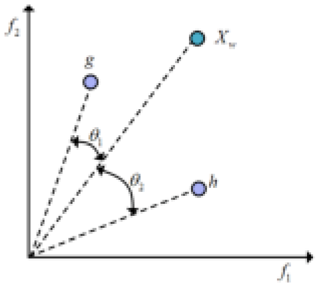


Fig. 3.5: The competitive strategy of CMOPSO algorithm

set; otherwise, return to the second step. However, the MOPSO algorithm did not solve the balance problem between population convergence and diversity, so the study introduced a competition mechanism and formed the CMOPSO algorithm. The competition mechanism mainly involves competitive learning strategies. The core idea is to let elite particles compete and then let the winning particles guide the update of the current particles. The competitive learning mechanism includes three aspects. The first aspect is elite particle selection, which is studied using the non dominated sorting genetic algorithm II. The second aspect is spatial competition, and the competition strategy of the CMOPSO algorithm is Fig 3.5.

In Fig. 3.5, both g and h are particles. θ_1 and θ_2 are both the angles between particles and X_w. If $\theta_1 > \theta_2$, h wins and is recorded as X_W. The third aspect is learning strategies. The effective solution position information expression when using the CMOPSO algorithm to solve the MAEED problem is consistent with MASED, and the repair method of constraint conditions is also the same as MASED. The process of using the CMOPSO algorithm to solve the MAEED problem is Fig 3.6.

In Fig. 3.6, the first step of the CMOPSO algorithm in solving the MAEED problem is to initialize the positions and velocities of all particles in the population. The second step is to fix the constraint violation of particles, and then calculate the target values of fuel cost and pollution emissions. The third step is to perform non dominated sorting and crowding distance sorting, and then determine the elite population. The fourth step is to select the winning particles. The fifth step is to update the velocity and position information of particles. The sixth step is to mutate and repair the particles, and calculate their objective function values. The seventh step is to determine the termination condition. If yes, output Pareto frontier particle information; otherwise, return to the third step.

4. Analysis of Economic Dispatching Results for MRPS Based on CMOPSO and ImCSO Algorithms. This study sets the simulation environment and running times for the ImCSO algorithm and CMOPSO algorithm, and verifies their effectiveness through algorithm comparison. The comparison indicators of the algorithm include fuel cost, cost convergence curve, Pareto frontier, distribution uniformity index, and comprehensive performance index inverse generation distance.

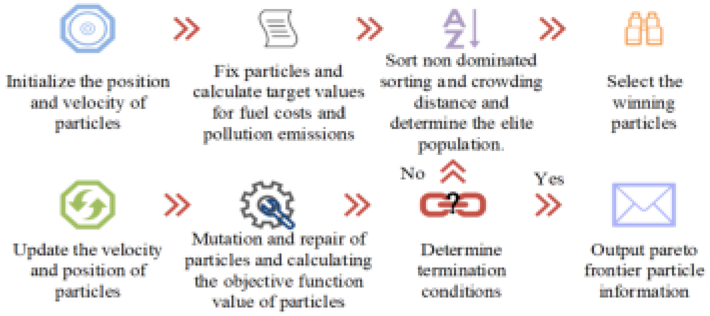

Initialize the position and velocity of particles
Fix particles and calculate target values for fuel costs and pollution emissions
Sort non dominated sorting and crowding distance and determine the elite population.
Select the winning particles
No
Yes
Update the velocity and position of particles
Mutation and repair of particles and calculating the objective function value of particles
Determine termination conditions
Output pareto frontier particle information

Fig. 3.6: The process of using CMOPSO algorithm to solve MAEED problems

Table 4.1: Comparison of fuel costs with different algorithms under MASED test system 1

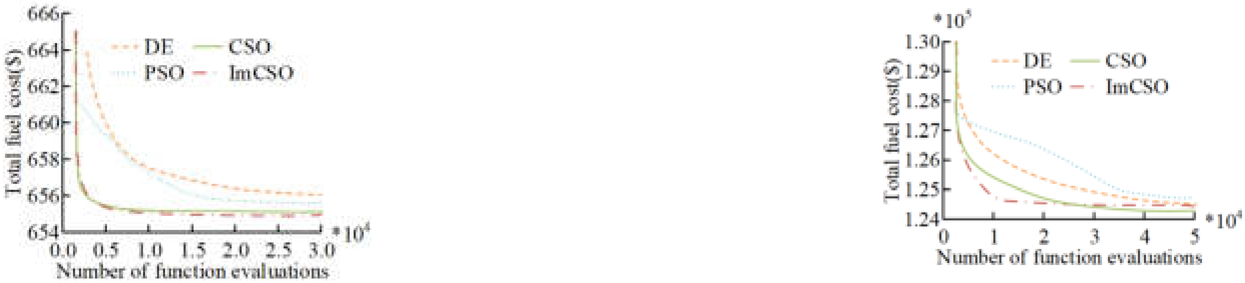
Algorithm	Minimum value($/h)	Mean value($/h)	Maximum value($/h)	Standard deviation	Time(s)
DE	656.2642	657.0810	658.5198	5.7913	1.43
PSO	655.9438	656.6521	661.2338	11.3302	1.64
CSO	655.8728	656.1905	657.0903	3.7779	1.58
ImCSO	655.8592	655.9866	656.2243	1.9642	1.60

4.1. Analysis of S-DE dispatch results in multiple regions based on ImCSO algorithm. In order to conduct simulation analysis on the ImCSO algorithm, two MASED testing systems and two MADED testing systems were selected for the study. A comparative analysis was conducted on the effectiveness verification of the ImCSO algorithm. Comparative algorithms include Competitive Swarm Optimizer (CSO), PSO, and Differential Evolution (DE). The simulation environment for comparing algorithms is MATLAB 9.6, with 10 runs. Table 1 shows the fuel cost comparison of different algorithms under MASED test system 1.

In Table 4.1, the maximum, minimum, average, and standard deviation of the fuel cost of the ImCSO algorithm are lower than those of other comparative algorithms, with values of 656.2243 $/h, 655.8592 $/h, 655.9866 $/h, and 1.9642, respectively. The running time of the ImCSO algorithm is 1.6 seconds, which is not significantly different from other comparative algorithms. The fuel cost values of the DE algorithm are 658.5198 $/h, 656.2642 $/h, 657.0810 $/h, and 5.7913, respectively, with a running time of 1.43 seconds. The four values of fuel cost for the PSO algorithm are 661.2338 $/h, 655.9438 $/h, 656.6521 $/h, and 11.3302, respectively, with a running time of 1.64 seconds. The relevant values for the fuel cost of the CSO algorithm are 657.0903 $/h, 655.8728 $/h, 656.1905 $/h, and 3.7779, respectively, with a running time of 1.58 seconds. Therefore, the performance of the ImCSO algorithm is better and more stable. The comparison of cost convergence curves of different algorithms under different MASED testing systems is Fig 4.2.

As Fig. 4.1a, with the increase of iterations, the total fuel cost of different algorithms gradually decreases. In Test System 1, the ImCSO algorithm flattened after nearly 4000 iterations, while the DE, PSO, and CSO algorithms stabilized after nearly 15000, 14000, and 5000 iterations, respectively. In test system 2 of Figure 4.1b, the ImCSO algorithm tends to flatten out after nearly 10000 iterations, while the DE, PSO, and CSO algorithms tend to flatten out after 30000, 32000, and 25000 iterations, respectively. From this, the ImCSO algorithm converges faster and has better performance. Table 2 is the fuel cost comparison of different algorithms under MADED test system 1.

In Table 4.2, the minimum fuel cost values for ImCSO, DE, PSO, and CSO algorithms are 13003.9526 $/h, 13166.7657 $/h, 13492.8771 $/h, and 13476.6407 $/h, respectively. The average fuel costs of the four algorithms are 13151.3299 $/h, 13291.1036 $/h, 14167.4435 $/h, and 16795.4673 $/h. The maximum fuel costs are 13299.2825 $/h, 13434.3578 $/h, 17826.6214 $/h, and 34830.2087 $/h. The standard fuel cost values are 78.1757, 69.2933762.1120, and 4659.4287. The running time of each algorithm is 81.427s, 179.1231s, 127.6157s,



(a) Cost convergence curve of MASED test system 1

(b) Cost convergence curve of MASED test system 2

Fig. 4.2: Comparison of cost convergence curves of different algorithms in different MASED test systems

Table 4.2: Comparison of fuel cost of different algorithms under MADED test system 1

Algorithm	Minimum value($/h)	Mean value($/h)	Maximum value($/h)	Standard deviation	Time(s)
DE	13166.7657	13291.1036	13434.3578	69.2933	179.1231
PSO	13492.8771	14167.4435	17826.6214	762.1120	127.6157
CSO	13476.6407	16795.4673	34830.2087	4659.4287	116.0529
ImCSO	13003.9526	13151.3299	13299.2825	78.1757	81.427

and 116.0529s respectively. Therefore, it can be concluded that the ImCSO algorithm has the smallest running time, and the maximum, minimum, and average fuel costs are lower than other algorithms. And this also indicates that the ImCSO algorithm performs better in solving MADED problems. In order to further validate the performance of the ImCSO algorithm, other algorithms were selected for comparison in the study. The selected algorithms for the study include Gravitational Search Algorithm (GSA), Gbest guided Artificial Bee Colony (GABC), Teaching Learning Based Optimization (TLBO), and Differential Evolution Algorithm with Strategy Adaptation (DESA). The F1 values and CPU utilization of different algorithms are compared in Table 4.3.

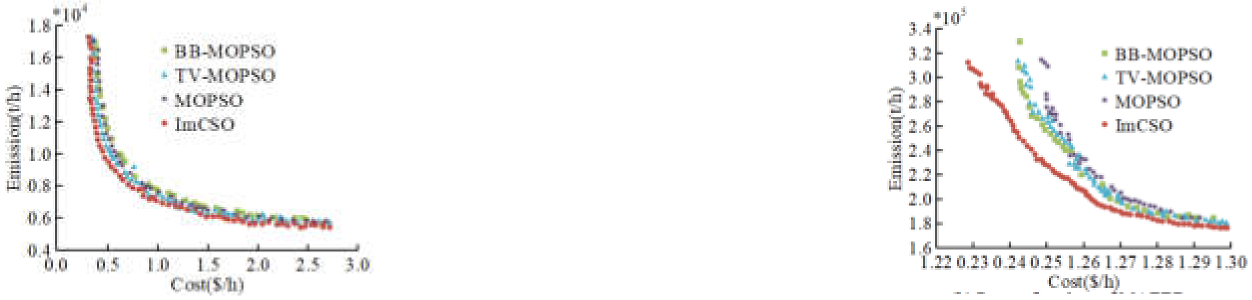
From Table 4.3, it can be seen that in terms of CPU utilization, the maximum value of the ImCSO algorithm is 17.5%, and the minimum value is 16.1%. The maximum values of DE, PSO, CSO, GSA, GABC, TLBO, and DESA algorithms are 28.2%, 22.8%, 21.5%, 20.8%, 24.8%, 23.6%, and 27.1%, respectively, while the minimum values are 26.7%, 21.3%, 19.2%, 18.5%, 23.6%, 22.1%, and 25.5%, respectively. On the F1 value, the maximum value of the ImCSO algorithm is 0.993 and the minimum value is 0.976. The maximum values of DE, PSO, CSO, GSA, GABC, TLBO, and DESA algorithms are 0.837, 0.938, 0.953, 0.966, 0.888, 0.924, and 0.861, respectively, while the minimum values are 0.811, 0.921, 0.938, 0.953, 0.857, 0.902, and 0.834, respectively. It can be seen that the ImCSO algorithm has advantages in CPU utilization and F1 value, indicating better performance of the algorithm.

4.2. Analysis of multi-regional environmental and economic dispatch results based on CMOPSO algorithm.. Two MAEED testing systems were also selected for the simulation analysis of the CMOPSO algorithm. To verify the effectiveness of the CMOPSO algorithm, a comparative analysis was conducted. Comparison algorithms include MOPSO, BB-MOPSO [20], and TV-MOPSO [12]. The comparison content includes Pareto Frontier, Distribution Uniformity Index, and Comprehensive Performance Index Reverse Generation Distance. The simulation environment for comparing algorithms is MATLAB 9.6, with 10 runs. The Pareto frontier comparison of different algorithms under different MAEED testing systems is shown in Fig 4.4.

From 4.3a, different algorithms have more repetitions in obtaining Pareto frontiers, but the CMOPSO

Table 4.3: Comparison of F1 values and CPU utilization of different algorithms

Algorithm	CPU utilization					F1				
	Number of experiments					Number of experiments				
	1	2	3	4	5	1	2	3	4	5
DE	26.7%	27.3%	26.8%	27.4%	28.2%	0.815	0.837	0.826	0.811	0.832
PSO	21.7%	21.3%	22.6%	22.8%	21.5%	0.925	0.934	0.938	0.921	0.933
CSO	20.8%	19.5%	21.5%	20.2%	19.2%	0.938	0.949	0.953	0.952	0.945
GSA	19.3%	20.8%	19.1%	18.5%	19.4%	0.957	0.966	0.955	0.961	0.953
GABC	23.7%	24.3%	24.5%	23.6%	24.8%	0.871	0.857	0.866	0.888	0.865
TLBO	22.5%	23.2%	23.6%	22.1%	23.4%	0.902	0.913	0.908	0.924	0.917
DESA	25.9%	26.7%	25.5%	27.1%	26.9%	0.855	0.842	0.857	0.861	0.834
ImCSO	17.2%	16.8%	17.5%	16.1%	17.1%	0.977	0.983	0.989	0.976	0.993



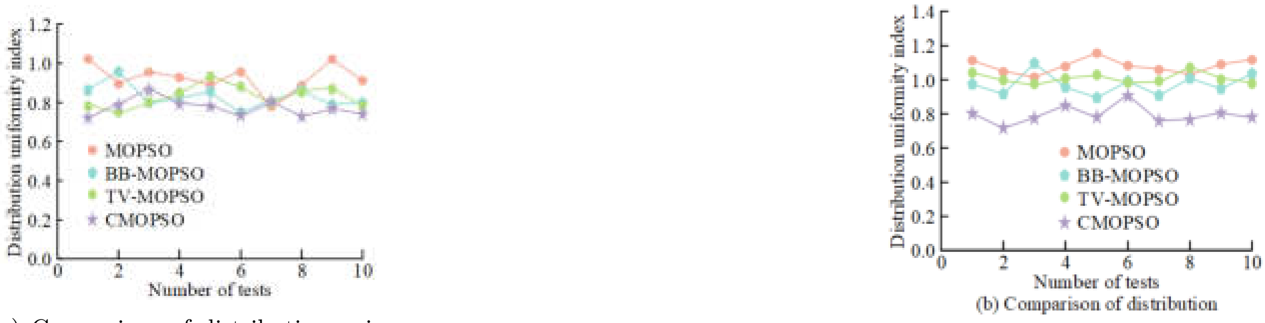
(a) Pareto frontiers of MAEED test system 1

(b) Pareto frontiers of MAEED test system 2

Fig. 4.4: Pareto frontier comparison of different algorithms in different MAEED test systems

algorithm has a wider distribution of Pareto frontiers than other algorithms, and the extreme solutions are also better than other algorithms. In Figure 4.3b, there is a significant difference in the Pareto frontier obtained by different algorithms. The distribution of Pareto frontiers obtained by the CMOPSO algorithm is significantly higher than other algorithms, and also higher. From this, the CMOPSO algorithm performs better and achieves better extreme solutions. The comparison of distribution uniformity indicators of different algorithms under different MAEED testing systems is Fig 4.6.

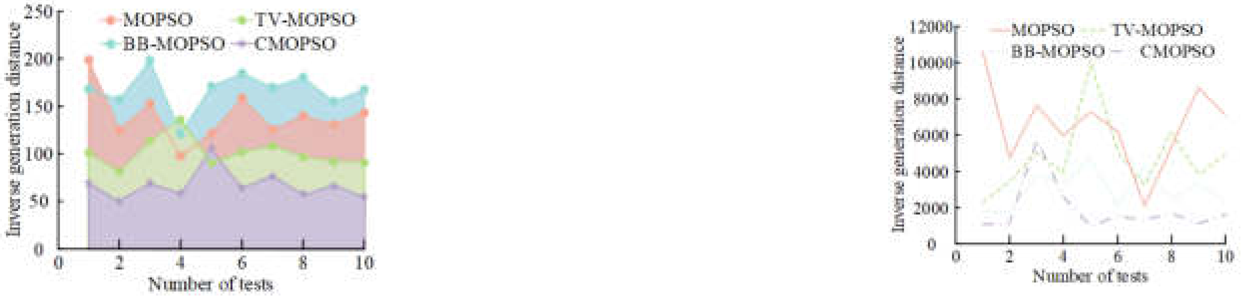
In Fig. 4.5a in MAEED test system 1, the maximum value of the MOPSO algorithm's distribution uniformity index is 1.0578, the min value is 0.7472, and the average value is 0.8669. The max values of the distribution uniformity indicators for the BB-MOPSO algorithm and TV-MOPSO algorithm are 0.9873 and 0.9625, respectively, and the min values are 0.7152 and 0.7187. The average values are 0.8358 and 0.8575. The max value of the distribution uniformity index of the CMOPSO algorithm is 0.9051, the minimum value is 0.6919, and the average value is 0.8058. According to 4.5b, in MAEED test system 2, the maximum value of the MOPSO algorithm's distribution uniformity index is 1.1872, the minimum value is 0.9815, and the average value is 1.0498. The maximum values of the distribution uniformity indicators for the BB-MOPSO algorithm and TV-MOPSO algorithm are 1.1308 and 1.1032, and the min values are 0.8617 and 0.9353. The average values are 0.9937 and 1.0149, respectively. The max value of the distribution uniformity index of the CMOPSO algorithm is 0.9605, the min value is 0.68, and the average value is 0.8457. From this, the distribution uniformity index of the CMOPSO algorithm is superior to the comparison algorithm, which also indicates that the algorithm has superiority in MAEED problems. The comparison of the comprehensive performance indicators of different algorithms under different MAEED testing systems in terms of inverse generation distance is shown in Fig 4.8.



(a) Comparison of distribution uniformity indicators of different algorithms in MAEED test system 1

(b) Comparison of distribution uniformity indicators of different algorithms in MAEED test system 2

Fig. 4.6: Comparison of distribution uniformity indexes of different algorithms in different MAEED test systems

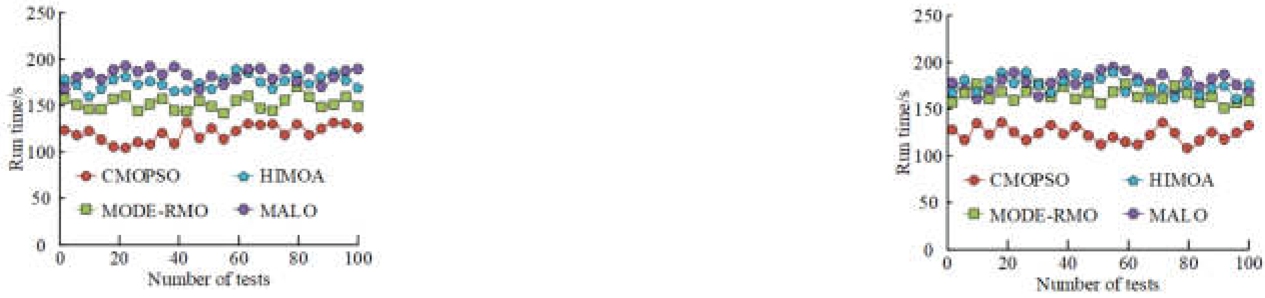


(a) Comparison of inverse generation distance between different algorithms in MAEED testing system 1

(b) Comparison of inverse generation distance between different algorithms in MAEED testing system 2

Fig. 4.8: Comparison of the comprehensive performance index of different algorithms in different MAEED test systems

From Fig. 4.7a, in MAEED test system 1, the maximum anti generation distances of the MOPSO algorithm, BB-MOPSO algorithm, and TV-MOPSO algorithm are 204.3648, 204.6281, and 143.2398, respectively. The minimum values are 93.1586, 116.6733, and 76.4959, respectively, and the average values are 123.6742, 155.883, and 95.2604, respectively. The maximum anti generation distance of the CMOPSO algorithm is 112.6341, the min value is 43.9868, and the average value is 67.6316. From Figure 4.7b, in MAEED test system 2, the maximum inverse generation distances of the three comparison algorithms are 10879.9507, 4959.483, and 10117. 1681, The minimum values are 1769.6697, 1473.3904, and 2036.2403, respectively; The average values are 6675.7363, 2898.1252, and 4685.6277, respectively. The maximum anti generation distance of the CMOPSO algorithm is 6121.3624, the minimum value is 702.5518, and the average value is 1664.0978. From this, the CMOPSO algorithm has the best comprehensive performance. In order to further verify the performance of the CMOPSO algorithm, other algorithms were selected for comparison in the study. Comparison algorithms include Multi Objective Differential Evolution with Ranking based Mutation Operator (MODE-RMO), Hybrid Immune Multi Objective Optimization Algorithm (HIMOA), and Multi Objective Ant Lion Optimization Algorithm (MALO). The experiment was conducted a total of 100 times. The comparison of runtime of different algorithms under different MAEED testing systems is shown in Figure 4.10.



(a) Comparison of runtime of different algorithms in the MAEED1 testing system

(b) Comparison of runtime of different algorithms in the MAEED2 testing system

Fig. 4.10: Comparison of runtime of different algorithms in different MAEED testing systems

From Figure 4.9a, it can be seen that in MAEED test system 1, the maximum running time of the CMOPSO algorithm is 127.8 seconds, and the minimum value is 101.3 seconds. The maximum running times of MODE-RMO, HIMOA, and MALO algorithms are 169.8 seconds, 180.2 seconds, and 188.9 seconds, respectively, while the minimum values are 147.4 seconds, 158.9 seconds, and 163.3 seconds, respectively. As shown in Figure 4.9b, in MAEED test system 2, the maximum running time of the CMOPSO algorithm is 139.6 seconds, and the minimum value is 118.4 seconds. The maximum running times of MODE-RMO, HIMOA, and MALO algorithms are 172.9s, 183.6s, and 190.8s, respectively, while the minimum values are 153.4s, 161.5s, and 165.2s, respectively. From this, it can be seen that the CMOPSO algorithm has obvious advantages in runtime and better performance.

5. Conclusion. In response to the improvement of energy utilization efficiency, this study innovatively proposes the use of CMOPSO algorithm to solve multi regional environmental and economic scheduling problems. Moreover, the ImCSO algorithm is adopted to handle S-DE scheduling matters in multiple regions. Research showed that the maximum, minimum, and average fuel costs of the ImCSO for solving MASED problems were 656.2243 \$/h, 655.8592 \$/h, and 655.9866 \$/h, respectively. The maximum, minimum, and average fuel costs of the ImCSO algorithm for solving MADED problems were 13299.2825 \$/h, 13003.9526 \$/h, and 13151.3299 \$/h, respectively. All values were smaller than other comparison algorithms. From this, the ImCSO algorithm performed better in solving MASED and MADED problems. The CMOPSO algorithm have a wider Pareto frontier distribution when solving MAEED problems. Under different testing systems, the maximum values of the distribution uniformity index of the CMOPSO algorithm were 0.9051 and 0.9605, the minimum values were 0.6919 and 0.68, and the average values were 0.8058 and 0.8457, respectively. Under different testing systems, the maximum anti generation distance of the CMOPSO algorithm was 112.6341 and 6121.3624, while the minimum value was 43.9868 and 702.5518, respectively. The average value was 67.6316 and 1664.0978. From this, the performance of the CMOPSO algorithm is superior to that of the comparison algorithm. However, there are also certain shortcomings in the research, which only considers multi-objective issues of the environment and economy, and does not involve many other factors, which is also an area for further research to improve.

REFERENCES

[1] Chen, S., Zhang, L., Yan, Z. & And, S. and robust security-constrained economic dispatch algorithm based on blockchain. *IEEE Transactions On Power Systems.* **37**, 691-700 (2021)

[2] Wen, G., Yu, X. & Liu, Z. Recent progress on the study of distributed economic dispatch in smart grid: an overview. *Frontiers Of Information Technology & Electronic Engineering.* **22**, 25-39 (2021)

[3] Huang, B., Li, Y., Zhan, F., Sun, Q. & Zhang, H. distributed robust economic dispatch strategy for integrated energy system considering cyber-attacks. *IEEE Transactions On Industrial Informatics.* **18**, 880-890 (2021)

[4] Younes, Z., Anharmonic, I., Makhila, S. & Reya Sudin, M. memory-based gravitational search algorithm for solving economic dispatch problem in micro-grid. *Ain Shams Engineering Journal.* **12**, 1985-1994 (2021)

[5] Lin, L., Guan, X., Peng, Y., Wang, N., Maharjan, S. & Ootsuka, T. Deep reinforcement learning for economic dispatch of virtual power plant in internet of energy. *IEEE Internet Of Things Journal.* **7**, 6288-6301 (2020)

[6] Goudarzi, A., Fahad, S., Ni, J., Ghayoor, F., Siano, P. & Achelous, H. sequential hybridization of ETLBO and IPSO for solving reserve-constrained combined heat, power and economic dispatch problem. *IET Generation, Transmission & Distribution.* **16**, 1930-1949 (2022)

[7] Lyu, C., Jia, Y. & Xu, Z. DRO-MPC-based data-driven approach to real-time economic dispatch for islanded microgrids. *IET Generation Transmission & Distribution.* **14**, 5704-5711 (2020)

[8] Wang, X., Liu, R., Wang, X., Hou, Y. & Bouffard, F. Data-Driven Uncertainty Quantification Method for Stochastic Economic Dispatch. *IEEE Transactions On Power Systems.* **37**, 812-815 (2022)

[9] Marco, A., Cecilia, M. & Guadalupe, C. power optimization model for the long-term planning scenarios: Case study of Mexico's power system decarbonization. *The Canadian Journal Of Chemical Engineering.* **99**, 884-897 (2020)

[10] Lev, K., Tang, H., Akenson, B., Pillai, J., Tan, Q. & Zhang, Q. Hierarchical learning optimization method for the coordination dispatch of the inter-regional power grid considering the quality-of-service index. *IET Generation, Transmission & Distribution.* **14**, 3673-3684 (2020)

[11] Xu, D., Liang, F., Cheng, Q., Zhou, J. & Tang, J. Research on joint optimization model and algorithm of multi-area generation and reserve with considering its availability. *IET Generation, Transmission & Distribution.* **16**, 2032-2048 (2022)

[12] Wang, S., Duan, J., Shi, D., Xu, C. & Wang, Z. Data-driven Multi-agent Autonomous Voltage Control Framework Using Deep Reinforcement Learning. *IEEE Transactions On Power Systems.* **35**, 4644-4654 (2020)

[13] Shaheen, A., Ragab, A., Elsayed, A. & Elattar, E. Multi-objective manta ray foraging algorithm for efficient operation of hybrid AC/DC power grids with emission minimization. *IET Generation, Transmission And Distribution.* **15**, 1314-1336 (2020)

[14] Shaheen, A., Elsayed, A., Gunadi, A., El-Shimmy, R., Alharthi, M. & Ghoneim, S. novel improved marine predators' algorithm for combined heat and power economic dispatch problem. *Alexandria Engineering Journal.* **61**, 1834-1851 (2022)

[15] Wang, G., Zhang, T., Yang, Y., Cheng, J., Chang, J., Liang, X. & Hou, Z. Cross-modality paired-images generation for RGB-infrared person reidentification. *Proceedings Of The AAAI Conference On Artificial Intelligence 34(7.* pp. 12144-12151 (2020)

[16] Barma, M. & Modibbo, U. Multiobjective mathematical optimization model for municipal solid waste management with economic analysis of reuse/recycling recovered waste materials. *Journal Of Computational And Cognitive Engineering.* **1**, 122-137 (2022)

[17] Abdollahzadeh, B. & Gharehchopogh, F. multi-objective optimization algorithm for feature selection problems. *Engineering With Computers.* **38**, 1845-1863 (2022)

[18] Wu, C. Li C Distribution uniformity of water-based binders in si anodes and the distribution effects on cell performance. *ACS Sustainable Chemistry & Engineering.* **8**, 6868-6876 (2020)

[19] Miguel Nacarlos, M., Wolgemuth, J., Haraf, S. & Fisk, N. Anti-oppressive pedagogies in online learning: A critical review. *Distance Education.* **41**, 345-360 (2020)

[20] Xie, Y., Du, L., Zhao, J., Liu, C. & Li, W. Multi-objective optimization of process parameters in stamping based on an improved RBM–BPNN network and MOPSO algorithm. *Structural And Multidisciplinary Optimization.* **64**, 4209-4235 (2021)

[21] Wang, S., Zhang, T., Kong, W., Wen, G. & Yu, Y. An improved MOPSO approach with adaptive strategy for identifying biomarkers from gene expression dataset. *Mathematical Biosciences And Engineering: MBE.* **20**, 1580-1598 (2022)

Edited by: Zhengyi Chai
Special issue on: Data-Driven Optimization Algorithms for Sustainable and Smart City
Received: Oct 16, 2023
Accepted: Mar 20, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 2034–2047, DOI 10.12694/scpe.v25i3.2701

PRODUCT DESIGN FOR THE ELDERLY BASED ON HUMAN-COMPUTER INTERACTION IN THE ERA OF BIG DATA

ZHIHONG LIU[*]

Abstract. Since 21st century, social aging trend has become more and more serious. It needs accurately identify the nursing needs of elderly disabled people, who are characterized by inconvenient movement and unclear speech. How to solve these problems has become a key point in the field of elderly care and medical care. In response to this problem, this research has designed a human-computer interactive gesture recognition system for elderly nursing beds in the context of big data. The recognition rate of the fusion feature + support vector machine (SVM) classifier adopted in this study is higher than 90% for each category of gesture. On the test set, this method has an average recognition rate of 96.35%, which is much higher than that of single feature + SVM classifier. While other methods' recognition rate is lower than 90%. The recognition rate of tag c (nursing bed posture turning left) with obvious gesture feature information is as high as 99.28%, and that of tag h (nursing bed posture bedpan lowering) with weak gesture feature is 93.65%. The human-computer interaction system has well realized the recognition intention of user's dynamic and static gestures, achieved the goal set by the research, and the interaction form is natural and reliable. In the later research, we can further realize a more comprehensive, accurate and natural human-computer interaction product design through the multi-channel joint decision-making scheme to meet the needs of the elderly.

Key words: Gesture recognition; Human-computer interaction; SVM algorithm; DTW algorithm; Feature fusion

1. Introduction. Since the 21st century, aging trend has become more and more serious in our country. By the end of 2017, the population aged 60 and over was 158 million and 241 million, respectively. That are accounting for 11.4% and 17.3% of the total population, both are 0.6 percentage points higher than 2016 [1-3]. The proportion of the elderly aged 60 and above in the total population in China has exceeded 10%, and has entered into serious aging. At the same time, the problem of population aging is increasingly obvious. Elderly people aged 80 and above in China are increasing at a rate of 5% every year, and will reach 74 and more million by 2040 [4]. The accelerated process of population aging and aging has led to an increasing number of "disabled elderly" who have lost their ability to take care of themselves and "mentally retarded elderly" who have lost their ability to speak clearly and memory decline. According to the latest data of the National Bureau of Statistics, the number of disabled and partially disabled people has exceeded 40 million, and the number of completely disabled people accounts for nearly 30% [5]. Because the elderly with "disability and dementia" have the characteristics of lisping and inconvenient to get up, they not only often need 24-hour care, food, drink and live carefully, but also need to accurately identify the needs of such elderly people, improve their quality of life while maintaining their dignity. The care of such elderly people has become the biggest "pain point" in the field of elderly care and medical care in China, It will bring a severe test to the social old-age medical system. With the significant improvement of computer computing ability, gesture recognition technology has made a major breakthrough in algorithm theory and become an important breakthrough to solve the above problems. To this end, based on the research of gesture recognition algorithm, according to the gesture behavior habits of the elderly disabled, volunteers set up a self-defined sample library, and designed a gesture recognition human-computer interaction system with nursing bed as the carrier. In this study, a human-computer interactive gesture recognition system for the elderly nursing bed was specifically designed on the basis of human-computer interactive product design for the elderly under the background of big data.

2. Related work. Due to the increasingly severe aging trend in our country, the number of "disabled elderly" who have lost the ability to take care of themselves and "mentally retarded elderly" who have lost their

[*] Ecological Livable College, Hunan Polytechnic of Environment and Biology, Hengyang, 421005, China (Zhihong_Liu2023@outlook.com

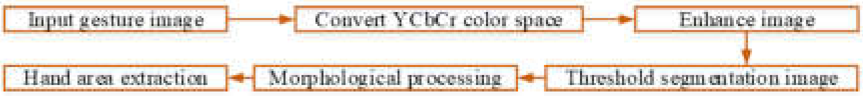


Fig. 3.1: Composition block diagram of hand area pretreatment

ability to speak clearly and memory decline is increasing. And other key issues that need to be solved urgently in the field of elderly care and medical care [6]. Nursing bed control system and human-computer interactive elderly nursing bed gesture recognition system [7]. Foreign gesture recognition technology developed earlier [8]. Mahmoud R et al. used depth optical flow estimation to calculate isolated gestures speed, and extracted relevant features from them. They input it into the linear SVM equally with the constructed gray sequence to achieve the classification of isolation segments. This model had good recall rate and robustness, and good performance and applicability [9]. Pan et al. proposed a hybrid flexible wearable system that can recognize complex gestures. The system used a simple dual-mode capacitive sensor algorithm to develop a low-power interface circuit. The dual-mode sensing platform could simultaneously perceive the space environment and identify local interactions [10]. Miao et al. proposed a new method for dynamic gesture recognition. The gesture recognition accuracy is higher than the existing methods [11]. Zhang et al. proposed a gesture recognition method. This method is based on unsupervised driving. It enabled the transfer of data without data label [12]. Liu et al. proposed a gesture recognition method, which could provide tactile and visual information and simulate neural morphology processing capabilities. This method could carry out 1000 tensile cycles and show excellent tensile endurance while maintaining the stability of electrical properties [13].

Hafsa et al. proposed a improved genetic algorithm (GA). By using non-blind search methods, this model could achieve image reconstruction. This model could use CNN for classification and achieve 92% accuracy [14]. Assisted by deep learning method, Yang J et al. proposed a wearable tactile sensor, which could realize gesture recognition and interaction. Experiments showed that this method could effectively balance the prediction accuracy of the model [15]. To sum up, many mathematicians' research on the design of gesture recognition model can provide reference for this study.

3. Human-computer interaction-based gesture recognition method for the elderly disabled.

3.1. Vision-based gesture image preprocessing method. As aging continues to deepen, it is particularly important to design intelligent elderly products, such as nursing beds and other intelligent medical products. Based on the human-computer interaction of the nursing bed, most elderly disabled people lie or sit on the bed. The complex and changeable background environment will inevitably affect the extraction of gesture target area. Therefore, pre-processing is carried out for collected information image. The block diagram of specific steps is displayed in Figure 3.1.

After the RGB color picture of the collected image is transformed into YCbCr space, the image is enhanced in Figure 1. Then, the single Gaussian model skin color model is used to complete the hand gesture segmentation. Then, the segmented binary image is processed with the morphology of closing first and then opening. The skin-like region is filtered by the area operator. Finally, the hand region obtained by the face elimination algorithm is used for the region graph with only the hand face. The YCbCr color space is derived from YUV (brightness - Y, chroma - U, concentration - V), which can describe digital video signals and optimize the transmission of video or pictures. Formula (3.1) shows the conversion formula from RGB to YCbCr color space.

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \begin{bmatrix} 65.481 & 128.553 & 24.996 \\ -37.797 & -74.203 & 112.000 \\ 112.000 & -93.786 & -18.214 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{3.1}$$

In the YCbCr color space, the single component Y is the brightness in formula (3.1). Cr and Cb are two color difference components, which are obtained by U and V through a little adjustment and used to store color information. Cb is the difference between the blue component and the reference value. Cr is the difference between the red component and the reference value. And YCbCr occupies less bandwidth. To achieve the best effect of image preprocessing. Based on the skin color details, the image data of Cr channel is concentrated on the image data of Cr channel, so the image data of Cr channel is taken as the test object in the image of color channel separation. Then the image is smoothed and sharpened to enhance the image. The gesture after image recognition is separated from the face in image segmentation. Otsu dynamic adaptive threshold method is selected for image segmentation. First, set all gray levels in the image as L levels, and the other gray level is the number of pixels. See formula (3.2) for the expression.

$$N=\sum_{i=0}^{L-1} n_i, P_i=\frac{n_i}{N}, \sum_{i=0}^{L-1} P_i=1 \tag{3.2}$$

The total number of all pixels in the image is N, and the probability of each gray level is P_i in formula (3.2). Then the gray level is divided by the threshold K, and the probability and average gray value of the two groups of gray levels $C_0=[0,1,...,k-1]$ and $C_1=[k,k+1,...,L-1]$ the average gray value of the whole image can be obtained, as shown in formula (3.3).

$$\begin{cases} \omega_0=\sum_{i=0}^{k-1} P_i=\sum_{i=0}^{k-1} n_i\Big/N, \omega_1=\sum_{i=k}^{L-1} P_i=\sum_{i=k}^{L-1} n_i\Big/N=1-\omega_0 \\ \mu_0=\sum_{i=0}^{k-1} iP_i\Big/\omega_0, \mu_1=\sum_{i=k}^{L-1} iP_i\Big/\omega_1 \\ \mu=\sum_{i=0}^{L-1} iP_i=\omega_0\mu_0+\omega_1\mu_1 \end{cases} \tag{3.3}$$

In formula (3.3), ω_0 is the probability of gray level generation, μ_0 is the average gray value, and μ is the average gray value of the entire image. The average variances of the two groups of C_0 and C_1 are shown in formula (3.4).

$$\sigma_B^2=\omega_0(\mu_0-\mu)^2+\omega_1(\mu_1-\mu)^2=\omega_0\omega_1(\mu_1-\mu_0)^2=(\mu_1-\mu)(\mu-\mu_0) \tag{3.4}$$

In formula (3.4), σ_B^2 is the variance between the two groups of gray levels. When inter-class variance owns the largest value, the corresponding optimal threshold is expressed by formula (3.5).

$$k^*=Arg\underset{0\le i\le L-1}{Max}\left[\sigma_B^2\right] \tag{3.5}$$

After obtaining the optimal threshold k^* in formula (3.5), this method is improved to get better adaptability. At the same time, it is necessary to ensure that the pixel mean distance between the target image and the whole image is $|\mu_0-\mu|$. The pixel mean distance between the background image and the whole image is $|\mu_1-\mu|$ as large as possible. Only when the above conditions are met can the weighted sum or product of the two be maximized. According to this idea, the pixel average variance is used to replace the pixel average. In formula (3.6), the improved new threshold formula is shown.

$$k^*=Arg\underset{0\le i\le L-1}{Max}\left[\left(\sigma_0^2-\sigma^2\right)^2\left(\sigma_1^2-\sigma^2\right)^2\right] \tag{3.6}$$

The improved new threshold has certain robustness to the impact of image contrast and brightness changes in formula (3.6). However, there are still burrs, holes and object noise in the binary image. Mathematical morphological transformation can eliminate noise and elements irrelevant to the image itself. Including corrosion, expansion and opening and closing operations. Suppose A is the target area on the (x,y) plane, S is the structural element with the size and shape set, and the corresponding area of the structural element S on this coordinate is $S(x,y)$. In formula (3.7), the results of corrosion, expansion and opening and closing operations in A zone are shown.

$$\begin{cases} (x,y)\left|(x,y)\in A, S(x,y)/A=\emptyset\right. \\ (x,y)\left|(x,y)\in A, S(x,y)\cap A\neq\emptyset\right. \\ A\circ S=(A-S)+S \\ A\bullet S=(A+S)-S \end{cases} \tag{3.7}$$

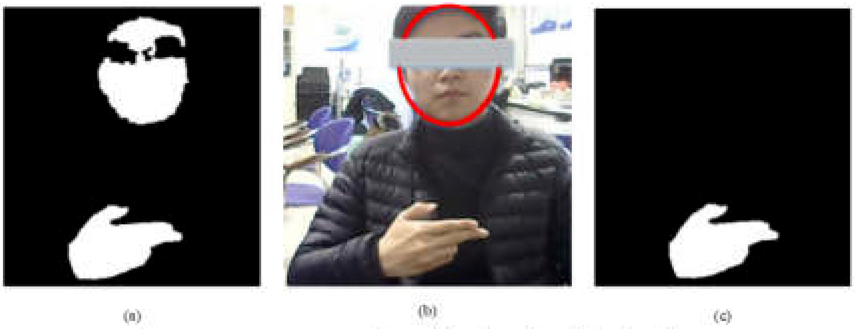


Fig. 3.2: Area operator operation and face detection elimination diagram

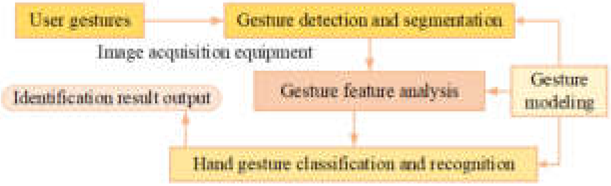

User gestures
Gesture detection and segmentation
Image acquisition equipment
Identification result output
Gesture feature analysis
Gesture modeling
Hand gesture classification and recognition

Fig. 3.3: Gesture recognition process

In formula (3.7), S is the structural element. First, in terms of corrosion, move from the upper left corner of the image in order. When S moves to a coordinate point and S is in the target area. The pixel points will be retained, otherwise deleted. Secondly, in terms of expansion, move the position of S from the top left corner of the image in sequence. When S moves to a coordinate point and the area of S and the target image intersect, it will be retained, otherwise it will be deleted. In the open and close operation, $A + S$ is the expansion of S pair. $A - S$ is the corrosion of S pair A, so A is defined as the open operation of S pair A. $A \bullet S$ is the closed operation that S does to A. Then, in Figure 3.2, the area operator is used to filter the specific effect.

First remove the non-skinned areas in Figure 3.2. Secondly, the face detection file method of OpenCV is used to draw ellipse for the face area detected by Cascade Classifier. Finally, the area operator is used to detect the face and eliminate the final effect.

3.2. Dynamic and static gesture feature extraction and recognition methods. In terms of dynamic and static gesture extraction, static gesture feature extraction is based on hand shape change, mainly using physical, geometric and mathematical features [16]. Dynamic gesture feature extraction is mainly through tracking the moving target, and the tracking algorithm obtains the track feature of the gesture area to recognize the gesture. Finally, the fusion of dynamic and static gesture features is realized. Figure 3.3 shows the overall process of gesture recognition.

First of all, on the extraction of static gestures, Fourier descriptor gesture contour extraction. Gesture structure feature extraction based on the ratio of perimeter to area of gesture contour. In formula (3.8), the perimeter of the gesture is calculated by the sum of the closed curve pixel points in the gesture contour area.

$$C = \sum \sum n(x, y) \tag{3.8}$$

In formula (3.8), C is the perimeter of the gesture contour. The pixel on the contour curve is 255. Calculate the number of white pixel points in all areas in the figure to get C. In formula (3.9), $n(x, y)$ is the number of white pixels on the contour.

$$n(x, y) = \begin{cases} 1 & ,if\ f(x, y) = 255 \\ 0 & ,others\ f(x, y) = 0 \end{cases} \tag{3.9}$$

In formula (3.9), $f(x,y)$ is the pixel value corresponding to point (x,y). Therefore, the processed image is always a binary image with 255 or 0 pixels. Scan the binary image of the gesture area. Formula (3.10) is the formula for calculating the area in the gesture contour area.

$$S=\sum\sum N\left(x,y\right) \tag{3.10}$$

In formula (3.10), $N(x,y)$ is the number of white pixels in the contour closed area, and the calculation formula of $n(x,y)$ is the same. The calculation of specific gesture area perimeter ratio M is shown in formula (3.11).

$$M=S/C \tag{3.11}$$

The final gesture area and perimeter ratio can be calculated by formula (3.11). The obtained gesture binary sample is as shown in Figure 3.2(c), and the size is uniformly adjusted to 320 * 240 resolution. Dynamic gesture feature extraction is carried out by using the hand gesture centroid of the track tracking algorithm Cam shift and the dynamic object of the optical flow method. Cam shift algorithm is a continuous adaptive Mean Shift tracking algorithm. In formula (3.12), Mean Shift tracking algorithm determines the zero-order moment M_{00}, first-order moment M_{01}, M_{10} and second-order moment M_{20}, M_{02}, M_{11} of the target color probability distribution.

$$\begin{cases} M_{00}=\sum_x\sum_y I(x,y) \\ M_{01}=\sum_x\sum_y xI(x,y) \\ M_{10}=\sum_x\sum_y yI(x,y) \\ M_{20}=\sum_x\sum_y x^2I(x,y) \\ M_{02}=\sum_x\sum_y y^2I(x,y) \\ M_{11}=\sum_x\sum_y xyI(x,y) \end{cases} \tag{3.12}$$

According to formula (3.12), we can get the centroid position (x,y). Formula (3.13) represents the bearing angle θ with the target.

$$\begin{cases} (x,y)=(M_{10}/M_{00}, M_{01}/M_{00}) \\ \theta=1/2\arctan\left[2(M_{11}/M_{00}-xy)/(M_{20}/M_{00}-x^2)-(M_{02}/M_{00}-y^2)\right] \end{cases} \tag{3.13}$$

After the centroid position is obtained in formula (3.13), it is necessary to judge whether the center position coordinate at this time has reached convergence. If not, return to the step of formula (3.12) and recalculate the area. The characteristics of gesture motion track mainly include: position information, speed and direction angle. Formula (3.14) is the expression formula of the trajectory direction angle of the gesture centroid.

$$\theta_t=\arctan\left[\left(y_t-y_{t-1}\right)/\left(x_t-x_{t-1}\right)\right] \tag{3.14}$$

In formula (3.14), θ_t and (x_t,y_t) are the azimuth and position coordinates at time t. (x_{t-1},y_{t-1}) is its position coordinate at the time of Dt-1. Dynamic gesture feature extraction based on optical flow method. LK optical flow algorithm is not only fast in operation, but also accurate in tracking target. The image pyramid is a sequence of images with scale changes obtained from a group of adjacent level images through a previous low-pass filter. Assume that the abscissa value of the image element is x, y is the ordinate value. Formula (3.15) is the inter-layer operation.

$$G_i\left(x,y\right)=\sum_{m=-2}^{2}\sum_{n=-2}^{2}w(m,n)G_{i-1}(2x+m,2y+n) \tag{3.15}$$

The lowest layer G_0 of the image is the original image, and G_i is the image of layer i in formula (3.15). The pixels of layer i are obtained by weighted average of the image matrix corresponding to the previous layer

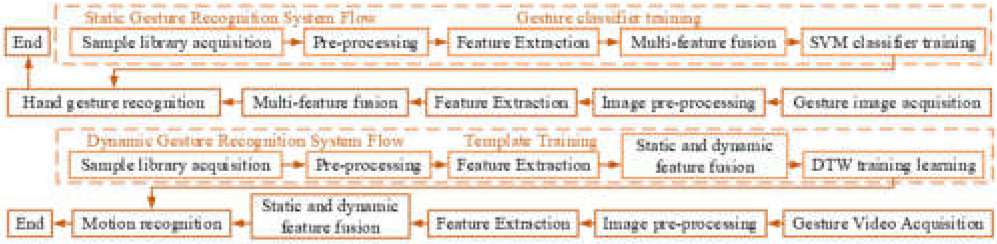


Fig. 3.4: Dynamic and static gesture recognition system flow

through Gaussian window function w. $w(m,n)$ is the window function. When the window size is 5 * 5, the four constraints in equation (3.16) are met.

$$\begin{cases} w(m,n) = w(m) * w(n), m \in [-2,2], n \in [-2,2] \\ \sum_{-2}^{2} w(m) = 1 \\ w(m) = w(-m) \\ w(-2) + w(2) + w(0) = w(-1) + w(1) \end{cases} \tag{3.16}$$

The window function can be obtained by equation (3.16). In the method of hand gesture target classification, we mainly adopt (dynamic) template matching method and (static) support vector machine method. Figure 4 describes the flow chart of dynamic and static gesture recognition system.

In the recognition of static gestures, the training set uses SVM classifier for parameter adjustment training in Figure 3.4. In the recognition of dynamic gestures, DTM algorithm is used to compare the obtained hand shape and trajectory fusion feature vectors with the training samples in the template library. The gesture with the shortest matching distance is the gesture to be recognized. Equation (3.17) describes the linear mapping function expression of specific SVM.

$$f(x_i, w, b) = w^T x_i - b \tag{3.17}$$

In formula (3.17), the penalty function $f(x_i, w, b)$ is defined as the scoring function of SVM. w^T and b are the parameters that SVM needs to train. There are 8 kinds of gestures to distinguish. Because the sample type is not large, the "one-to-one" multi-classification strategy is selected, which can ensure high accuracy and will not cause too much loss in speed. The final decision function of "one-to-one" multi-classification is expressed by formula (3.18).

$$\begin{cases} y = \arg \max\limits_{i \in [1,2,\ldots,n]} \left(\sum_{j \neq i} count_{ij} \right) \\ count_{ij} = \begin{bmatrix} 1 & SVM_{ij}(x) \rightarrow i \\ 0 & SVM_{ij}(x) \rightarrow j \end{bmatrix} \end{cases} \tag{3.18}$$

In formula (3.18), the SVM classifier decision samples composed of $SVM_{ij} \rightarrow i$ and j belong to class i. For 8 different types of gestures, 28 SVM classifiers need to be constructed, and the decision function is used to determine which type of gesture to be recognized is the closest among the 8 gesture categories. In the training stage of SVM, it is still necessary to manually input some super parameters. The change of decision function after the introduction of kernel function is shown in formula (3.19).

$$f(x) = sgn \sum_{i-1}^{n} a_i * y_i K(x_i *, x) + b* \tag{3.19}$$

The RBF kernel function is selected to map the samples to the high-dimensional space in formula (3.19), so better training and classification results can be obtained. The optimization of dynamic gesture recognition is using DTW algorithm. The global optimal path of the DTW algorithm is only related to the region within the parallelogram, and the feature vector sequence to be recognized does not need to consider the element matching outside the region. The dynamic path is divided into three parts $(1, X_a)$, (X_{a+1}, X_b) and (X_{b+1}, N). In Formula (3.20), through the slope of the parallelogram, the values of D and E can be calculated.

$$\begin{cases} X_a = 1/3(2M - N) \\ X_b = 2/3(2N - M) \end{cases} \tag{3.20}$$

In formula (3.20), the length condition of M, N is limited by $2M - N \geq 3$ hand $2N - M \geq 2$. In this constraint, each frame on the i axis only needs to be compared with the frame between $[y_{\min}, y_{\max}]$ on the j axis. When $X_a = X_b$, the comparison is divided into two sections. See Formula (3.21) for details.

$$\begin{cases} 1/2x \sim 2x, & x \leq x_a \\ 2x + (M - 2N) \sim 1/2x + (M - 1/2N), & x > x_b \end{cases} \tag{3.21}$$

When $X_a < X_b$, the comparison is divided into three sections. See formula (3.22) for details.

$$\begin{cases} 1/2x \sim 2x, & x \leq x_a \\ 1/2x \sim 1/2x + (M - 1/2N), & x_a < x \leq x_b \\ 2x + (M - 2N) \sim 1/2x + (M - 1/2N), & x > x_b \end{cases} \tag{3.22}$$

In the comparison of formula (3.22), when $X_a > X_b$, the situation is similar. In template training and recognition, the matching distance of DTW algorithm is calculated by Euclidean distance. Set the eigenvector of the input unknown gesture value as $X = (x_1, x_2, ..., x_m)$, and the eigenvector of a gesture sample in the template library as $G = (g_1, g_2, ..., g_m)$. See Formula (3.23) for the calculation of specific Euclidean distance.

$$D(X, G) = \sqrt{\sum_{i=1}^{m} (x_i - g_i)^2} \tag{3.23}$$

At the same time, select $\alpha(M + N)$. And α is a positive proportion coefficient, whose value is 0.25.

4. Test and analysis of gesture recognition system for elderly nursing bed based on human-computer interaction.

4.1. Analysis of test results of dynamic and static gesture recognition.. To verify the effectiveness of gesture recognition system for elderly nursing bed based on human-computer interaction. The customized gesture information includes 8 different categories. The software environment of Visual Studio 2015 (community) and OpenCV3.4.00 is adopted. The camera captures video at a frame rate of 30 frames/s. The extracted two-dimensional image has a resolution of 640 * 480. For the "one-to-one" strategy, 28 SVM two-classifiers are constructed and the classification of gestures is determined by voting. The obtained gesture classification results correspond to the command information of a bed posture control respectively in the text, achieving the purpose of human-computer interaction. Tn Figure 4.1, the binary diagram of some samples in the dataset is shown.

The number of gesture values that can be selected in the experiment is 8 in Figure 4.1, and the specific gesture shape can be customized according to the user's actual situation. The final classification of the SVM classifier is obtained from the "one-to-one" voting results of each test image. In order to protect the privacy of users, the collected data was anonymized and encrypted. First, hashing algorithm is used to convert personal identifiable information (such as name, email, mobile phone number) and other data into irreversible anonymous identifiers to achieve data anonymity. Secondly, AES strong encryption algorithm is used to encrypt the whole data set to ensure the security of data in storage and transmission. Finally, assign data access to authorized users or roles, and ensure that only those who need to know specific information can access relevant data. Compare it with the real category of the image and calculate the recognition rate of each gesture category. As shown in Table 4.1 below, the total number of training samples is 8 * 1280, and the number of test samples is 8 * 640.

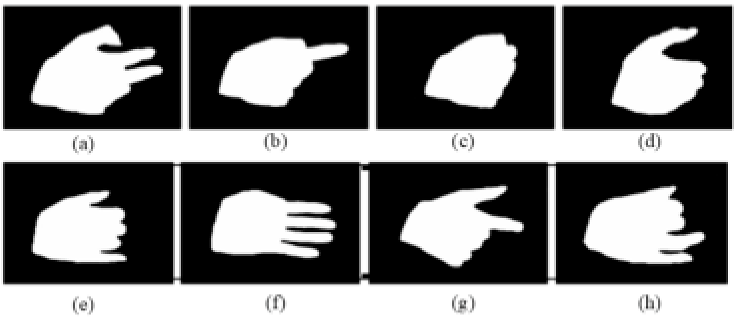


Fig. 4.1: Schematic diagram of binary sample

Table 4.1: Static gesture recognition rate

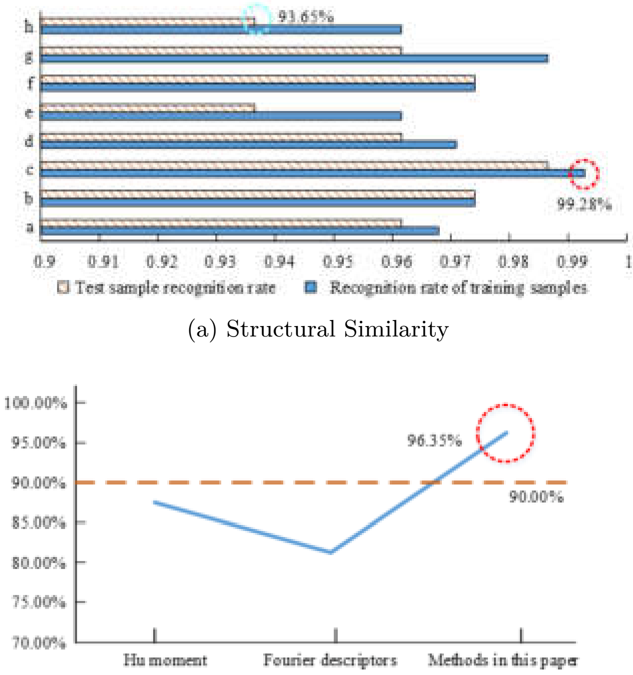
Gesture label	Recognition rate of training samples	Test sample recognition rate
a	0.9678312	0.9615375
b	0.974025	0.974025
c	0.9928062	0.9865125
d	0.9709281	0.9615375
e	0.9615375	0.9365625
f	0.974025	0.974025
g	0.9865125	0.9615375
h	0.9615375	0.9365625

The recognition rate of fusion feature + SVM classifier for each type of gesture is greater than 90% in Table 1. The recognition rate of tag c with obvious gesture feature information is as high as 99.28%, and that of tag h with weak gesture feature is 93.65%. The specific effect is shown in Figure 4.3 (a). However, advantages and disadvantages' comparison of the fusion feature algorithm is difficult only by the recognition rate of the data set. So the recognition rate test of single feature extraction for the control variables of the data set was conducted. After RGB image preprocessing and gesture segmentation, the sample library of the data set in the previous article is extracted using moment invariants and single features of Fourier descriptors. At the same time, conduct SVM training and testing. Figure 4.3(b) records the comprehensive average recognition rate comparison of 8 * 640 sample test sets of eight gestures.

This fusion feature extraction is very effective in classification and can well achieve static gesture recognition in Figure 4.3. The fusion feature + SVM classifier in this study is comprehensive. On the test set, this method's average recognition rate is 96.35%, which is much higher than that of single feature + SVM classifier. Other methods' recognition rate are lower than 90%. To some extent, it alleviates the problem of low accuracy caused by information waste. In the comparison of dynamic gesture recognition rate, the test set of volunteer gesture video data set is tested. First, four kinds of gestures in 1280 video images are processed frame by frame, then the image area of the gesture is extracted and the optical flow of the two adjacent frames is calculated, and the hand shape is normalized by feature extraction. The trajectory features and fusion features obtained only by pyramid optical flow method are compared and verified. Table 4.2 describes the comparative recognition rate of the two methods.

When faced with the distinction between eating and drinking gestures, the classification effect of static hand shape feature and dynamic track feature fusion is better than that of using dynamic track feature alone in Table 4.2. Figure 4.4 shows the specific effect analysis.

Figure 4.4 shows that less than 60% of gesture recognition optical flow features for eating and drinking, while the recognition rate of fusion features is greater than 80%. The essential reason is that the movement tracks



(a) Structural Similarity

(b) Information Entropy

Fig. 4.3: Static gesture recognition rate

Table 4.2: Dynamic gesture recognition rate

Gestures	Optical flow feature recognition rate	Fusion feature recognition rate
Eating	0.5369625	0.8366625
Drinking water	0.574425	0.8616375
Urination	0.824175	0.8866125
Defecation	0.874125	0.924075

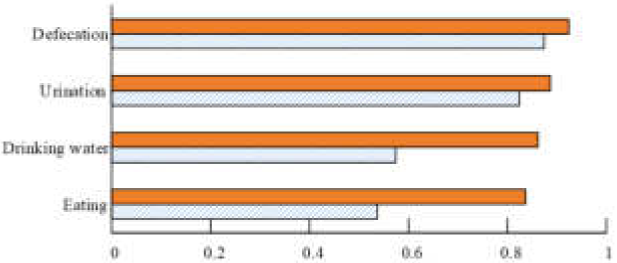


Fig. 4.4: Static gesture recognition rate

Table 4.3: The 8 action command codes correspond to the table

Action Instructions	Hand gestures	Command Code
Backlift	a	0X01
Back Down	b	0X02
Left flip	c	0X03
Right flip	d	0X04
Back up leg bend	e	0X05
Back flat leg extension	f	0X06
Potty up	g	0X07
Potty down	h	0X08

of eating and drinking are quite close, but the hand shape is different, so the individual track characteristics cannot accurately distinguish them.

4.2. Test result analysis of human-computer interactive gesture recognition system for elderly nursing bed. . To better reflect the feasibility of gesture recognition technology, the effect of the system is tested by building a QT interface, and static gestures are used to achieve bed posture control. Dynamic gestures are used to identify basic physiological needs. Through dynamic and static gesture recognition for the elderly with disabilities, the demand results are displayed on the interface to realize the auxiliary nursing effect of human-computer interaction elderly nursing bed. Firstly, a common monocular camera (CMOS sensor) with a maximum frame of 30 FPS is used as the video acquisition hardware device and installed on the inclined top of the nursing bed (fixed on the ceiling) to facilitate the acquisition of video information; Winowds 10 (Enterprise x64) operating system; DDr4 2400Hz (2 * 8GB) memory and Intel (R) Core (TM) i3-8100 CPU @ 3.60GHz CPU. The software is based on OpenCV3.4.0 open-source computer vision library, programmed under the Visual Studio 2015 (Commuity) platform, and the QT programming framework is used to display the interface effect. In Table 4.3, based on the recognition of static gestures, the nursing bed has 8 action commands, and the corresponding relationship is set.

The gesture corresponding to the command code of 0X01-0X08 controls the posture of the nursing bed in Table 4.3. In dynamic gesture recognition, the customized action template is four basic physiological needs. Before analyzing the characteristics of gestures of elderly disabled people, users need to be trained, and the training process is as follows. First of all, a phased training method is adopted in the implementation of training, so that users can gradually start from basic instructions, understand the relationship between various gestures and nursing bed operation, and finally be able to skillfully complete the expression of bed posture control and basic physiological needs. Second, the time from the beginning of the user's exposure to the system to being able to operate independently is recorded, and the user will begin to be able to perform simple operations on the nursing bed, such as back rising or leg bending, after 1-3 hours of training. However, it can take anywhere from a few days to a week for a user to fully master all of the gesture instructions, depending on individual differences such as age, learning ability and sensitivity to gesture movements. Finally, in order to improve the training process and increase adoption by older users, regular return visits will also be conducted to evaluate the effectiveness of the user's application and collect feedback as a reference for subsequent research and system optimization. After the training, the recognition results are presented in the human-computer interface by analyzing the characteristics of gestures of the elderly disabled. Figure 4.5 shows the test results of the static gesture recognition system.

After recognizing the static gestures in Figures 4.5(a) and 4.5(b), the nursing bed completes the gesture actions of turning right (c) and turning left (d) according to the gesture commands. In dynamic gesture recognition, the human-computer interaction of the nursing bed is tested. Testing in real environment can evaluate the adaptability of the system to complex and changing environments. There may be different illumination, noise levels and background interference in places such as nursing homes, so the adaptability of the test system under these conditions is very important for its practical application. In order to evaluate the performance of the system under different environmental factors, the system was tested in a more realistic environment and

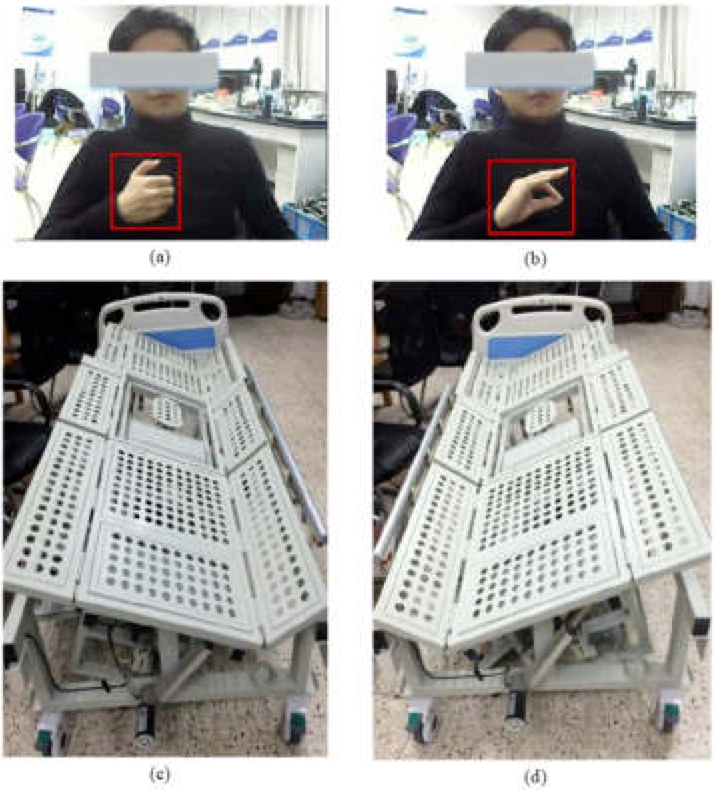


Fig. 4.5: Static gesture recognition test chart

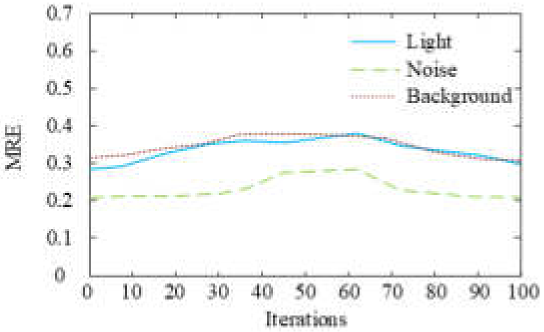


Fig. 4.6: MRE value of the system under different environmental factors

experimented under different environmental conditions. The experimental results are shown in Figure 4.6.

As can be seen from Figure 4.6, the average MRE value of the evaluation system is 0.32 under the interference of lighting environment factors. Under the interference of noise environmental factors, the average MRE value of the evaluation system is 0.25. Under the interference of background interference and environmental factors, the average MRE value of the evaluation system is 0.34. On the whole, the MRE values are all between 0.25 and 0.35, and these smaller values indicate that the prediction results of the evaluation system are close to the real values. The evaluation system performs well under the interference of environmental factors such

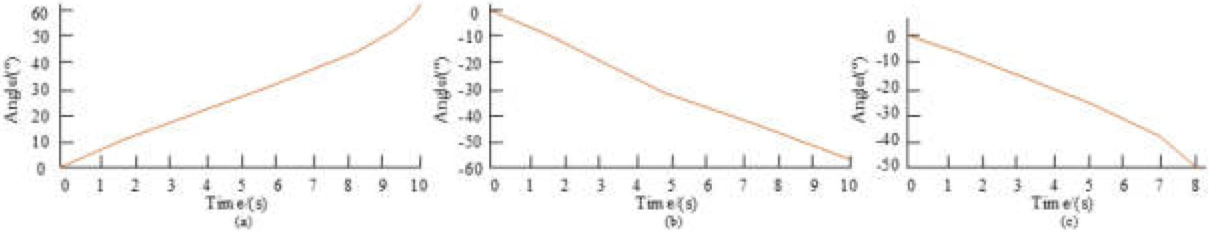


Fig. 4.7: Motion simulation results under dynamic gesture

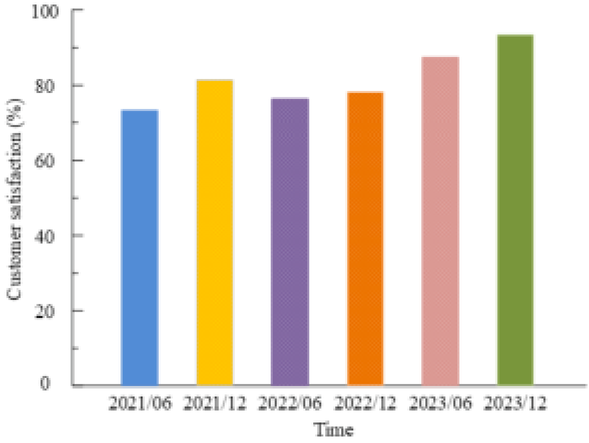


Fig. 4.8: The result chart of user satisfaction survey with time.

as lighting, noise and background interference, and can maintain high accuracy and stability. In Figure 4.7, kinematics simulation is carried out under the action commands of back up, back up, leg bending and left and right turning.

Figure 4.7(a) shows the relationship between time and back lifting angle. The rotation angle of the joint of the back lifting part is 0 ° ∼ 62 °. That meets the design requirements and the angular velocity is relatively gentle, and will not cause secondary injury due to the unstable velocity. That verifies the correctness of the kinematics modeling of the back lifting part of the nursing bed. See Figure 10(b) for the kinematics simulation results of the back-lift leg curve. The joint of the leg bending part can reach an angle of 0 ° ∼ 58 °, and the angular velocity changes smoothly and smoothly, meeting the design requirements of the elderly products. The kinematics simulation results of left and right rolling are shown in Figure 4.7(c). The joint of the rolling part can reach an angle of 0 °∼50 °, and the speed is stable, without obvious speed mutation. The human-computer interaction system has well realized the intention recognition of user's static and dynamic gestures. And it achieved the bed posture control based on static gesture recognition and the demand recognition based on dynamic gestures for users with relatively clear awareness, achieving the goal set by the research, and the interaction form is natural and reliable. User satisfaction can promote the long-term use of nursing bed gesture recognition system by elderly users. To assess the long-term usability of the system, a user satisfaction survey was conducted, the results of which are shown in Figure 4.8.

According to Figure 4.8, in June 2021, the user satisfaction reached 74.52%. This figure rose to 81.49% in December 2021, an increase of 6.97%. However, in June, 2022, user satisfaction dropped to 77.15%, a decrease of 4.34%. However, in December 2022, user satisfaction increased slightly to 78.61%, and in June 2023, it increased significantly to 87.86%. Finally, in December 2023, the user satisfaction reached 92.37%. From the data analysis, with the passage of time, the overall user satisfaction showed a trend of first rising, then slightly falling, and then continuing to rise. The decrease of satisfaction in the intermediate stage may be caused by the corresponding adjustment of the system. The continuous increase of satisfaction shows that users have a good acceptance and experience of the gesture recognition system. This can not only meet the actual needs of users, but also provide feedback for system developers to improve system functions and performance. This trend reveals that the system has certain attraction and long-term use potential.

5. Conclusion. In view of the increasingly severe trend of social aging in recent years, the number of elderly people with disabilities characterized by lisping and inconvenient movement is increasing. How to meet their nursing needs has become an urgent problem in the field of medical care and old-age care. In response to this problem, this research has designed a gesture recognition system for the elderly nursing bed based on human-computer interaction in the context of big data. The fusion feature + SVM classifier adopted in this study has a recognition rate of more than 90% for each category of gesture. The recognition rate of tag c (nursing bed posture turning left) with obvious gesture feature information is as high as 99.28%, and that of tag h (nursing bed posture bedpan lowering) with weak gesture feature is 93.65%. At the same time, this method can alleviate the problem of low accuracy caused by information waste to a certain extent. The average recognition rate of this method reaches 96.35% on the test set, which is much higher than that of single feature + SVM classifier. While other methods' recognition rate is lower than 90%. Kinematics simulation is carried out under the action command after gesture recognition. The human-computer interaction system can better realize the intention recognition of user's dynamic and static gestures, achieve the expected goal of the research. And the interaction form is natural and reliable. Since older adults may have different physiological and cognitive abilities, future research could examine how the system performs under both ability and physical conditions. The nursing bed gesture recognition system is customized according to the individual needs and wishes of users. In addition to gesture features, consider incorporating other information such as eye tracking, facial expressions, etc., into the system to provide more comprehensive and accurate intent recognition. This enhances the interactivity and user experience of the system.

REFERENCES

[1] Wang, X., Shi, R. & Niu, F. Optimization of furniture configuration for residential living room spaces in quality elderly care communities in Macao. *Frontiers Of Architectural Research.* **11**, 357-373 (2022)

[2] Faba, B., Yecl, A., Tvs, A., Ssm, A., Jt, A., Vs, C., Mg, C., Jdw, C., Ja, C. & Cgc, D. Clinical effectiveness of music interventions for dementia and depression in elderly care (MIDDEL): Australian cohort of an international pragmatic cluster-randomised controlled trial. *The Lancet Healthy Longevity.* **3**, 153-165 (2022)

[3] Mu, Q., Guo, P. & Wang, D. Optimal Subsidy Support for the Provision of Elderly Care Services in China Based on the Evolutionary Game Analysis. *IJERPH.* **19**, 1-20 (2022)

[4] Wang, X., Ma, S., Hu, D., Xu, C., Luo, X. & Tang, Z. Effect of aging temperature on the fatigue properties of shot-peened single crystal superalloy at intermediate temperature. *International Journal Of Fatigue.* **2022**, 6675-10 (0)

[5] Wang, L. Cooperative Elderly Care Services in the Greater Bay Area. *China Report ASEAN.* **7**, 50-52 (2022)

[6] Hu, J., Zhang, Y., Wang, L. & Shi, V. An Evaluation Index System of Basic Elderly Care Services Based on the Perspective of Accessibility. *IJERPH.* **19**, 1-16 (2022)

[7] Trujillo, J., Asli, Z., Kan, C., Irina, S. & Harold, B. Differences in functional brain organization during gesture recognition between autistic and neurotypical individuals. *Social Cognitive And Affective Neuroscience.* **17**, 1021-1034 (2022)

[8] Suni, S. & Gopakumar, K. Extracting Multiple Features for Dynamic Hand Gesture Recognition. *International Journal Of Engineering And Advanced Technology.* **10**, 71-75 (2021)

[9] Mahmoud, R., Belgacem, S. & Omri, M. Towards wide-scale continuous gesture recognition model for in-depth and grayscale input videos. *International Journal Of Machine Learning And Cybernetics.* **12**, 1173-1189 (2021)

[10] Pan, J., Li, Y., Luo Y, Z., Wang, X., Wong, D., Heng, C., Tham, C. & Thean, A. Hybrid-Flexible Bimodal Sensing Wearable Glove System for Complex Hand Gesture Recognition. *ACS Sensors.* **6**, 4156-4166 (2021)

[11] Miao, Y., Li, J. & Sun, S. Dynamic Gesture Recognition Combining Global Gesture Motion and Local Finger Motion. *Journal Of Computer-Aided Design & Computer Graphics.* **32**, 1492-1501 (2020)

[12] Zhang, Y., Wu, L., He, W., Zhang, Z., Yang, C., Wang, Y., Wang, Y., Tian, K., Liao, J. & Yang, Y. An Event-Driven Spatiotemporal Domain Adaptation Method for DVS Gesture Recognition. *IEEE Transactions On Circuits And Systems, II.* **69**, 1332-1336 (2022)

[13] Liu, L., Xu, W., Ni, Y., Xu, Z., Cui, B., Liu, J., Wei, H. & Xu, W. Stretchable Neuromorphic Transistor That Combines Multisensing and Information Processing for Epidermal Gesture Recognition. *ACS Nano.* **16**, 2282-2291 (2022)

[14] Hafsa, M., Atitallah, B., Salah, T., Amara, N. & Kanoun, O. genetic algorithm for image reconstruction in electrical impedance tomography for gesture recognition. *Technisches Messen: Sensoren, Gerate, Systeme.* **89**, 310-327 (2022)

[15] Yang, J., Liu, S., Meng, Y., Xu, W., Liu, S., Jia, L., Chen, G., Qin, Y., Han, M. & Li, X. Self-Powered Tactile Sensor for Gesture Recognition Using Deep Learning Algorithms. *ACS Applied Materials & Interfaces.* **14**, 25629-25637 (2022)

[16] Envelope, M., Rlpdm, B. & Vfdlj, B. Gesture recognition of wrist motion based on wearables sensors. *Procedia Computer Science.* **210**, 181-188 (2022)

Edited by: Zhengyi Chai
Special issue on: Data-Driven Optimization Algorithms for Sustainable and Smart City
Received: Oct 17, 2023
Accepted: Mar 21, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 2048–2061, DOI 10.12694/scpe.v25i3.2708

DETECTION METHOD OF TOURIST FLOW IN SCENIC SPOTS BASED ON KALMAN FILTER PREDICTION

XIAOYAN XU*AND LI ZHANG†

Abstract. The tourism industry has developed rapidly, but it is always limited by the environmental carrying capacity and cannot receive too many tourists at the same time. Therefore, it is very necessary to limit the number of tourists visiting at the same time based on traffic detection. To this end, the tourist scenic spot (TSS) traffic statistics system was designed. The system performed graying, binarization, image denoising, and morphological processing on the image. The pre-processed image used the background difference method based on mixed Gaussian background modeling to detect moving objects. The improved Hough transform circle detection method was used to identify the head target, and the Kalman filter (KF) was used to complete the target tracking. KF could predict the target trajectory accurately, and the improved Hough transform circle detection method could recognize the head under occlusion. The maximum missed detection rate of the statistical system was 3.2%, the minimum is 0, and the overall detection accuracy was the highest. The error rate of inbound passenger flow was 4.10%, and the error rate of outbound passenger flow was 3.0%. Using this system can control the tourist flow (TF) in the scenic spot and avoid safety accidents due to excessive passenger flow. And it is conducive to the sustainable development of the scenic spot.

Key words: Moving target detection; KF; Tourist flow in scenic spots; Hough transform circle detection

1. Introduction. In recent years, tourism has gradually become a leisure and lifestyle of people around the world. According to the figures released by the World Tourism Association, tourism ranked third in the global GDP growth in 2019. It is reported that the tourism industry (TI) has grown at a rate of 3.5%, far exceeding the growth rate of 2.5% of the world's GDP. Taking China as an example, TI provides nearly 80 million jobs, equivalent to 10.3% of the whole country. Meanwhile, the total output value of China's TI is about 10.9 trillion-yuan, accounting for 11.3% of China's economy. The rapid development of the global tourism market has led to the rapid development of China's domestic TI [1]. The TI of China has stepped into ”mass tourism”, and tourists' willingness to travel is increasing. It can be expected that TI will continue to thrive even in the post-epidemic period. However, due to the sharp increase in the number of tourists, the increase in tourist experience and traffic safety accidents of TSS has brought a great negative impact on TSS. Relevant departments have proposed measures to strengthen TF control to solve the hidden danger of urban traffic safety. As the carrying capacity of the urban population exceeds the carrying capacity, the tourist experience will be reduced. This has led to deterioration of the ecological environment and aggravation of environmental pollution, and it even triggers social security incidents [2-3]. If the TSS-TF can be accurately detected, it can prevent a large number of tourists from staying, achieve peak travel, improve the experience of TSS tourists, and promote the comprehensive development of TSS. In addition to tourism TSS, due to the development of the economy, various large transportation hubs and public places will be crowded, which will have a great impact on urban transportation and public places [4]. Therefore, if the passenger flow statistics are carried out for commercial locations, the operator can make reasonable decisions based on the changes in passenger flow, which leads to more benefits for the operator. The flow of visitors can be reflected in real-time through the statistics of passenger traffic at TSS and other cultural and entertainment venues. Thus, the development trend of tourism off-season and peak season is obtained, which is convenient to establish the corresponding safety warning mechanism. Real-time dispatching and management of passenger flow can be realized through the statistics of passenger flow at airports, subway stations and other transportation hubs. The passenger flow statistics of mobile vehicles such as buses and subways can be used for early warning management of overload.

*School of Tourism Management, Xinyang Agriculture and Forestry University, Xinyang, 464000, China
†School of Tourism Management, Xinyang Agriculture and Forestry University, Xinyang, 464000, China. Corresponding E-mail: li_zhang2023@outlook.com

Traditional TF detection mainly relies on empirical method, but it is difficult to meet the current demand due to its short time and low accuracy [5]. So how to detect TF faster and more accurately has become an important part of TSS management. This requires further deepening and improving the tourist experience, avoiding large-scale tourist detention, improving the quality of tourist service and ensuring the safety of tourists' lives and property. Therefore, the research will detect TF of tourism TSS by moving target detection method based on mixed Gaussian background modeling background difference method, Hough transform circle detection method, and KF moving target tracking method.

2. Related works. The research on monitoring methods of tourist TF in the industry has produced many research results. Rogowski M et al. proposed the MSTT tourist behavior monitoring system. The system used 39 pyroelectric sensors and survey data to obtain the space-time characteristics of TF. It realized the monitoring of mountaineering, automobile traffic, and the evaluation of illegal tourism. The system was feasible [6]. Kumaran N et al. proposed an abnormal behavior detection scheme in human group activities based on social behavior. The method included a feature description method using covariance matrix coding to realize optical flow expression and a covariance matrix model using mixed optical flow. This method could better simulate the group activities and identify the monitored abnormal events [7]. Lev et al. proposed a K-Means-based method to identify the spatial distribution characteristics of TSS TF. This method combined the location big data features to express the data source in a temporal and spatial manner, which analyzed the TF temporal features in combination with the distance measurement of dynamic time warping and K-Means. This method could analyze the flow time series type and identify the flow spatial distribution characteristics. The analysis results were helpful to the internal traffic and facility management of TSS [8]. Zimoch M et al. established a TF recognition system based on thermal map and SSD method. The system used edge image analysis, pre-trained SSD method, and displacement vector centroid algorithm combined with Siam network. In the Market 1501 data set, the system had a level 1 efficiency of 84.6% [9]. Du S et al. proposed a TSS tourist feature analysis method based on wavelet nonlinear transform. In spring and autumn, the TF changed greatly, while in winter it is relatively small. Inbound tourists showed obvious seasonal characteristics and showed clear characteristics of off-season and peak season [10].

Classic methods such as KF are often used in the industry to track human targets to improve the accuracy of TF detection. The research on KF has also produced many results. He L et al. combined adaptive KF with transient flow model and adaptive control to propose a real-time unsteady flow estimation method for multi-product pipelines. In this method, the linear flow field model combined with difference conversion and frequency response was used to replace the nonlinear flow model, which was applied to the linear compensation of the unsteady flow model. Under unpredictable conditions, this method could make the transient flow estimation error of linear model and conventional nonlinear model less than 0.5%, effectively overcoming the problems of traditional methods [11]. Zhang HF et al. proposed a detection method based on KF and wavelet transform for pitch target recognition. This method could well suppress the mixed white noise in the music. Compared with the fundamental frequency detection algorithm based on Fourier transform, its average recognition error rate in the full frequency range decreased by 0.89%, and its robustness was more ideal [12]. Yu X et al. applied the KF estimation model to cycle slip detection and repair. In this study, a cycle slip detection and repair method with strong adaptability was proposed, and then the KF estimation model was established by connecting the cycle slip detection equation and the state equation. KF estimation model could detect and repair all simulated cycle slips [13]. Yang J et al. adopted a new method based on differential KF to track the lag behavior of civil engineering structures. This method combined QR decomposition with strong tracking filtering to track the changes of structural parameters. The hysteresis curve obtained by the proposed algorithm was in good agreement with the test curve [14]. Zhou P et al. proposed a robust model-free adaptive predictive control method based on KF for quality control of molten iron in blast furnace. This control strategy could better overcome data loss, measurement noise and other problems [15].

To sum up, previous passenger flow statistics methods often use sensors and monitoring images to identify human targets. However, due to the complex situation of pedestrians entering side by side and people moving in both directions, these passenger flow statistics methods are not competent. Based on the complexity of TF motion, this research will propose a moving target detection method based on head tracking, which can improve the accuracy of moving human detection through real-time tracking of human head.

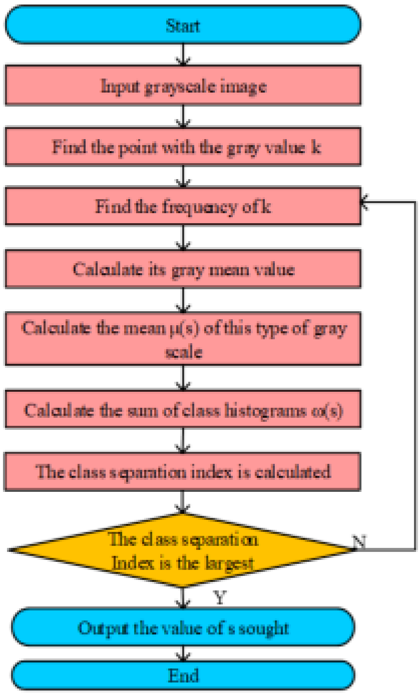

Start
Input grayscale image
Find the point with the gray value k
Find the frequency of k
Calculate its gray mean value
Calculate the mean μ(s) of this type of gray scale
Calculate the sum of class histograms ω(s)
The class separation index is calculated
The class separation Index is the largest
N
Y
Output the value of s sought
End

Fig. 3.1: The method of image graying

3. Tourism TSS-TF detection method based on KF prediction.

3.1. Video image preprocessing method.. In the passenger flow monitoring of TSS, the video image of TSS should be obtained first. When acquiring video images, the image has a certain degree of distortion due to the influence of various environmental factors such as noise, light and so on, so it must be preprocessed. The image preprocessing methods used in this study include image binarization, image graying, image denoising, and image morphology processing [16].

The method of image graying is to transform RGB three-channel image data of color image into single-channel image data for processing. The research uses the weighted average method to weighted average the RGB three components to obtain a more reasonable gray image. Image binarization is the process of changing the gray value of an image to make a color image into a black and white image. Its core is the selection of threshold. Otsu method is selected for threshold selection. The flow of Otsu algorithm is shown in Figure 3.1. First of all, it needs to find the point in the gray image with the gray value of k and find the frequency of the point. Then it needs to calculate its gray mean value, the sum of the gray mean value $\mu(s)$ and the class histogram $\omega(s)$ to calculate the class separation index. When the class separation index is maximum, the s value obtained is the best threshold of the image.

There are many methods for image denoising, among which median filtering can remove impulse noise and salt and pepper noise very well, and the edge of the image is not affected. Therefore, median filtering is used for denoising [17]. If the position of the pixel points to be processed in the image is (m,n), and the gray value of the point is shown in $g(m,n)$. Then the processing result of median filtering for the point is shown in formula (3.1).

$$g(m,n) = med\{f(m-v)(n-v),...,f(m-1)(n-1),f(m,n),f(m+1)(n+1),...,f(m+v)(n+v)\} \quad (3.1)$$

where $f(m,n)$ is the gray value at position (m,n), and $2v$ is the number of pixels in the filter window. Morphological processing refers to the use of a structural element in the image to perform a set operation on the original binary image to extract useful information of the image. The basic morphological operations include expansion, corrosion, open, and close operations. The expansion process is that image a is expanded by structural element b, which is recorded as $a \oplus b$, where $\oplus$ is the expansion operator. Formula (3.2) is the expansion expression. w

$$a \oplus b = \{x \Big| (\hat{b}_x \cap a) \neq \emptyset\} \tag{3.2}$$

where x is the distance b first maps about the origin and then translates. Formula (3.3) is the expression of corrosion.

$$a \odot b = \{x \Big| (\hat{b}_x \cap a) \neq \emptyset\} \tag{3.3}$$

where $\odot$ is the operator of corrosion. Operation calculation is to corrode the binary image first and then expand it. The expression is shown in formula (3.4).

$$a \circ b = (a \circ b) \oplus b \tag{3.4}$$

where $\circ$ is the operator of the open operation. The close operation is to expand the binary image first and then corrode it. The expression is shown in formula (3.5).

$$a \bullet b = (a \oplus b) \odot b \tag{3.5}$$

where $\bullet$ is a closed operator. The above processing measures are conducive to extracting the required image information. It can reduce irrelevant information in video images, enhance useful information, and simplify image data, laying a foundation for subsequent target detection.

3.2. Moving target detection method based on background difference method of mixed Gaussian background modeling.. The pre-processed image is detected to recognize the moving human object in the image. Moving human detection is divided into two parts: moving target detection and pedestrian target recognition. Moving target detection is to detect the moving object in the sequence image, so as to obtain the region of the moving target. The commonly used moving target detection methods include optical flow method, inter-frame difference method, and background difference method. The optical flow method has many iterations and complex calculation, and it is generally difficult to realize real-time detection. The detection threshold of the inter-frame difference method is selected by manual experiments, and the adjustment steps are cumbersome [18]. So, the background difference method is selected.

The principle of background difference method is to subtract the current frame from the background image to get the difference value. The threshold is set in advance. If the difference value is greater than the set threshold, it is considered as a moving target. Otherwise, it is not a moving target. The algorithm expression is shown in formula (3.6) [19].

$$d(m,n) = \begin{cases} 0, |f_k(m,n) - b(m,n)| \leq t \\ 1, others \end{cases} \tag{3.6}$$

where $d(m,n)$ is the binary differential image, $f_k(m,n)$ represents the current frame image, $f_k(m,n)$ is the background image, and t is the threshold. The background difference method will change the image background in practical application, resulting in inaccurate target detection. Therefore, the background difference method generally needs to establish a background model to achieve real-time background update. In the monitoring area, most of the background images are modal. At this point, the background of each mode can be represented by a Gaussian function. Therefore, a variety of Gaussian models are used in the experiment to describe the background of multiple different modes. The gray value of the sequence image shall meet the Gaussian distribution of formula (3.7) below. w

$$p(x_t) = \sum_{i=1}^{k} w_{it}(x_t, g_{it}, \sum{}_{it}) \tag{3.7}$$

where, x_t is the pixel value at time t. w_{it}, g_{it}, and $\sum_{it}$ are the weight, mean, and covariance of the i-the Gaussian distribution at time t. The process of background difference method based on mixed Gaussian

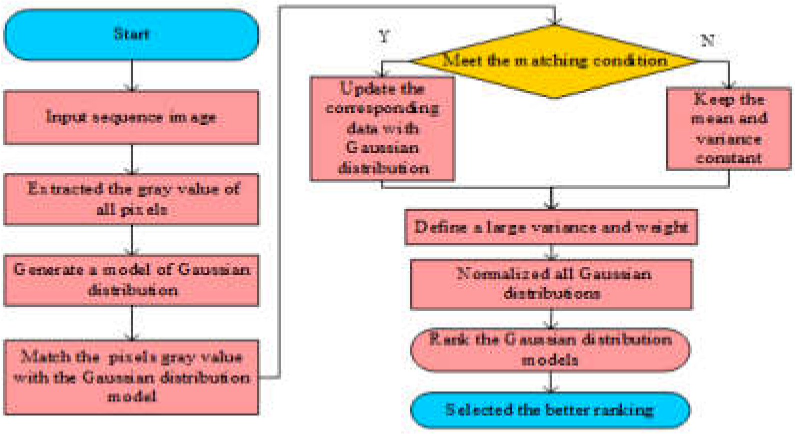

Start
Input sequence image
Extracted the gray value of all pixels
Generate a model of Gaussian distribution
Match the pixels gray value with the Gaussian distribution model
Y
N
Meet the matching condition
Update the corresponding data with Gaussian distribution
Keep the mean and variance constant
Define a large variance and weight
Normalized all Gaussian distributions
Rank the Gaussian distribution models
Selected the better ranking

Fig. 3.2: Process of background difference method based on mixed Gaussian background modeling

background modeling is as follows: the sequence image needs to be input to extract the gray value of all pixels. Then the Gaussian distribution model is generated, and the gray value of the matching pixel is matched with the Gaussian distribution model. The update that meets the matching conditions corresponds to the data of the first Gaussian distribution. Otherwise, it needs to keep the mean and variance of the Gaussian distribution corresponding to the pixel unchanged, and then define a larger variance and weight. Then all the Gaussian distributions are normalized, and the Gaussian distribution models are sorted, and the better one is selected as the background model.

After the moving target in the sequence image is detected, the next step is to identify whether it is a pedestrian from the detected moving target. First of all, it needs to extract the edge of the detected binary image. Combining the scene of TF statistics, the Canny algorithm is selected for edge extraction. The Hough detection circle algorithm is used for head recognition in this study since the pedestrian head information in the monitoring area is approximately circular when the camera collects video images.

The main reason for choosing the Canny algorithm is its excellent edge detection performance, especially robustness in noisy environments [20-21]. The Canny algorithm reduces the effect of noise on edge detection by smoothing the image using a multi-stage algorithm. First, it smoothes the image using a Gaussian filter to filter out the noise. The potential edges are then found by calculating the gradient strength and direction of the image. Next, these edges are refined using non-maximal suppression (NMS) technique [22-23]. Finally, the Canny algorithm is able to effectively differentiate between true and false edges with the dual-thresholding algorithm and edge-connection technique. This series of fine-tuning steps gives the Canny algorithm a significant advantage in the accuracy of edge detection and the precision of edge localization. In this study, the clear and accurate head contour features extracted by the Canny algorithm lay a solid foundation for subsequent head recognition and target tracking, thus improving the overall system performance.

In the surveillance area, the head target can be effectively identified using the Hough transform when the camera captures the video image since the pedestrian head information usually presents an approximate circle. The core of the Hough transform lies in the parameter estimation, which realizes the double mapping from pixel coordinates to parameter space by converting the coordinate problem in the image into a point and line problem in the parametric coordinate system. However, it is more difficult to find complete circles in practical detections, so this study combines the geometrical properties of circles to make improvements. First, the radius threshold of head-like circles is set to exclude the maximum and minimum boundaries of the image. Second, all the points on each edge need to be searched. With these improvements, the Hough transform is able to exclude non-human targets from the moving targets recognized by the background difference method based on hybrid Gaussian background modelling, thus improving the accuracy of pedestrian recognition.

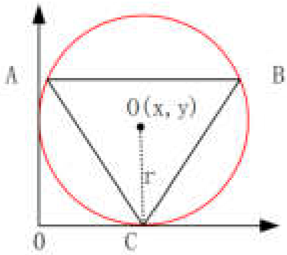


Fig. 3.3: Arc division diagram

When improving the Hough transform, it is necessary to use the idea of arc division to enhance the geometric features of incomplete circles to accurately calculate the coordinates of the centre of the circle and the radius. The structure diagram of arc division is shown in Figure 3.3. Firstly, the target boundary is equally divided into three arc segments AB, BC, and AC. This step is based on the geometrical properties of circles, and the maximum and minimum boundaries in the image are excluded by setting the radius threshold for head-like circles to ensure that a target close to the real head shape is detected. Next, all points on each arc segment need to be detected one by one to compute the centre coordinates and radius of the circle that can be determined through these three points. The advantage of the arc division method is that it can more accurately identify and label the circular target of a traveling human head by accurately calculating the features of each arc segment. This method can effectively identify the target even when the targets are partially overlapped, thus avoiding the detection omission due to the overlapping of images. This arc segmentation and computation method enables the detection algorithm to effectively exclude non-human targets in complex surveillance environments. It significantly improves the accuracy of pedestrian head recognition and provides reliable data support for subsequent target tracking and pedestrian traffic statistics.

3.3. Tracking method of head contour feature frame matching based on KF.. The background difference method based on mixed Gaussian background modeling can recognize the moving objects in the image. Canny algorithm and improved Hough detection circle algorithm can recognize the head objects and determine the detected moving objects as pedestrians [24-25]. The pedestrian is a moving target, so it is necessary to estimate and predict the direction of the moving target for tracking and matching.

In the research on pedestrian target tracking and matching, KF is selected as the core algorithm, mainly based on its excellent prediction and estimation capabilities. KF is a recursive algorithm that can effectively handle the state estimation problem of dynamic systems and is particularly well-suited to deal with noisy signal data. In pedestrian head contour feature matching tracking, KF can utilize the dynamic changes of pedestrian head contour features to provide accurate target state prediction through iterative calculations. In addition, KF does not need to save multiple previous input signals, so it takes up less memory and is very suitable for scenarios that require real-time processing of large amounts of data. Pedestrian targets can be effectively tracked and matched by using the edge detection results of pedestrian head contour features directly as the input of KF prediction, thereby improving the performance and efficiency of the entire system. The basic idea is to estimate the signal trend by recursion based on the characteristics and change rules of the dynamic signal itself. The algorithm has two steps: prior state prediction and observation correction. The equation of state is shown in formula (3.8) [26-27].

$$x_t = \alpha x_{t-1} + \beta s_{t-1} + z_{t-1} \tag{3.8}$$

where x_t is the state at time t. α and β are coefficients. s_{t-1} and $z_{t-1}$are the system control vector and process noise at $t-1$ time. Formula (3.9) is the observation equation. w

$$G_t = \delta x_t + v_t \tag{3.9}$$

where G_t is the observation variable at time t, G_t is the coefficient, and v_t is the observation noise. According to the state matrix and state estimation matrix of the current data, KF calculates the minimum mean square error

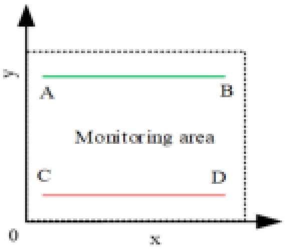


Fig. 3.4: Schematic diagram of tracking area division

matrix in the prediction matrix and corrects it. The minimum error after correction is the optimal prediction result. Formula (3.10) is the state prediction equation.

$$\hat{X_t}^{-} = \alpha \hat{X}_{t-1} + \beta s_{t-1} \tag{3.10}$$

where $\hat{X_t}^{-}$ is the optimal state prediction of the state variable. The covariance equation corresponding to the prediction matrix is shown in formula (3.11).

$$P_t^{-} = \alpha p_{t-1} \alpha_t^T + Q \tag{3.11}$$

where P_{t-1} is the $(t-1) * (t-1)$ covariance matrix, and p_{t-1} is the $p * p$-dimensional symmetric non-negative definite variance matrix. To make the prediction equation the best prediction result, the error gain can be minimized, and the error gain is formula (3.12).

$$K_t = P_t H^T R^{-1} \tag{3.12}$$

where H is the coefficient and R is the symmetric positive definite variance matrix. P_t is the covariance of the optimal estimate at the time of t as shown in formula (3.13). w

$$P_t = P_t^{-1} + HR^{-1}H^T? \tag{3.13}$$

The optimal estimation equation at the time of t is recorded as $\hat{X}_t$, and the calculation process of $\hat{X}_t$ is shown in formula (3.14).

$$\hat{X}_t = \hat{X_t}^{-} + K_t(z_t - H\hat{X_t}^{-}) \tag{3.14}$$

where K_t is the weight of the estimated value at the last time and the current measured value. KF prediction is also the inter-frame matching of head contour features. Its main function is data association, and then real-time update of target tracking chain. In the scope of photography, there are two most common conditions for pedestrians, one is ordinary access, the other is forward, backward, stop, and turn back. In these two cases, the state of the target tracking chain is: the establishment of the target tracking chain, the matching of the target tracking chain, the completion of the target tracking chain, and the disappearance of the target tracking chain. The monitoring area is divided into the tracking area as shown in Figure 3.4, and two entry and exit sign lines AB and BC are set. The positive direction of the y-axis is defined as in, and the negative direction is defined as out.

When a moving target chain passes AB sign line first and then BC sign line, the pedestrian is considered to enter. When the moving target chain passes through the BC marker line first and then AB marker line, it is considered as outflow. If there is a moving target chain entering and leaving from AB or BC, it is considered that the number of entry and exit remains unchanged. To sum up, the main flow chart of updating the target tracking chain is shown in Figure 3.5. After inputting the sequence image, it is necessary to recognize the head target, and then match the head with the pedestrian. The head feature contour data of the next frame is predicted by KF, and the target tracking chain is updated according to the optimal prediction result.

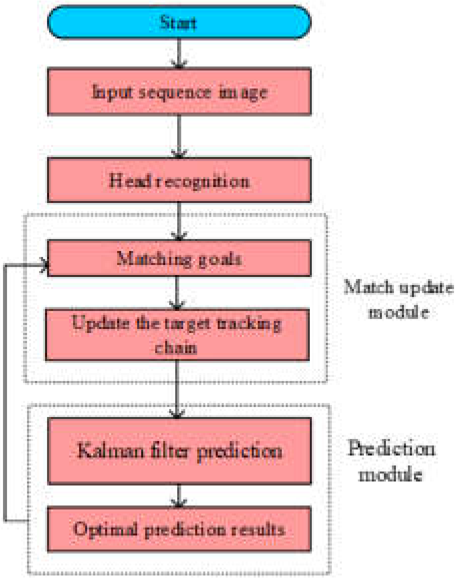


Fig. 3.5: Update moving target chain flowchart

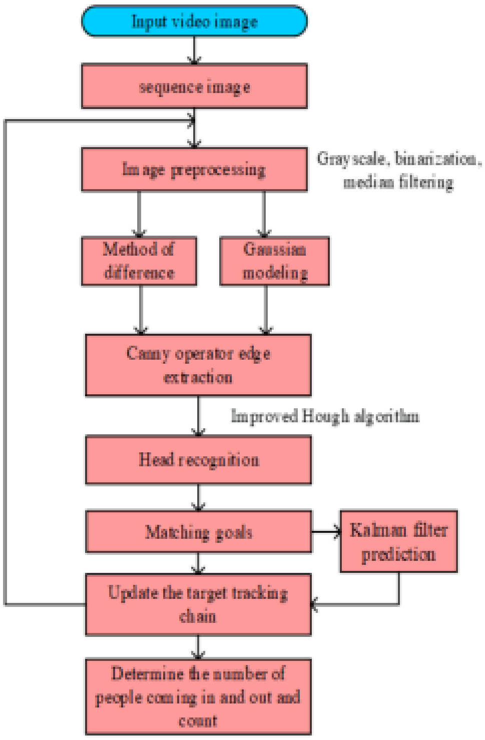
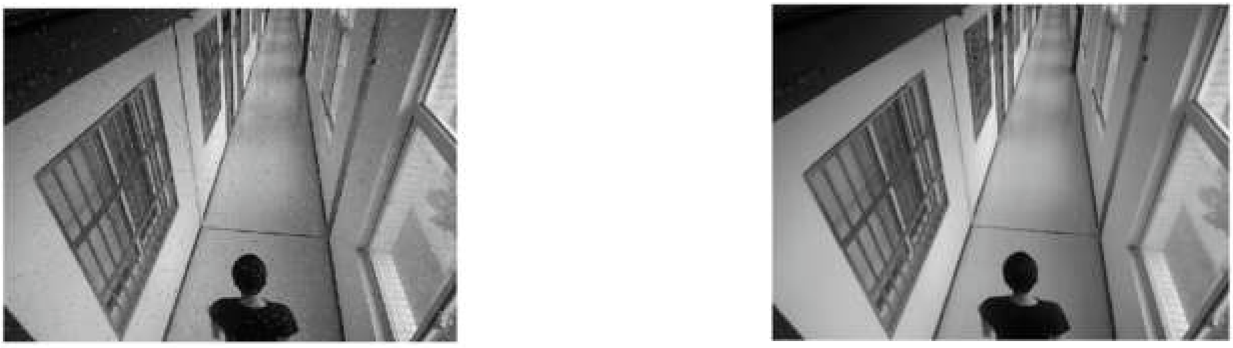


Fig. 3.6: Scenic spot human flow detection system framework



(a) Grayscale Image with Noise

(b) The filtered Grayscale Image

Fig. 4.2: Noise removal effect picture

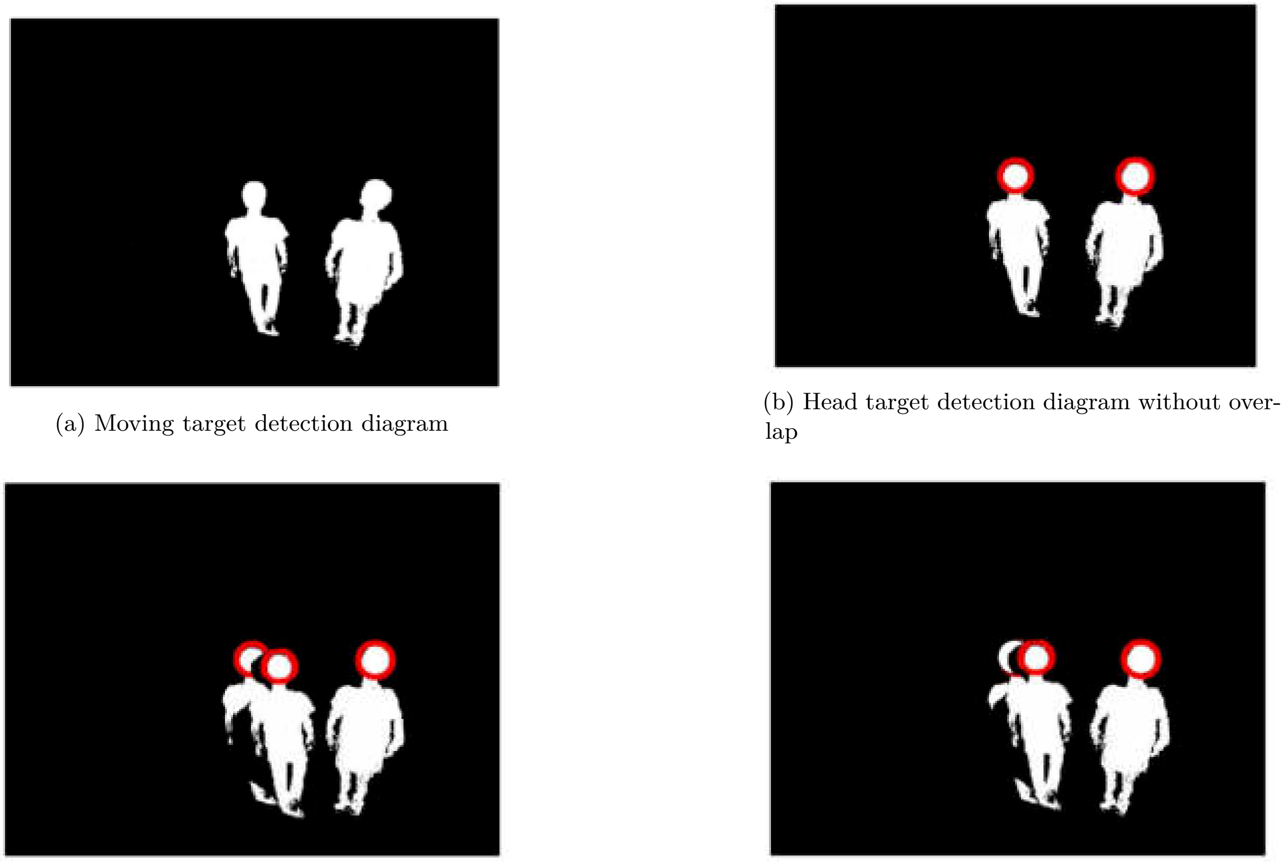
The following passenger flow statistics system architecture is designed to realize the function of expected passenger flow statistics as shown in Figure 3.6. The architecture includes video image acquisition, and then it needs to be converted into sequence diagram and processed by graying, binarization, and median filtering. The processed image is recognized by the background difference method based on Gaussian modeling. Then the edge feature of the image is extracted by Canny operator, and the head is detected by improved Hough transform. Finally, KF is used to match pedestrian targets and count the number of people entering and exiting the monitoring sign line.

4. Analysis of application results of tourism TSS-TF detection system.. The fixed-focus camera HD720P USB is used for acquisition, and the image data is transmitted to the computer using the signal shielded wire. The image data are stored, processed and displayed by the computer. A high-performance computer is used as the core of the entire image signal analysis. The computer CPU is AMD Sempron (tm) Processor2800+Intel (R) Core (TM) i5-5200U CPU @ 2.20GHz2.2GHz, and the memory is 4G. After the completion of the hardware and software of the passenger flow statistics system, the actual measurement is carried out to check whether the system meets the system requirements. The test software environment is Visual Studio 2015 under Windows 7 operating system.

According to the application scenario, the relevant parameters of the passenger flow statistics system are adjusted and set according to the fixed angle of the camera and the height from the ground before the system runs. The TF at gate s1 and 2 of a TSS is counted in the experiment. The cameras of the No. 1 and No. 2 doors are fixed diagonally above the channel. The height of the installed camera from the ground is 3m, and the fitting radius of the system head is 18.3. Relevant scene videos need to be collected for statistics. In the process of statistics, each step of the system is processed separately, and the error causes are analyzed, and comparative experiments are carried out to verify the system performance. The comparison method is the micro-Doppler human motion detection method based on KF and convolution neural network (CNN) in [20] and the target detection algorithm based on improved Gaussian mixture model in [21].

The comparison of denoising effect between median filtering and other methods is shown in Figure 4.2. The median filter removes the noise of the original image well, the pedestrian target and background are more distinct. And the filtered image is clearer, which can help the subsequent system to better perform the target recognition task.

The result of target detection is shown in Figure 4.4. The images show that the edge detection results of this research method are more complete and clearer. The results of head target recognition are shown in the figure. Under the condition that pedestrian targets do not overlap, the improved Hough transform circle detection method in this study can completely recognize and mark the head circle of pedestrians. The contour feature extraction of pedestrians is also relatively complete. When the pedestrian head is partially overlapped,



(a) Moving target detection diagram

(b) Head target detection diagram without overlap

(c) Head target detection diagram in the case of partial overlap

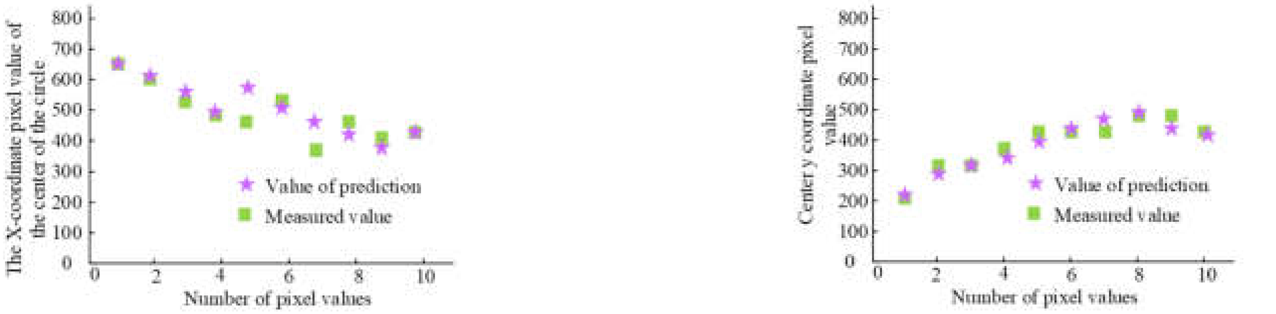
(d) Image of head target detection in most overlapping cases

Fig. 4.4: The result graph of target detection

the head target can also be recognized, which can avoid the problem of missing detection caused by the overlap of pedestrian images to a certain extent. However, when the pedestrian image completely overlaps or mostly overlaps, the method cannot recognize the head target because the overlapping part exceeds the recognition threshold.

Several frames of images are selected to be input into the KF device, and the pixel values of the center position x and y in the image are calculated. As time goes by, the predicted and actual values can be plotted. The x-coordinate trajectory and y-coordinate trajectory as shown in Figure 4.6 are obtained. In the prediction results of the head center coordinate predicted by KF, the error of the prediction of the center x coordinate is greater than that of the prediction of the y coordinate. When the number of pixel values is 5 and 7, the predicted value of x coordinate has a large deviation from the actual value. But in most cases, the prediction is more accurate. The predicted value of y coordinate has no big error all the time, which indicates that KF can better achieve the inter-frame matching of moving objects.

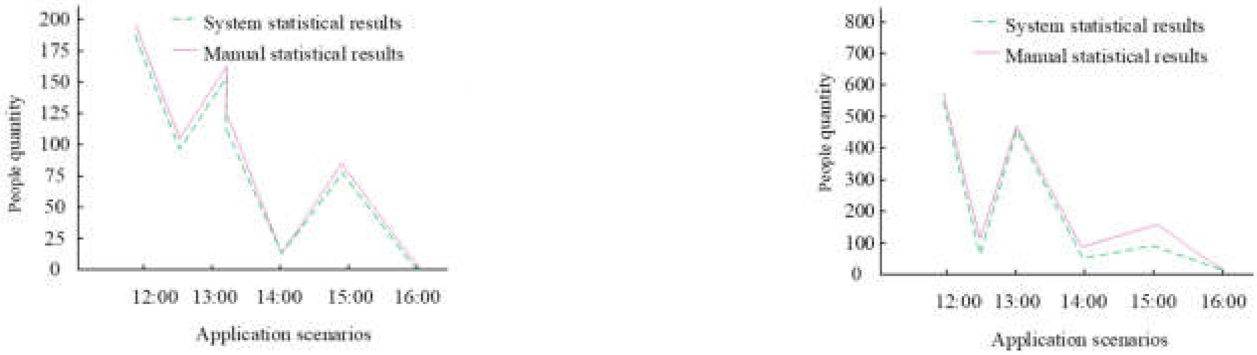
TF statistics at the entrance and exit of Gate 1 and Gate 2 from 12:00 to 16:00 are selected, and the statistical results are shown in Figure 4.8. The statistical results of gates 1 and 2 have missed inspection. However, the error between the TF statistical results of the system and the manual statistical results of gate 1 is slightly smaller than that of gate 2. This may be related to the different size of TF at the entrance and exit. The number of people in and out of gate 2 is far more than that of gate 1. In the case of large crowd density,



(a) Actual and predicted trajectories in X-coordinates of the center of the circle

(b) Actual and predicted trajectories of Y-coordinates of the center of the circle

Fig. 4.6: KF center coordinates predicted trajectory



(a) Passenger flow statistics of Gate 1

(b) Passenger flow statistics of Gate 2

Fig. 4.8: Statistical results of inbound and outbound traffic during 12:00 to 16:00

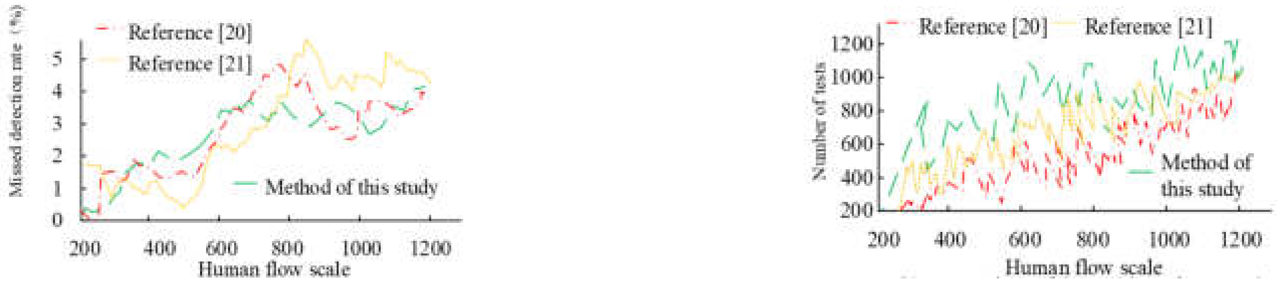
the problem of pedestrian head target occlusion is more serious. This is not conducive to the target recognition of the system, and the phenomenon of missing detection occurs from time to time.

The specific results of TF statistics of 9 sampled videos from two doorways are shown in Table 1. In this study, although the numbers of inflows and outflows are manually calculated, a number of measures are taken to minimize calculation errors to ensure the accuracy of the statistics. Firstly, each counting point is monitored by at least two specially trained observers who are responsible for recording the number of pedestrians passing through a particular entrance and exit. These observers use pre-designed counting equipment to minimize human error. Second, a video surveillance system is used in the study as a secondary validation tool to further validate the accuracy of the data. They can be corrected and validated against each other by comparing the manual counting data with the statistics automatically generated by the video analysis system. Finally, repeated measurements and cross-checks are performed on all data to ensure consistency and reliability. Overall, the error rate of the system's inbound passenger flow is 4.10%, and the error rate of the outbound passenger flow is 3.0%. The outflow statistics error of video number 9 is the largest, 0.3. The inflow statistics error of No. 8 video is the largest, 0.36. The statistical results of inflow and outflow of No. 1 video are both zero errors. The greater the number of people entering and leaving, the greater the statistical error.

Under different TF scales, the statistical results of this research system are compared with the micro-

Table 4.1: Sampling video traffic statistics result

Collect video number	The number of manual inflows is counted	Inflow system count	Inbound passenger flow error rate	The outflow volume is counted manually	System counts the outgoing volume	Error rate of outflow volume
1	22	22	0	5	5	0
2	55	54	0.035	10	10	0
3	111	110	0.036	35	34	0.027
4	80	78	0.005	5	5	0
5	126	120	0.032	6	6	0
6	320	324	0.031	31	30	0.096
7	680	660	0.028	240	236	0.029
8	519	500	0.036	330	319	0.085
9	841	820	0.034	1216	1200	0.03
Total	2760	2677	0.041	1216	1200	0.03



(a) Comparison of missed detection rates under different human flow scales

(b) Comparison of the number of tests under different human flow scales

Fig. 4.10: The statistical comparison results of the human flow of each method

Doppler human motion detection method based on KF and CNN and the target detection algorithm based on improved Gaussian mixture model, as shown in Figure 4.10. With the increase of TF, the number of missed inspections of each method gradually increases. However, the growth rate of the system in this study is slightly smaller than that of other methods, with a maximum miss rate of 4.0% and the minimum of 0. The rate of missed detection of other methods has risen rapidly. Its detection performance is relatively stable.

Single Shot MultiBox Detector (SSD), Mask Region-based Convolutional Neural Network (Mask R-CNN), and Multi-Scale Shared and Independent Feature Network (MSSIF-Net) were selected as the comparison algorithms to further prove that the proposed method in this study has better detection effect. The leakage rate, false detection rate, and accuracy rate of several algorithms are obtained as shown in Table 4.2. The leakage rate, false detection rate and accuracy rate of SSD were 5.96%, 6.11%, and 89.21%, respectively. The leakage rate, false detection rate and accuracy rate of Mask R-CNN were 3.87%, 3.95%, and 92.35%, respectively. The leakage rate, false detection rate and accuracy rate of MSSIF-Net were 2.30%, 2.57%, and 96.54%, respectively. The leakage rate, false detection rate and accuracy rate of the method proposed in the text were 1.25%, 1.34%, and 98.67%, respectively. In conclusion, the method proposed in the text has better detection effect.

5. Conclusions. The accurate prediction of the passenger flow of TSS can enable the relevant departments to efficiently reallocate various resources and public services, so as to organically integrate the resources in the region. For this reason, the TSS-TF statistical system was designed based on KF. The median filter had

Table 4.2: Detection performance of different algorithms

Methods	Leakage rate/%	Misdetection rate/%	Accuracy rate/%
SSD	5.96	6.11	89.21
Mask R-CNN	3.87	3.95	92.35
MSSIF-Net	2.30	2.57	96.54
Text methods	1.25	1.34	98.67

good denoising effect. The improved Hough transform circle detection method in this study could completely recognize and mark the head circle of the traveler without overlapping pedestrian targets. When the pedestrian's head overlapped partially but did not exceed the threshold, the head target could also be recognized. The total error rate of inbound passenger flow of TF statistical system was 4.10%, and the error rate of outbound passenger flow was 3.0%. With the increase of TF, the system in this study gradually raised, with the maximum missed detection rate of 3.2% and the minimum of 0. In the comparison method, the overall detection accuracy of the system was the highest. The system can be used to detect TSS-TF, help TSS control passenger flow, and promote the coordinated development of economic society and TI. However, in this study, human body recognition is based on head feature circle recognition. For objects with irregular shapes such as hair and hat, or objects with circles, the accuracy of human recognition will be reduced, thus affecting the accuracy of statistics. In future research, it is necessary to eliminate the interference of irregular object images.

Fundings. The research is supported by: Xinyang Agriculture and Forestry University Scientific Research Fund Project for Young Teachers (NO.20200214).

REFERENCES

[1] Hcia, E. The role of tourism in the development of the city. *Transportation Research Procedia.* **39** pp. 104-111 (2019)

[2] Aly, S. & Gutub, A. Intelligent recognition system for identifying items and pilgrims. *NED University Journal Of Research.* **15**, 17-23 (2018)

[3] Kim, S., Guy, S., Hillesland, K. & Others Velocity-based modeling of physical interactions in dense crowds. *The Visual Computer.* **31** pp. 541-555 (2015)

[4] Aly, S., AlGhamdi, T., Salim, M. & Others Information gathering schemes for collaborative sensor devices. *Procedia Computer Science.* **32** pp. 1141-1146 (2014)

[5] Sridevi, N. & Meenakshi, M. Efficient reconfigurable architecture for moving object detection with motion compensation. *Indonesian Journal Of Electrical Engineering And Computer Science.* **23**, 802-810 (2021)

[6] Rogowski, M. Monitoring System of tourist traffic (MSTT) for tourists monitoring in mid-mountain national park, SW Poland. *Journal Of Mountain Science.* **17**, 2035-2047 (2020)

[7] Kumaran, N. & Reddy, U. Classification of human activity detection based on an intelligent regression model in video sequences. *IET Image Processing.* **15**, 65-76 (2020)

[8] Lv Q. G, W., Yu, Q., Spatial, L. & Characteristics, T. Identification of scenic flow based on Location Big Data. *Information Technology.* **2022**, 93-97 (0)

[9] Zimoch, M. & Markowska-Kaczmar, U. Human flow recognition using deep networks and vision methods. *Engineering Applications Of Artificial Intelligence.* **104**, 1-10434 (2021)

[10] Du, S., Bahaddad, A., Jin, M. & Model, Z. - A CASE OF ZHANGJIAJIE. *Fractals: An Interdisciplinary Journal On The Complex Geometry Of Nature.* **2022**, 1-22401 (0)

[11] He, L., Gong, J., Wen, K., Wu, C. & Min, Y. New Method Based on Model-Free Adaptive Control Theory and Kalman Filter for Multi-Product Unsteady Flow State Estimation. *Journal Of Energy Resources Technology.* **143**, 2-10 (2021)

[12] Zhang H. F, Z. Fundamental Frequency Wavelet Autocorrelation Detection Method Based on Kalman Filter. *Electronic Design Engineering.* **30**, 77-81 (2022)

[13] Yu, X. & Xia, S. highly adaptable method for GNSS cycle slip detection and repair based on Kalman filter. *Survey Review.* **53** pp. 169-182 (2020)

[14] Yang, J., Xia, Y., Yan, Y., Yan, Y., Sun, L. & Sun, L. Modified Strong Tracking System Identification Method Based on Square Root Center Difference Kalman Filter for Civil Structures. *International Journal Of Structural Stability And Dynamics.* **21**, 1354-1361 (2021)

[15] Zhou, P., Zhang, S., Wen, L., Fu, J. & Wang, H. Kalman Filter-Based Data-Driven Robust Model-Free Adaptive Predictive Control of a Complicated Industrial Process. *IEEE Transactions On Automation Science And Engineering.* **19**, 788-803 (2021)

[16] Zhu, H., Yan, X., Tang H. , Y. & Yuan, F. Moving object detection with Deep CNNs. *IEEE Access.* **8** pp. 29729-29741 (2020)

[17] Kaysi, I., Alshalalfah, B., Shalaby, A. & Others Users' evaluation of rail systems in mass events: case study in Mecca, Saudi Arabia. *Transportation Research Record.* **2350**, 111-118 (2013)

[18] And, M. and collection algorithms for collaborative sensor devices using dynamic cluster heads. *Trends In Applied Sciences Research.* **8**, 55-72 (2013)

[19] Singh, A., Satapathy, S., Roy, A. & Others Ai-based mobile edge computing for iot: Applications, challenges, and future scope. *Arabian Journal For Science And Engineering.* pp. 1-31 (2022)

[20] Saeed, T., Mutashar, S., Abed, A., Mahmuod, H. & Abd-Elghany, S. Human Motion Detection through Wall Based on Micro-Doppler Using Kalman Filter Combined with Convolutional Neural Network. *International Journal Of Intelligent Engineering And Systems.* **12**, 317-327 (2019)

[21] Su, Y. Target detection algorithm and data model optimization based on improved Gaussian mixture model. *Microprocessors And Microsystems.* **81**, 1-10379 (2021)

[22] Roy, P., Saumya, S., Singh, J. & Others Analysis of community question-answering issues via machine learning and deep learning: State-of-the-art review. *CAAI Transactions On Intelligence Technology.* **8**, 95-117 (2023)

[23] Alharthi, N. & Gutub, A. Data visualization to explore improving decision-making within Hajj services. *Scientific Modelling And Research.* **2**, 9-18 (2017)

[24] Curtis, S., Zafar, B., Gutub, A. & Others Right of way: Asymmetric agent interactions in crowds. *The Visual Computer.* **29** pp. 1277-1292 (2013)

[25] Abdelgawad, H., Shalaby, A., Abdulhai, B. & Others Microscopic modeling of large-scale pedestrian-vehicle conflicts in the city of Madinah, Saudi Arabia. *Journal Of Advanced Transportation.* **48**, 507-525 (2014)

[26] Sufi, F., Alsulami, M. & Gutub, A. Automating global threat-maps generation via advancements of news sensors and AI. *Arabian Journal For Science And Engineering.* **48**, 2455-2472 (2023)

[27] Singh, A., Gutub, A., Nayyar, A. & Others Redefining food safety traceability system through blockchain: findings, challenges and open issues. *Multimedia Tools And Applications.* **82**, 21243-21277 (2023)

Edited by: Zhengyi Chai
Special issue on: Data-Driven Optimization Algorithms for Sustainable and Smart City
Received: Oct 18, 2023
Accepted: Mar 1, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 2062–2075, DOI 10.12694/scpe.v25i3.2717

ARIMA-BPNN BASED STOCK PRICE PREDICTION MODEL BASED ON FUSION NEWS SENTIMENT ANALYSIS

XIAOZHE GONG*

Abstract. In recent years, the prediction of stocks has mainly focused on improving and combining stock prediction algorithms, or analyzing news sentiment tendencies to simulate subjective investor consciousness. However, both methods have shortcomings in practicality and comprehensiveness. Therefore, based on the use of stock data, the sentiment propensity of vocabulary in the article was processed, and a new algorithm model was obtained by combining the differential integration moving average autoregressive model and backpropagation feedforward neural network model. Finally, sentiment propensity was integrated into the combination model to obtain an algorithm model that integrates sentiment analysis. After optimizing the sentiment vocabulary of news. The algorithm has improved its ability to recognize emotional tendency words, while traditional algorithms have been improved to improve the accuracy of stock prediction, further verifying the relationship curve between emotional tendency and stock prediction fluctuations. The experimental results show that the combined model of sentiment analysis is close to the true value in predicting stock results, with an error of less than 1.5%. The accuracy and stability of the model's prediction results are significantly better than the uncombined model and traditional prediction models. The new combination model provides better judgment basis for investors through experimental prediction results, creating conditions for investors to avoid stock market risks and improve investment value.

Key words: Stock prediction; News information; Emotional tendencies; Combination model; accuracy

1. Introduction.

As the country's economic strength improves, stock investment has gradually penetrated into ordinary people's hearts. But the variability and uncontrollability of stock market have left many investors at a loss. Therefore, timely prediction of stocks is an important part of achieving control over stock market. At present, the most commonly used methods for stock analysis are basic analysis and technical analysis. The basic analysis method is to use one's own knowledge of economics, finance, and other aspects to analyze stock market through understanding. However, the relative basic analysis is influenced by stock market fluctuations, resulting in a significant deviation in analysis accuracy [2]. Technical analysis is mainly conducted through methods such as historical data and mathematical analysis, which can achieve most predicted results. However, technical analysis also has significant technical limitations and cannot predict stock data from multiple dimensions. Therefore, on the basis of traditional analysis methods, the sentiment analysis section is designed, which combines the differential integration moving average autoregressive model and backpropagation feedforward neural network model to obtain a new combination algorithm model for sentiment analysis fusion, thereby improving the accuracy of stock prediction by the network model. At the same time, the neural network model after sentiment analysis is fused, can further verify the direct relationship between stock market changes and emotional tendencies [4]. There are many current studies on stock prediction, and most of them only improve the prediction model or perform sentiment analysis prediction from the perspective of sentiment tendency, and very few studies that combine the two exist. For this reason, this study adds the new concept of positional weights and punctuation weights, and integrates the sentiment value obtained from sentiment analysis as a feature into the combination model, realising a prediction method that combines numbers and words. This method not only improves the existing theory, but also proposes new directions and research ideas for the research in this field. This study consists of four parts. Firstly, it mainly introduces the research achievements of various experts and scholars at home and abroad. Secondly, the method and structure of building the entire model were introduced. Next, comparative experiments were conducted on the accuracy and feasibility of the model using stock data. Finally, there is a summary of this study and prospects for future research directions.

[*] Advertising Institute, Communication University of China, Beijing, 100024, China (Corresponding author, gongxiaozhe1996@126.com)

Table 1.1: Comparative Analysis of Related Studies

Author Name	Research Title	Research Areas	References
Somesh Yadav et al.	Stock price forecasting and news sentiment analysis model using artificial neural network	Data network, finance, news analysis	References [5]
Zitnik, Slavko et al.	Target-level sentiment analysis for news articles	News sentiment, Internet of Things	References [6]
Mohan B R et al.	Hybrid ARIMA-deep belief network model using PSO for stock price prediction	Internet of Things, Computers, Data Forecasting	References [7]
Colasanto F et al.	AlBERTino for stock price prediction: a Gibbs sampling approach	Computers, Internet of Things, Finance	References [8]
Vara P V et al.	Sree R M. Sruthi, Nishanthi, K.Prediction of Stock Prices Using Statistical and Machine Learning Models: A Comparative Analysis	Statistics, Computers, Stocks	References [9]
Chen Y et al.	Stock Price Forecast Based on CNN-BiLSTM-ECA Model	Statistics, Computers, Stocks	References [10]
Shapiro A H et al.	Measuring news sentiment	Statistics, stocks, news	References [11]

2. Related works. The ever-changing stock market has attracted many investors, and the prediction of stock price fluctuations has also deeply attracted domestic and foreign scholars to explore and research in this field. Somesh Yadav et al. used artificial neural networks (ANNs) to predict stock prices, and studied the impact of past stock trends, daily opening prices, and news sentiment on investors' investment directions. They proposed a model based on ANNs. Artificial neural networks for predicting stock prices is feasible and effective [5]. Zitnik, Slavko, and others created a new article dataset in their research on news and social media information. The dataset was analyzed and evaluated using deep NN algorithms and machine deep learning algorithms. The new dataset performed better than traditional datasets and had significantly higher accuracy [6]. After analyzing the operational planning and sales of enterprises, Shaikh Sahil Ahmed et al. found that the accuracy of stock prediction (SP) for listed companies is an important factor. Therefore, to improve SP accuracy on the foundation of deep NN, the sequence model was enhanced and particle swarm optimization algorithm was used. The improved new model outperforms other models in terms of market prediction accuracy [7]. Colasanto, Francesco and others believe that traditional encoders have some uneven classification problems when classifying language emotions and words. Therefore, a new emotion classification framework is proposed to join the traditional encoder. The new encoder can use emotional orientation to improve SP, and use Monte Carlo method to generate a series of feasible paths. The new encoder can solve classification inequality and SP [8].

Prasad, Venkata Vara et al. compared traditional SP methods and found that among three main algorithms currently used. Kalman filter can consider market fluctuations and greatly improve the accuracy of predictions. XGBoost can collect data and provide a dataset, effectively capturing time series. The main function of Autoregressive Integrated Moving Average Model (ARIMA) is to eliminate stationarity and improve prediction performance. In combination with the advantages of three algorithms, a new algorithm Mixture model is proposed. The new hybrid algorithm's performance is significantly better than other three algorithms, and its clarity is higher [9]. Chen et al. found in their study that SP is a time series problem. However, due to the instability and variability of SP, successfully predicting stock prices has become a highly challenging issue. Therefore, to improve prediction accuracy, they proposed a new model based on convolutional NN and long short-term memory network. The new model utilizes NN characteristics to reduce noise impact in SP. This new model is significantly superior to other traditional models in predicting stock prices [10]. Shapiro A H et al. believe that a new sentiment analysis method has been discovered for stock analysis and prediction. By using this method, some stock data can be predicted to improve the accuracy of stock prediction. The experimental results indicate that this model has better accuracy than existing models and can predict news emotions [11].

In summary, in many experts and scholars' research, SP has low accuracy and precision in prediction algorithms due to its nonlinearity and high volatility. Many traditional algorithms currently in use can only solve a single problem. Therefore, combining multiple NN algorithms on top of traditional algorithms can greatly improve SP accuracy. This experiment uses a combination of ARIMA algorithm and Back Propagation (BP) algorithm to improve the algorithm using sentiment analysis methods, resulting in improved accuracy and prediction accuracy.

3. Method of Stock Price Prediction Based on ARIMA-BPNN Fusion News Emotional Tendency Analysis. This chapter mainly introduces the theoretical methods used, such as web crawler technology, data pre-processing technology, ARIMA model and BP model, to classify and analyze the news words with news emotional tendency. The experiment introduced how to combine multiple methods and apply them to stock price prediction.

3.1. Introduction to the Method of Combining Models Based on Fusion News Emotional Tendency Analysis. Web crawler is a means that can freely grab information data from the World Wide Web, and is a commonly used means to grab network information [12]. Network crawlers first apply for data crawling on the network, then analyze the captured data to extract useful data information, and finally store the obtained data to achieve the purpose of data crawling. When crawling news information, web crawler technology is an indispensable means. Data preprocessing technology is a means of processing and analyzing the obtained data, usually used in large datasets, mainly including filtering and cleaning the data, and deleting pauses in sentences. Since the current study is really a study conducted by investors, it is unrealistic for investors to access many aspects such as weather, economy, company data and many other factors, so in the study only the time series related data predictive analytical model ARIMA model is selected for analysis and prediction. ARIMA is a model that can perform differential classification processing on autoregressive models, with the core content of establishing, testing, and validating methods. By utilizing data that can be crawled, a random sequence is formed and arranged in chronological order. Equation 3.1 is its main parameter.

$$[p, d, q, AR(P), MA(q)] \tag{3.1}$$

In equation 3.1, $AR(P)$ represents the autoregressive model, and $MA(q)$ represents the moving average model. p represents autoregression order, d represents the difference number during data processing, and q represents the moving average order. Equation 3.2 is an explanation of [13].

$$x_t = \sum_{i=1}^{p} \mu_i x_{t-i} + \varepsilon_t \tag{3.2}$$

In equation 3.2, μ_i represents the autoregressive coefficient, ε_t represents noise sequence, and x_t represents the autoregressive process at p order. Equation 3.3 represents the moving average model.

$$x_t = \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} + \varepsilon_t \tag{3.3}$$

In equation 3.3, x_t represents that the interference at different periods is random, θ_i represents the autoregressive coefficient, and represents the noise sequence. When ARIMA model's time series is stationary, then equation 3.4 is its expression formula.

$$x_t = \sum_{n=1}^{q} \theta_n \varepsilon_{t-n} + \sum_{i=1}^{p} \mu_i x_{t-i} + \varepsilon_t \tag{3.4}$$

In equation 3.4, when $q = 0$, its model is represented as an autoregressive model. When $p = 0$, its model is represented as a moving average model. ARIMA model does not need to choose time development due to its own time series. Moreover, its plasticity is strong and can be modified until the model is satisfactory [15].

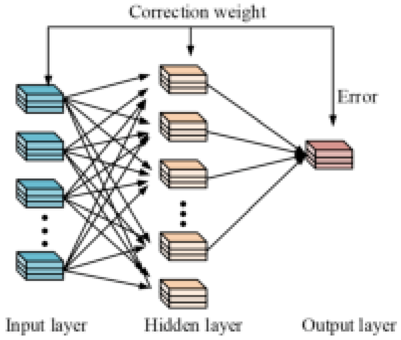


Fig. 3.1: BP Neural Network Backpropagation

Table 3.1: News Information Retrieval Content

/	1	2	3	4	5
Content captured	Financial news broadcast	Individual stock information review	Headline	News hour	Content

And this model can also compare past and present time data, thus possessing better data accuracy. However, the same ARIMA model can only deal with the original data of stable topics, and cannot analyze the data set of nonlinear relations. BPNN is currently a widely used NN, mainly composed of input layer, output layer, and hidden layer. When trained through the algorithm, it mainly propagates through forward and backward in Figure 3.1.

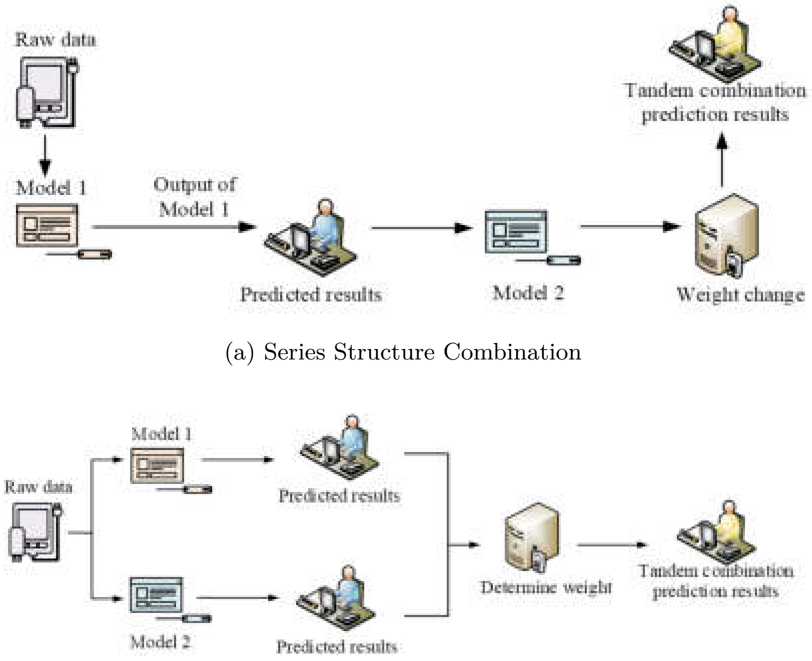
In Figure 3.1, BPNN's backpropagation process is mainly to compare the obtained data results with the real values. And activation function is used to compare the neuron parameters. The neurons gradient is reduced by using loss function, and then the neurons are updated [3]. BPNN is a mapping network for parameter data, with excellent processing ability for nonlinear relationships. BPNN can also process and classify data in many different states. At the same time, it also has the ability to apply the resulting data to another new knowledge. However, due to BPNN structure's large size, there is a tendency for non-convergence. Moreover, when the weight value is too large, BPNN will fall into local extremum, leading to training failure. A combination prediction model is a model that combines two or more models, including series combination and parallel combination. Figure 3.3 shows the series combination and parallel combination diagrams.

In Figure 3.3, the series combination is to transfer the previous model's results to the next model, and the calculated results' transfer is used from the next model to the previous one to calculate the model weight and obtain the results. The parallel combination model predicts data values through multiple models and compares them with actual values. The error is larger, the weight is smaller [4].

3.2. Emotional Analysis Based on News Text Information. The stock market volatility is not only related to the listed company's operation, but also influenced by domestic and foreign policies, news, and other factors. The research on news text information's emotional analysis will greatly improve the model accuracy and stability. Table 3.1 shows the content crawled through web crawlers in news and finance networks.

In Table 3.1, the latest financial content is obtained through the broadcast of financial news content. The current individual stocks' latest data information is obtained through the individual stock's comments. The data is filtered by Headline. And the feelings are analyzed by news time and content. When preprocessing the algorithm, it is necessary to annotate the vocabulary in news and filter out useful information. Some data belongs to junk data, such as advertising, copyright, personalized information, etc., which needs to be filtered.

When analyzing emotions in news information texts, it is necessary to analyze and construct the quality of each individual word [18]. Such as basic emotional vocabulary, financial direction vocabulary, modified vocabulary, etc. Basic emotive words include some daily vocabulary with positive and negative emotions.



(a) Series Structure Combination

(b) Combination of Parallel Results

Fig. 3.3: Two Combination Structures

Financial vocabulary is mainly commonly used in financial field, and modifying vocabulary includes some degree adverbs, negative words, affirmative words, turning words, and so on. When analyzing emotional vocabulary, it is necessary to divide some paragraphs in news, compare and match them with vocabulary in dictionary, and finally determine emotional inclination degree through vocabulary use in equation 3.5.

$$T = \frac{p - N}{p + N} \tag{3.5}$$

In equation 3.5, p represents the positive emotional vocabulary frequency, N represents the negative emotional vocabulary frequency, and T represents the emotional value brought by emotional vocabulary. But traditional calculation formulas cannot calculate article expression's words emotional weight. Therefore, on the basis of the basic formula algorithm, add vocabulary's basic weight value itself, vocabulary location information's weight value, and punctuation position's weight value. By increasing weight, each news vocabulary's emotional orientation can be calculated, and then the article's order and paragraphs can be quantified through emotional orientation. Equation 3.6 is the emotional paragraphs tendency.

$$\text{para} = \frac{\sum_{i=1}^{M} \text{Sent}_i}{M} \tag{3.6}$$

In equation 3.6, M represents the total sentences with emotional tendencies, and Sent_i represents sentences' emotional tendency values. Equation 3.7 is the average emotional tendency value.

$$\text{Con} = \frac{\sum_{i=1}^{Z} \text{para}_i}{Z} \tag{3.7}$$

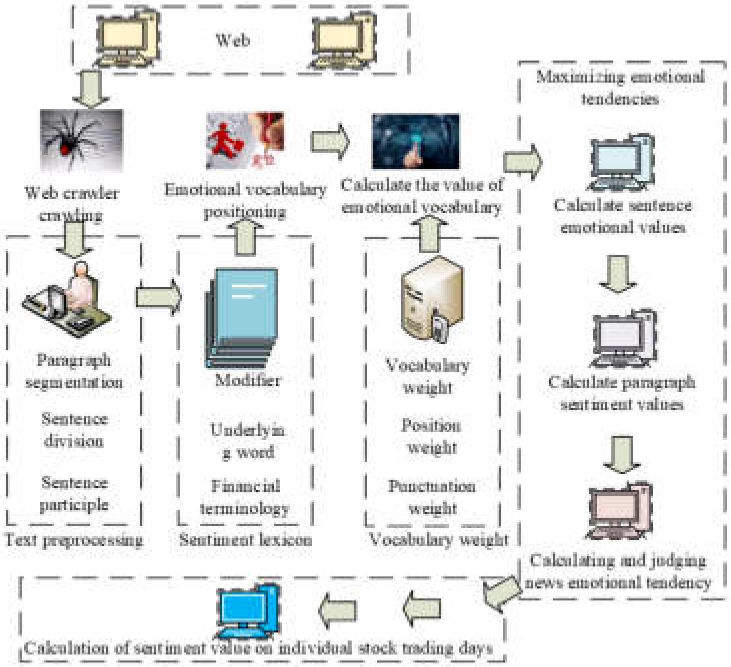

Web
Web crawler crawling
Emotional vocabulary positioning
Calculate the value of emotional vocabulary
Maximizing emotional tendencies
Calculate sentence emotional values
Paragraph segmentation
Sentence division
Sentence participle
Text preprocessing
Modifier
Underlyin g word
Financial terminology
Sentiment lexicon
Vocabulary weight
Position weight
Punctuation weight
Vocabulary weight
Calculate paragraph sentiment values
Calculating and judging news emotional tendency
Calculation of sentiment value on individual stock trading days

Fig. 3.4: Flow Chart of Emotional Tendency Analysis

In equation 3.7, Z represents the sum of all paragraphs. to meet the calculation requirements, it is also necessary to calculate and divide each trading day and trading time in equation 3.8.

$$\text{Every-stock} = \frac{\sum_{i=1}^{S} \text{Con}_i}{S} \tag{3.8}$$

In equation 3.8, S represents the total news that appears on each trading day. Every-stock represents the average cumulative increase in financial news on the trading day. The financial news sentiment tendency analysis chart in Figure 3.4 was obtained through the above calculation formula.

In Figure 3.4, a web crawler is first used to crawl news data, and the captured data is preprocessed for text. A vocabulary of emotional tendencies was constructed and the emotional tendency vocabulary in each financial news was determined. The vocabulary weight has been enhanced, and the emotional vocabulary's emotional value has been calculated. By quantifying emotional vocabulary's tendency value, the financial news's emotional tendency value on the trading day was calculated.

3.3. Methods Based on ARIMA Model and BP Model. During the process of capturing real-time stock data, there may be data loss or garbled code. Therefore, it is necessary to preprocess the data first. Firstly, the captured data should be cleaned up to exclude information such as advertisements and stocks. Secondly, the missing data should be supplemented [21, 22, 23]. Finally, the captured information should be formatted to make it easier for computer recognition. In the stock market, many investors analyze individual stocks through various indicators. RSI is a commonly used indicator for analyzing stocks strength, and its formula is expressed as equation 3.9.

$$\text{RSI}(t) = 100 - \frac{100}{1 + \frac{a}{b}} \tag{3.9}$$

In equation 3.9, a represents an average increase in the closing price on t day, b represents an average decrease in the closing price on t day, and t represents a time cycle. The change rate was calculated by

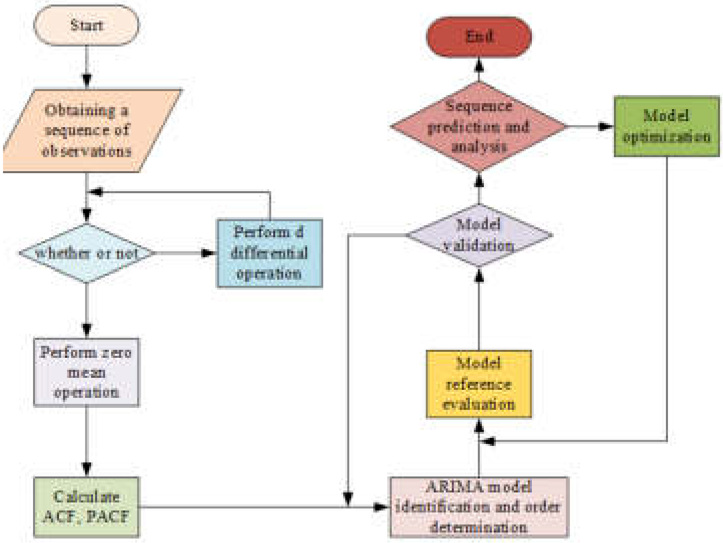

Start
Obtaining a sequence of observations
whether or not
Perform d differential operation
Perform zero mean operation
Calculate ACF, PACF
ARIMA model identification and order determination
Model reference evaluation
Model validation
Sequence prediction and analysis
Model optimization
End

Fig. 3.5: ARIMA flowchart

comparing the current closing price with the past t weeks' price within a given time period in equation 3.10.

$$\mathrm{ROC}(t) = \frac{u-k}{k \cdot c} \tag{3.10}$$

In equation 3.10, u represents the current closing price. k represents the closing price before t days. c represents a fixed value of 100. t represents the calculation parameter. The common stock market will analyze low data characteristics based on the above two indicators, so as to reflect the daily fluctuations of stocks well. ARIMA model in Figure 3.5 was constructed based on stocks characteristics.

In Figure 3.5, stock prices stability was obtained by capturing data from stocks. Through unity value's root test, it can distinguish whether price is stable or not. Mean difference operations were performed on non-stationary time series until data was stable. The stationary time series was subjected to zero mean operation and then fitted to the data. Finally, the model feasibility is tested until it is determined that model parameters are feasible. If it is not feasible, it needs repeat step four [18]. The combination model in series can achieve two models' combination in terms of feasibility. However, the concatenated results cannot separate the linear and nonlinear parts of NN, and cannot achieve mutual superposition effect. Therefore, a parallel combination model was used in this experiment to predict and analyze data in equation 3.11.

$$f(x) = \omega_1 G_1 + \omega_2 G_2 \tag{3.11}$$

In equation 3.11, ω_1 represents a single model weight, G_1 represents ARIMA model's predicted value, and G_2 represents BP model's predicted value. In parallel combination structures, this model is mainly influenced by weight value, which determines its prediction results. Only through reasonable weight calculation can the model be more stable and reasonable. Equation 3.12 represents the average weight.

$$\omega_i = \frac{1}{n} \tag{3.12}$$

In equation 3.12, ω_i represents the weight, and n represents the weight calculating number. According to the error variance in equation 3.13, the weight value can be calculated.

$$\omega_i = \frac{i}{\sum_{i=1}^{n} i} = \frac{2i}{n(n+1)} \tag{3.13}$$

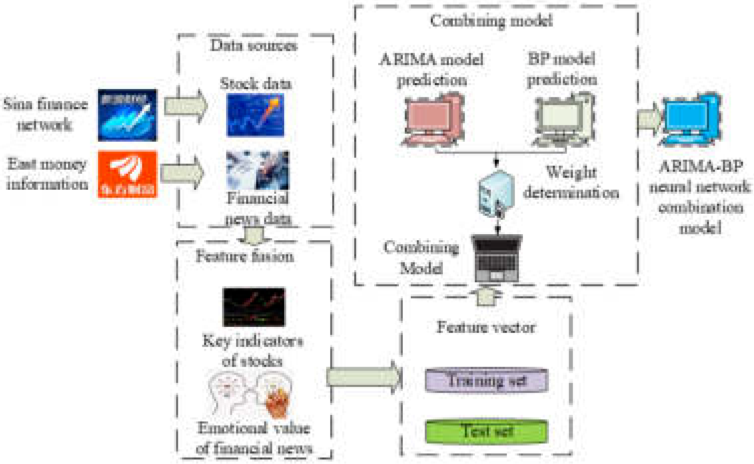


Fig. 3.6: Framework of Fusion Emotional Analysis Combination Model

In equation 3.13, $\sum_{i=1}^{n} \omega_i = 1$, $\omega_i > 0$, $i = 1, 2, 3, \ldots, n$. Calculating the weight value through relative error can reflect model's predictive accuracy. The weight is larger, the error is smaller in equation 3.14.

$$\omega_i = \frac{E_i^{-1}}{\sum_{i=1}^{n} E_i^{-1}} \tag{3.14}$$

In equation 3.14, E_i^{-1} represents the i-th model's relative error, and other parameters are the same as equation 3.13. The mean square error formula compares the true and predicted values to reflect model accuracy in equation 3.15 [16].

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.15}$$

In equation 3.15, n represents the sample number, y_i represents the actual value, and $\hat{y}_i$) represents the predicted value. By calculating model's accuracy stability and error size, the model feasibility for news sentiment analysis and stock price prediction can be determined. Figure 3.6 shows the model prediction diagram fused with ARIMA-BPNN.

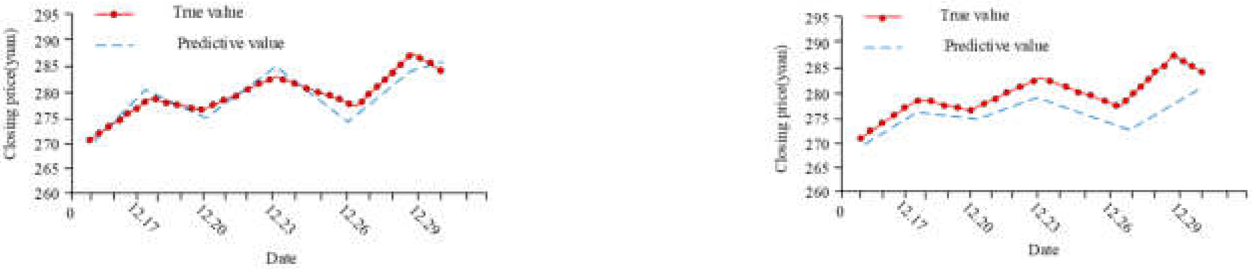
In Figure 3.6, the combined model first preprocesses the collected and captured data. Key indicators that can reflect current day's stock information were selected, such as trading volume, opening information, closing information, etc. Emotional tendencies were judged based on news' emotional information. Two types of information data were subjected to normalization analysis and processing. The data was further divided into training and testing sets based on time. Finally, the weights of the predicted results of the two models are determined to obtain the ARIMA-BPNN combination model.

4. Stock Price Prediction Results Analysis Based on ARIMA-BPNN Fusion News Sentiment Tendency Analysis. The programming language for this experiment is Python 3.8.6, the compiler is PyCharm, the MongoDB database is used for storage, and TXT and CSV documents are used for data preprocessing. Table 4.1 presents an emotional orientation analysis of news data from three listed companies and a comparative experiment on the emotional orientation of the model.

In Table 4.1, the tendency values of traditional and improved vocabulary for Boss Electric are -2.43 and -1.75, respectively. The reason may be that according to news reports, the current boss of Electric Appliances' earnings are not ideal. Hualan Biology and Jiangsu Hengrui Medicine both have positive increases in the

Table 4.1: Results of the Emotional Tendency Comparison Experiment

Stock Name	**Robam**	**Hualan Biology**	**Hengrui Medicine**
Stock code	002508	002007	600276
Headline	Stable profit margin	Complete vaccine release	Approved anti liver cancer treatment
Content	Slowing growth rate	Blood bank restore balance	Benefiting from the approval of new drugs
Improve emotional orientation values	-2.43	3.03	2.93
Traditional emotional tendency value	-1.75	2.67	2.13
/	Emotional tendency value		Basic-exp
R_active	79.24%		73.14%
A_negative	85.23%		81.24%
R_negative	82.14%		74.25%
A_negative	76.21%		69.27%



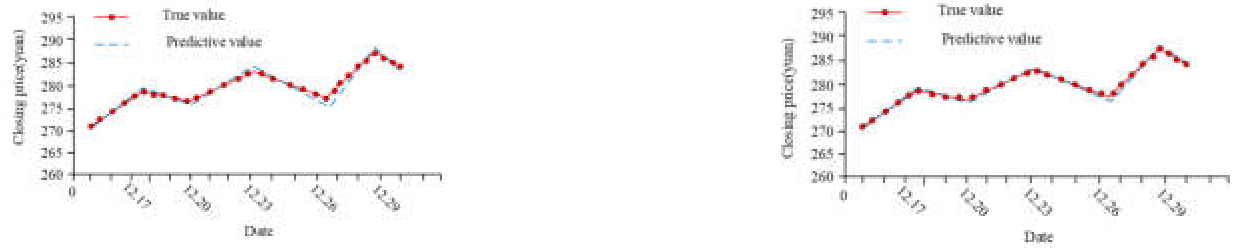
(a) BP neural network predicted value and true value

(b) ARIMA neural network predicted value and true value

Fig. 4.2: Prediction Results of Wuliangye Yibin

improved emotional orientation value and the traditional emotional orientation value. It may be that the news shows that two listed companies' earnings are relatively good, resulting in a higher emotional tendency. Traditional emotion dictionary's calculation value is better than the improved emotion dictionary. A represents accuracy, R represents news recall, and $BAISC_{exp}$ represents emotional tendency comparison. Table 3.1 shows that the recall rate and accuracy rate of traditional emotional orientation comparison are 73.14% and 81.24%, respectively. The combined model's recall rate is 79.24%, its accuracy rate is 85.23%. Two models' difference is 6.1% and 3.99%. The difference in negative recall rate and accuracy rate is more significant, with a difference of 7.86% and 6.94%, respectively. When expressing and conveying different emotions in news, emotional inclination degree varies. It is not comprehensive to only consider vocabulary in the emotional analysis of the news itself. Therefore, optimizing the weight of vocabulary is a necessary condition to increase the accuracy and feasibility of the model. When forecasting the stock data of BPNN and ARIMA models, Wuliangye Yibin Group was selected as SP data from December 15, 2020 to December 30, 2020. Figure 4.2 is the predicted results.

In Figure 4.2, ARIMA's predicted results are roughly similar to the true curve trend. But the error results fluctuate between $\pm 7.4 \sim 2.5$, with significant fluctuations in the upper and lower errors. This suggests that when forecasting stock prices, the use of a single model for price forecasting can result in large deviations from the true value. BPNN's prediction curve results are significantly better than ARIMA, with errors fluctuating between $\pm 2.2 \sim 1.8$ yuan. The fluctuation trend of ARIMA model curve is significantly smaller than the true



(a) ARIMA-BP neural network predicted value and true value

(b) Fusion of emotional analysis ARIMA-BP neural network predicted values and real values

Fig. 4.4: Fusion sentiment analysis ARIMA-BP and ARIMA-BP model prediction results

Table 4.2: Four Model Error Indicators

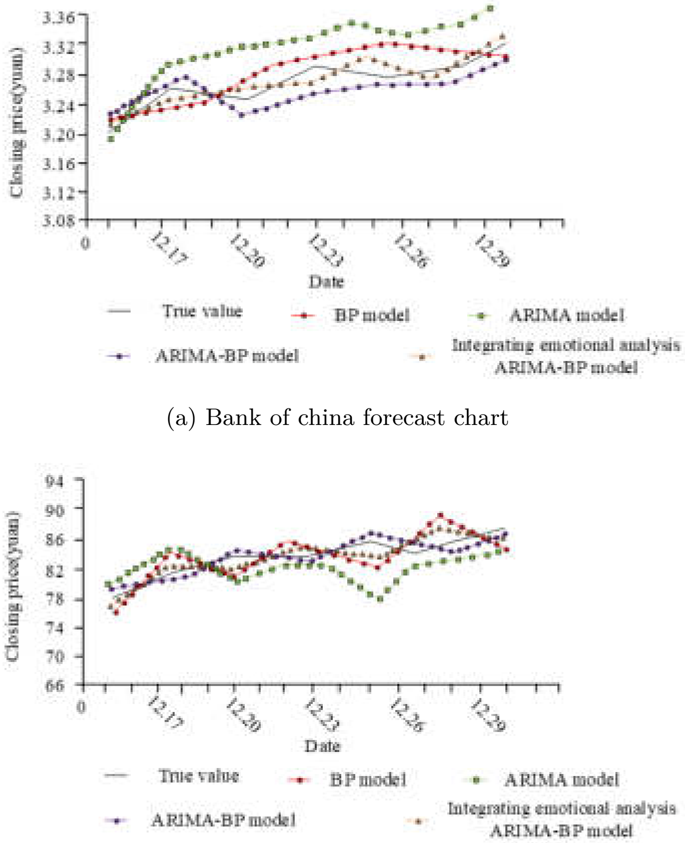
Time	12.15	12.16	12.17	12.18	12.21
True value	270.8	274.33	277.81	276.86	281.91
ARIMA error	-2.56	1.42	2.66	2.98	5.05
BP error	1.2	-1	-0.4	1.14	-1.54
ARIMA-BP error	0.56	-0.59	0.12	1.45	-0.42
Integrating Emotional AnalysisARIMA-BP error	-0.31	1.03	1.18	0.75	-0.12

value. BP not only greatly improves prediction accuracy, but also has a significant improvement in predicting trend changes, Also the BP model has a significant advantage in the predictive effect of the data. After performing algorithmic predictions on both models, a single weight value was reassigned and the ARIMA-BPNN parallel combination model was used to perform algorithmic predictions on data. The combined algorithm model integrated with sentiment analysis was used for data prediction in Figure 4.4.

In Figure 4.4, the combined algorithm model error fluctuates between $\pm 2.1 \sim 0.5$ yuan. Most data prediction errors are between $\pm 1.1 \sim 0.9$ yuan, almost overlapping with true value's fluctuation curve. Its prediction results are significantly better than BPNN and ARIMA models. When predicting the model after fused sentiment analysis, the error value significantly decreased from error $\pm 2.1 \sim 1.5$ to $\pm 2.0 \sim 0.5$ yuan, which is the smallest in four prediction results. The predicted curve is almost consistent with the actual curve, and there are more overlapping parts compared to models that do not integrate sentiment analysis. The model that integrates sentiment analysis outperforms other three models in terms of predictive performance. This suggests that after combining the two models, the model is able to combine the strengths of the two models to improve the processing of the time series, and thus improve the prediction of the data. Table 4.2 compares four models' evaluation indicators.

In Table 4.2, the combination model in which some of Wuliangye Yibin's values are mixed with feelings has the lowest error under relative conditions, and the lowest error was -0.31% on December 15. Each model has the lowest error moment, so it is necessary to further judge whether each model can predict the stock price but is relatively optimal. After combining the two algorithmic models, the new algorithmic model appeared to have fewer prediction errors, which suggests that by combining the two algorithmic models, the prediction accuracy of the model can be improved and the errors appearing in the prediction can be reduced, which may be due to the fact that the model combines the advantages of the two separate models. To test model's ability to deal with different data, the stock data of Bank of China and Juewei Duck Neck are compared and analyzed in Figure 4.6.

In Figure 4.6, regarding the prediction results of Bank of China, among four models, ARIMA model has the largest deviation from the actual curve trend, while other three models have consistent trends with the



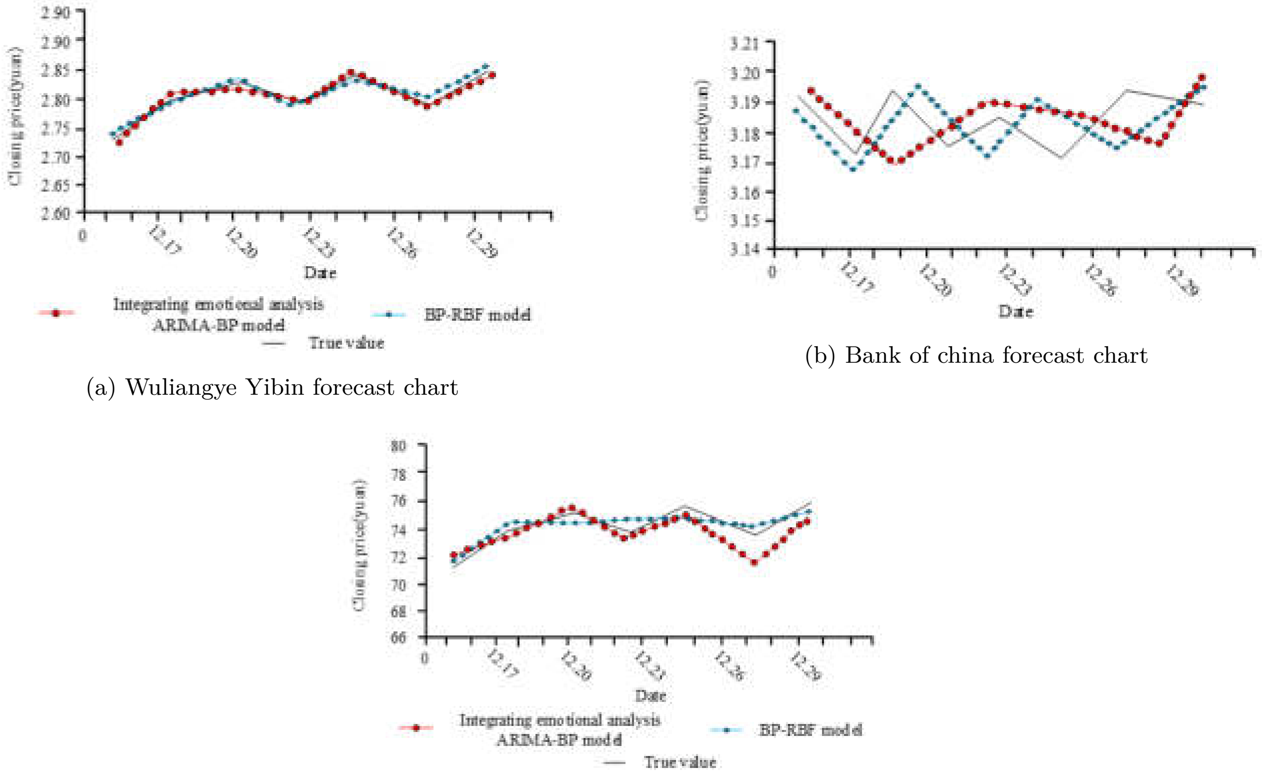
(a) Bank of china forecast chart

(b) Prediction diagram of juewei duck neck

Fig. 4.6: Forecast of Bank of China and Juewei Duck Neck

actual values. The change curve of the combined model that integrates emotional analysis almost overlaps with the true value change curve. When analyzing Juewei Duck Neck's prediction curve, compared with the change trend of Bank of China, four models' prediction curves are almost consistent with the real curve. This shows that four models perform very well in Juewei Duck Neck's data predicting. However, the combination model that integrates emotional analysis still has the highest overlap degree with the true values, and its error change is also the smallest. This suggests that only fusing the models does not lead to better predictions, and with the incorporation of sentiment analysis it is possible to analyse the situation from more perspectives and subjective realities, which can enhance the predictions of the current study. Io further verify the combined model's superiority for integrating sentiment analysis, it was compared with the traditional commonly used BP-RBF combined model in Figure 4.8.

In Figure 4.8, in Wuliangye's stock data forecast curves, two models are consistent with the real results in terms of change trend. But the combination model that integrates emotional analysis has slightly higher overlap than traditional models. When comparing the data of Juewei Duck Neck and Bank of China, it shows a clear gap. The combination model that integrates emotional analysis is more in line with the true value curve, and its change trend is also consistent with the true curve. Therefore, the combination model that integrates sentiment analysis has more advantages in predicting data values, and its performance is also significantly better than traditional algorithm models. In order to compare the predictive effectiveness of different models on the same stock price, Autoregressive Moving Average (ARMA), Generalised Autoregressive Conditional Heteroskedastic-



(a) Wuliangye Yibin forecast chart

(b) Bank of china forecast chart

(c) Prediction diagram of juewei duck neck

Fig. 4.8: Prediction charts of three companies using ARIMA-BP and BP-RBF models fused with sentiment analysis

ity (GARCH), Generalised Autoregressive Conditional Heteroskedasticity (EGARCH), and Graphical Neural Network Models (GNNM) were compared with the models used in the study, and were obtained as shown in Table 4.3.

As can be seen from Table 4.3, in the price prediction of Wuliangye in the first half of December, the price of the ARIMA-BP model used by the study is closer to the true value, and the price prediction results of several of these months are consistent with the true closing price, which indicates that the study uses the algorithmic model with a higher prediction accuracy. At the same time, when several other models were analysed, it was found that the predictions in the other models deviated more from the true results, but there were also a few days where the predictions were consistent with the true results, but the number of occurrences was less. This indicates that the other algorithmic models are also capable of predicting stock prices, but their accuracy is lower.

5. Conclusion. At present, SP has become an important research direction in financial issues, and how to improve SP accuracy and precision is the current research focus. This experiment uses ARIMA algorithm and BP algorithm model to combine, and then uses news information for emotional orientation analysis to conduct stock data analysis. The study optimized the weight of news vocabulary by adding positional and punctuation weights, demonstrating a significant improvement in the accuracy of the emotional orientation values of the current optimized vocabulary weight. Simultaneously, an ARIMA-BP neural network combination model was constructed to predict non-stationary stock data. Finally, the sentiment analysis of financial news texts was integrated into the ARIMA-BP neural network combination model, further confirming that the sentiment

Table 4.3: Comparison of Five Different Models for Predicting Wuliangye Prices

/	Algorithm model	True value	ARMA	GARCH	EGARCH	GNNM	Research Use Model
Price	12.01	262	254	253	254	253	261
	12.02	268	256	255	255	257	267
	12.03	270	268	266	263	267	271
	12.04	276	268	273	270	270	275
	12.05	279	270	271	272	273	279
	12.06	276	272	274	276	271	277
	12.07	275	270	264	268	271	275
	12.08	281	273	270	271	272	280
	12.09	283	276	273	271	278	283
	12.10	286	281	281	280	279	286
	12.11	288	282	281	279	278	287
	12.12	285	276	275	279	274	286
	12.13	291	284	286	287	284	290
	12.14	290	290	276	286	284	290
	12.15	284	276	284	274	284	283

tendency of news texts is positively correlated with stock market volatility. The research results indicate that, When analyzing data of Wuliangye Yibin Group, the improved model's minimum root mean square error is 0.879%, and ARIMA model's maximum variance error is 3.342%. The combined model's average percentage error value for sentiment analysis is the lowest at 0.27%, while the ARIMA model's average percentage error value is the highest at 1.04%. The combination model that integrates sentiment analysis has the lowest error result of only 1.5%. And after analyzing the data of Juewei Duck Neck and Bank of China, the combined algorithm of emotion analysis is the same as the real value in change trend and coincidence degree. Its error is less than 1.5 yuan and 0.05 yuan, which is closer to true value compared to other models. And when compared with the traditional combination model, its performance prediction ability is also better than the traditional combination model. The errors of Wuliangye Yibin, Juewei Duck Neck and Bank of China are less than 0.03 yuan, 3.01 yuan and 0.02 yuan respectively. This study optimized the vocabulary use by designing modifications to the emotional vocabulary of the news, so that the algorithm's recognition ability for the emotionally inclined vocabulary was improved, while the improved algorithm was able to enhance the accuracy of the stock prediction, and further verified the relationship curve between the emotionally inclined and the fluctuation of the stock prediction.The algorithm studied in this study can improve the prediction ability of stocks, but in other aspects, the algorithm is limited by the ARIMA model, which can lead to missing data time series in data processing, thereby reducing the model's effectiveness. The data used in the study is only a part of the stock data, and in subsequent research, it is necessary to analyze the larger stock data. This study only used two models for combination, and more models will be combined in the future. At the same time, the models used in the study can also improve the prediction error of stocks, reducing the prediction error can more accurately predict the current stock price. At the same time, in the research, the model should also consider more advantages of the model and combine more stock factors for model improvement and analysis.

REFERENCES

[1] Ribeiro, G. André Alves Portela Santos, Mariani V C, LDS Coelho. Novel hybrid model based on echo state neural network applied to the prediction of stock price return volatility. *Xpert Systems With Applications.* **184**, 2-14 (2021)

[2] Shapiro, A., Sudhof, M. & Sentiment, W. Journal of Econometrics. (2022)

[3] Gurrib, I. & Kamalov, F. Predicting bitcoin price movements using sentiment analysis: a machine learning approach. *Tudies In Economics And Finance.* **39**, 347-364 (2022)

[4] Fedorova, E., Druchok, S. & Drogovoz, P. Impact of news sentiment and topics on IPO underpricing: US evidence. *Nternational Journal Of Accounting And Information Management.* **30**, 73-94 (2022)

[5] Yadav, S., Suhag, R. & Sriram, K. Stock price forecasting and news sentiment analysis model using artificial neural network. *Nternational Journal Of Business Intelligence And Data Mining.* **19**, 113-133 (2021)

[6] Zitnik, S. & Blagus, N. Baje cM. *Target-level Sentiment Analysis For News Articles.* **249**, 2-15 (2022)

[7] Mohan, B., Ahmed, S. & Kankar, M. Biju R. *Mohan.Hybrid ARIMA-deep Belief Network Model Using PSO For Stock Price Prediction.* **71**, 66-81 (2021)

[8] Colasanto, F., Grilli, L. & Santoro, D. Villani, Giovanni. *AlBERTino For Stock Price Prediction: A Gibbs Sampling Approach.* **597**, 341-357 (2022)

[9] Vara, P., Srinivas, G., Venkataramana, L., Srinethe, S., Sruthi, S. & Nishanthi, K. of Stock Prices Using Statistical and Machine Learning Models: A Comparative Analysis. *He Computer Journal.* **5**, 1338-1351 (2021)

[10] Chen, Y., Fang, R., Liang, T., Sha, Z., Li, S., Zhou, Y. & Song, H. Stock Price Forecast Based on CNN-BiLSTM-ECA Model. *Cientific Programming.* **2021**, 2-21 (2021)

[11] Shapiro, A., Sudhof, M. & Wilson, D. Measuring news sentiment. *Ournal Of Econometrics.* **228**, 221-243 (2022)

[12] Kumar, C. Hybrid models for intraday stock price forecasting based on artificial neural networks and metaheuristic algorithms. *Attern Recognition Letters.* **147**, 124-133 (2021)

[13] Hanif, R., Mustafa, S., Iqbal, S. & Piracha, S. A study of time series forecasting enrollments using fuzzy interval partitioning method. *Journal Of Computational And Cognitive Engineering.* **2**, 143-149 (2023)

[14] Ribeiro, G. André Alves Portela Santos, Mariani V C, LDS Coelho. Novel hybrid model based on echo state neural network applied to the prediction of stock price return volatility. *Expert Systems With Applications.* **184**, 2-14 (2021)

[15] Rezaei, H., Faaljou, H. & Mansourfar, G. Stock price prediction using deep learning and frequency decomposition. *Expert Systems With Applications.* **2020**, 5 (0)

[16] Gurrib, I. & Kamalov, F. Predicting bitcoin price movements using sentiment analysis: a machine learning approach. *Studies In Economics And Finance.* **39**, 347-364 (2022)

[17] Fedorova, E., Druchok, S. & Drogovoz, P. Impact of news sentiment and topics on IPO underpricing: US evidence. *Nternational Journal Of Accounting And Information Management.* **30**, 73-94 (2022)

[18] Banerjee, A., Dionisio, A., Pradhan, H. & Mahapatra, B. Hunting the quicksilver: Using textual news and causality analysis to predict market volatility. *Nternational Review Of Financial Analysis.* **77**, 2-13 (2021)

[19] Zhang, Y. & Hamori, S. Do news sentiment and the economic uncertainty caused by public health events impact macroeconomic indicators?. *Evidence From A TVP-VAR Decomposition Approach.* **82**, 145-162 (2021)

[20] Ray, P., Ganguli, B. & Chakrabarti, A. Hybrid Approach of Bayesian Structural Time Series with LSTM to Identify the Influence of News Sentiment on Short-Term Forecasting of Stock Price. *IEEE Transactions On Computational Social Systems.* **8**, 1153-1162 (2021)

[21] Vijayalakshmi, B., Ramar, K., Jhanjhi, N., Verma, S., Kaliappan, M., Vijayalakshmi, K., Vimal, S., Kavita & Ghosh, U. An attention-based deep learning model for traffic flow prediction using spatiotemporal features towards sustainable smart city. *International Journal Of Communication Systems.* **34**, e4609 (2021)

[22] Nanglia, S., Ahmad, M., Khan, F. & Jhanjhi, N. An enhanced Predictive heterogeneous ensemble model for breast cancer prediction. *Biomedical Signal Processing And Control.* **72** pp. 103279 (2022)

[23] Lim, M., Abdullah, A., Jhanjhi, N., Khan, M. & Supramaniam, M. Link prediction in time-evolving criminal network with deep reinforcement learning technique. *IEEE Access.* **7** pp. 184797-184807 (2019)

Edited by: Zhengyi Chai
Special issue on: Data-Driven Optimization Algorithms for Sustainable and Smart City
Received: Oct 20, 2023
Accepted: Feb 3, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 2076–2087, DOI 10.12694/scpe.v25i3.2720

TEXT EMOTION CLASSIFICATION SYSTEM INTEGRATING VISUAL COMMUNICATION AND DEEP LEARNING FOR SOCIAL PLATFORM

YAN LIU*

Abstract. With the development of modern information technology, social networks have become an im-portant platform for people to express, and at the same time, a large number of texts have been produced. However, the comment text has the characteristics of randomness and colloquialism in the way of expression, and also contains a lot of non-text data. Therefore, manually analyzing the emotional information in the text will consume a lot of time and the accuracy will be limited. To solve emotion classification, this study proposes a knowledge enhanced double loop emotion classification neural network model with attention mechanism. The study first preprocesses text data using a full sentence vector word vector model, then uses convolutional neural networks to recognize emotions in emoticons and emoticons in the text. Finally, the classification results are integrated using algorithms such as a dual loop sentiment classification neural network with pool-ing layers and attention mechanisms, K-nearest neighbors, and decision trees. The final compre-hensive expression recognition and text recognition results are used to obtain the text sentiment classification results. The experimental data shows that the model proposed in this study has an accuracy of 0.947 in the training set test, which is significantly better than other models. In da-tasets A and B, the accuracy of the research design model was 0.958 and 0.924, and the recall was 0.964 and 0.986, respectively. Compared to the baseline method or existing research models, the values of each indicator were significantly higher. The recall rate is the proportion of instances correctly identified as positive by the model to all actual positive instances, which can reflect the emotional classification performance of the research and design model. The higher the value, the better the performance of the model. In practical applications, the positive review rate of this model is above 0.9, which has obvious advantages compared to other models. This study utilizes deep learning techniques to classify sentiment in comment texts, providing reference for the field of text sentiment classification. In the e-commerce industry, it is possible to identify the emotions in user comments on products, further understand the product situation on the platform, and make targeted planning for product reserves, specifications, and so on.

Key words: Text sentiment classification; Deep learning; Visual communication; Fusion model

1. Introduction. As 5G era emerges, more and more people choose to communicate on social media and network platforms. Internet technology has been integrated into people's daily life and has had a significant impact on production and lifestyle. Due to the large number of Internet users in China, a large number of comment texts have been produced. Sentiment analysis of these massive comment texts can obtain valuable user sentiment information, which contains great value [1-2]. Emotion classification has its goal to extract emotional information from subjective texts and make accurate judgments [3-5]. The emotional classification of comment text usually includes positive and negative, and it can be classified at aspect level, sentence level and text level according to the processing level. As Internet users and comment texts increase, these texts show a huge and rapid growth trend [6-8]. At the same time, the comment text has the characteristics of unstructured and colloquial, and also contains elements such as network buzzwords and expression packs. Therefore, manually analyzing the emotional information in these massive texts will be time-consuming and limited accuracy. In the process of e-commerce operation, emotional recognition of user comments is crucial. Merchants analyze the sentiment of massive comment text data to understand user preferences and the quality of their products. This is of great significance for merchants to further improve their products and predict the market with targeted measures. To better understand the emotions of user comments and avoid the problem of inaccurate recognition caused by colloquial and image style comments, a new text emotion recognition model has been studied and constructed. The research contribution includes proposing a knowledge enhanced dual loop sentiment classification neural network model, which effectively improves the accuracy of text sentiment classification by adding attention mechanisms. At the same time, combining visual communication technology, emotion recognition is performed on emoticons and emoticons in the text, further enriching the dimension

* School of Arts and Design, Henan Institute of Technology, Xinxiang, 453003, China (liuyan@hait.edu.cn)

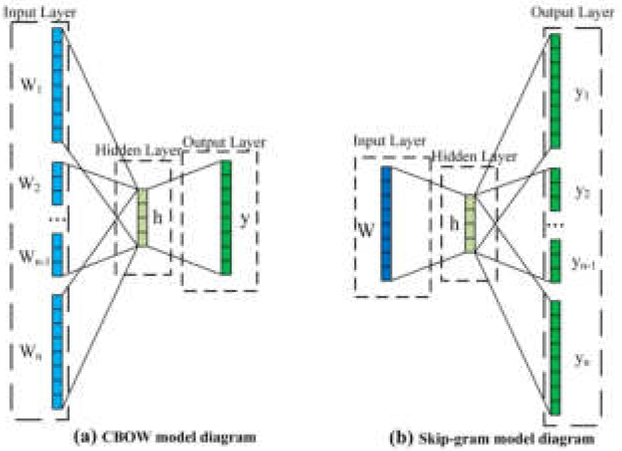
of emotion classification. And a global vector word vector model was adopted for text data preprocessing, effectively reducing the dimensionality of text data and improving computational efficiency. This study aims to use deep learning related technologies to classify the emotion of comment text, Knowledge Enhancement Of Double Loop Emotion Classification Neural Network By Adding Attention Mechanism (KEECA) based on Enhanced Representation Through Knowledge Integration (ERNIE) and Bidirectional Recurrent Neural Network (BiGRU) is proposed, And it is tested on different data sets, and finally applied in the actual sales of goods, which provides new ideas and references for text emotion classification.

The study includes four parts. The second part is text emotion classification research summary. The third part proposes a knowledge enhanced double cycle emotion classification neural network model with attention mechanism. The first section preprocesses the text, the second section identifies the emotion of the expression package, and the third section establishes the classification model. The fourth part is a comparative experiment between the two models. The last part is results.

2. Related work. Benefiting from the continuous maturity, more and more natural language processing fields have begun to adopt deep learning models, especially in emotion classification tasks. Guo y et al. proposed a detection method, using the pre-trained Bert model and machine learning algorithm for classification. The evaluation results showed that it had the best classification performance on two data sets [9]. Context free word scoring, proposed by Nimrah et al., served as an effective alternative method for querying. Results revealed that such approximations were highly efficient in attacking black box neural networks [10]. Anuratha et al. proposed a syntactic senti-rule predictive classifier based on the social spider Lex feature set model to improve accuracy. Results showed that the classifier was superior to other emotion classification models, reaching 94.1% performance score [11]. Wang et al. proposed an ATT algorithm and improved converter bidirectional encoder representation (BERT), which was used to solve the redundancy and noise problems in long text sentiment analysis. Compared with typical convolutional neural model, the algorithm significantly improved the relevant evaluation indicators, and exceeded original Bert model [12]. Wang and others compared the performance of traditional machine learning method in financial text classification, using LSTM as deep learning method and xgboost as traditional machine learning method. Research results showed that LSTM model is superior to traditional machine learning methods in all indicators [13]. Arya et al. Analyzed the feature selection technology of heterogeneous text data for emotion classification, compared the bag of words, TF-IDF and word2vector technology, and found that TF-IDF performed best in SVM classifier, which played an important role in developing adaptive system model for heterogeneous sources [14].

In terms of emotional analysis, Singh S K and others developed a good emotional analysis (SA) system, which was suitable for data sets in many fields. The system was tested on four different data sets, and it showed that it had better accuracy than the existing technology on social media data sets, which increased by 3%, 1.5%, 1.35% and 4.56% [15]. Bie et al. proposed an end-to-end model that comprehensively used for emotion analysis. The experimental results showed that the model had advantages in using text information and can utilize syntactic and semantic information for emotion analysis [16]. Sahu et al. proposed a framework that combines emotional analysis and hybrid recommendation system to recommend upcoming films. This study combined emotional analysis and recommendation system to provide personalized film recommendation by using the public data of film database [17]. Capecchi et al. Used the data of TripAdvisor platform to study the success factors of Tuscany wine tourism by combining the methods of text mining and emotion analysis. The results of the study identified six success factors, including guide, logistics, wine quality, food quality, complementary tourism activities and landscape historical villages [18]. Han et al. proposed a dual-mode fusion network to overcome the limitations of dynamic utilization of independence and pattern correlation in multimodal emotion analysis. Results verified the significant advantages on selected data sets [19]. Kota et al. introduced a method combining CNN, bi-LSTM and attention method for emotion analysis. This method used CNN to reduce complexity. Results showed that it was effective and provided a new emotion analysis method [20].

In summary, sentiment recognition in existing literature can be divided into aspect level sentiment classification and document level sentiment classification. Although aspect level sentiment classification can pay attention to sentiment tendencies in text at a fine-grained level, it is prone to ignoring the interaction between aspect words and context. Therefore, research adopts a dual loop approach to further enhance the connection



(a) CBOW model diagram

(b) Skip-gram model diagram

Fig. 3.1: CBOW and skip gram model diagram

between global features and local features. Document level sentiment classification mainly uses static word vectors for modeling, which will lead to words with similar context but different sentiment polarities being mapped to adjacent positions, further affecting classification performance. It introduces attention mechanism into text feature extraction, which can further improve classification performance. In this study, attention mechanism will be considered for application. Based on this, this study combines the advantages of the ERNIE and BiGRU models, and introduces attention mechanisms to propose a more effective KEECA model.

3. Text classification emotion model construction based on visual communication and deep learning algorithm. In the text classification, the initial text needs to be preprocessed, and the convolutional neural network (CNN) is used to classify text expression package emotion. Then, the processed text data set is trained according to the deep learning algorithm, and finally the text emotion classification results are output by the trained model.

3.1. Preprocessing and vectorization of text sentiment classification. The early text quantification method is one hot coding, but with the increase of text complexity, the number of words to be labeled becomes more and more, and One-Hot method is difficult to meet the requirements. Word2vec is a distributed representation method, which maps words to vector space and distinguishes word semantics by word vector distance and vector space region, making up for the deficiency of One-Hot coding. Word2vec has Continuous bag-of-words model (CBOW) and Skip-gram model[21-22]. Their structure is shown in Figure 3.1.

The CBOW uses context to predict words. The CBOW model is shown in Figure 1 (a). The input layer is vector obtained by one hot encoding, which is multiplied by weight matrix w in hidden layer to obtain corresponding vectorized representation. Finally, hidden layer output vector is obtained by summing and averaging it, as shown in formula (3.1).

$$h = \frac{1}{n} W^T (W_1 + W_2 + ... + W_n) \tag{3.1}$$

In formula (3.1),h is hidden layer output vector, $W^T$is weight,n is number of hot coded vectors,$W_1 ... W_n$ is the hot coded vector. The output vector is calculated in the output layer, as shown in formula (3.2).

$$y = soft\max(W^T h) \tag{3.2}$$

In formula (3.2), y is the output vector after normalizing the product of $h$andW^T. The structure of Skip-gram is opposite to that of CBOW. It predicts the context in the window according to the head word. Its

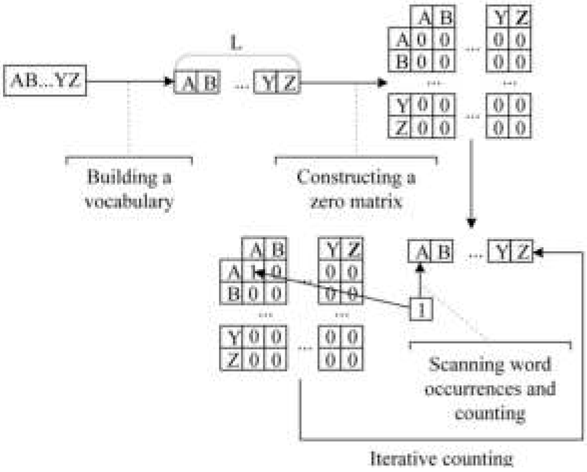

AB...YZ
L
Building a vocabulary
Constructing a zero matrix
Scanning word occurrences and counting
Iterative counting

Fig. 3.2: Flowchart of the glove word model

structure is shown in Figure 3.1 (b) and formula (3.3).

$$P(c|w)\frac{e^{u_c\cdot u_w}}{\sum e^{u}c'^{u_w}} \tag{3.3}$$

In formula (3.3), P is conditional probability, c is central word, w is the context word, $u_w$and u_c is the vectorized representation of w and c, U is the collection of all context words, and c' represents a word in the collection U. In the skip gram model, each word acts as a central word and a context word. Greater$P$ shows higher similarity between $c$and$w$, which means that they have similar meanings. The global vectors for word representation (GloVe) word vector model for word representation incorporates statistical information based on word2vee, which can reflect the co-occurrence of words in the context before and after, and more accurately express the global meaning[23]. The process of GloVe word model is shown in Figure 3.2.

As can be seen from Figure 3.2, column J in row I represents the co-occurrence times of words w_i within the window with words w_j as the center word, and x_i represents the total number of occurrences of any word within the window with words as the center word, then the co-occurrence probability is shown in formula (3.4).

$$P_{i,j}{=}P(w_i|w_j)={\frac{x_{i,j}}{x_i}} \tag{3.4}$$

When there is a given wordw_k, formula (3.5) is used to judge its correlation with w_i and w_j.

$$F(w_i,w_j,w_k)=\frac{P_{\text{ik}}}{P_{\text{ik}}} \tag{3.5}$$

If the value F is large, the correlation between the expression w_k and w_i is high; If the value is small, the correlation between the expression w_k and w_j is high; If the value is close to 1, the indicator$w_k$may be associated with both w_i and w_j, or not associated with them.

3.2. Visual communication emotion recognition based on CNN. Visual communication is a behavior of using visual means to actively disseminate information. It transforms a single text into an attractive image, thus providing a variety of information access for the audience. As a new way of communication, expression pack has been widely popular on the Internet. Emoticon is a kind of non-verbal language expression in modern network communication, which is used to convey the image symbol of emotion. CNN is used as a tool to combine feature extraction of expression package and emotion classification for an end-to-end network. The specific network structure includes convolution layer, relu activation function, pooling layer, local response normalization (LRN), full connection layer and softmax layer[24-25]. When extracting features of expression packs, CNN performs convolution operations by using convolution kernels to extract features, as shown in formula (3.6).

$$x_j^k = f\left(\sum_{i\in R_j} x_i^{k-1} * w_{ij}^k + b_j^k\right) \tag{3.6}$$

In formula (3.6), x_j^k is j convolution neuron value in layer$k$, w_{ij}^k is convolution kernel weight, b_j^k is j offset value in k, R_j is characteristic graphs, and f is activation function. After completing the convolution operation, after simplifying the extracted features in pooling layer, features with the most effective information are input into the classifier for result. The LRN calculation formula is shown in formula (3.7).

$$b_{x,y}^i = a_{x,y}^i/(k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a_{x,y}^j)^2)^\beta \tag{3.7}$$

In formula (3.7), i represents subscript, that is, subscript that calculates the pixel value from 0; j is the square cumulative index, indicating the sum of the squares of pixel values from $j$to$i$; x, y is the coordinate position of the pixel; a is specific value of pixel$i$; N is number of columns of inner vector. k, α, $n/2$, and $\beta$are super parameters specified by the prototype. Softmax improves the calculation speed and accuracy through normalization. Suppose there are 100 different types of pictures, and a 100-dimensional vector is output through the Softmax layer. In the vector, each element denotes the probability of the current image belonging to a specific category, where the summation of all elements equals 1. Its calculation is presented in equation (3.8).

$$f(z_j) = \frac{e^{z_j}}{\sum_{i=1}^n e^{z_j}} \tag{3.8}$$

In formula (3.8), z is the neuron emotion parameter. Because this study identifies seven expressions, and the softmax layer has seven output neurons. The program returns seven probability values, indicating the possibility of neutral, surprise, sadness, happiness, fear, disgust and anger. The maximum probability corresponds to the most likely emotion.

3.3. ERNIE and BiGRU fusion text sentiment classification system with attention mechanism. This study proposed KEECA model by improving ERNIE and BiGRU and adding attention mechanism layer and pooling layer to improve classification by combining advantages of different layers. KEECA structure is shown in Figure 3.3.

In Figure 3.3, $\{E_1, E_2...E_n\}$ represents the original text statement after vector expression in ERNIE layer, and then complete the feature extraction in BiGRU layer, and then input it into the attention layer to express the feature vector according to different weights, that is$\{A_1, A_2...A_n\}$. After the feature extraction is completed in the pooling layer, it goes to softmax layer. When recognizing text emotion, the premise of classification is that each feature of the text sample is independent of each other, and its probability distribution is shown in formula (3.9).

$$P(c|x) = \frac{P(c)P(x|c)}{P(x)} = \frac{P(c)}{P(x)}\prod_{i=1}^d P(x_i|c) \tag{3.9}$$

In formula (3.9), $x$is the text to be classified$X = (x_1, x_2, x_3...x_d)$, $c$is the text category$C = \{c_1, c_2, c_3...c_k\}$, $d$is attributes number of the text sample, $x_i$is attribute value of text sample$i$. The probability value of the text sample$x$ contained in the category$c$ is shown in formula (3.10).

$$c_{nb} = \arg\max_{c\in C} P(c)\prod_{i=1}^d P(x_i|c) \tag{3.10}$$

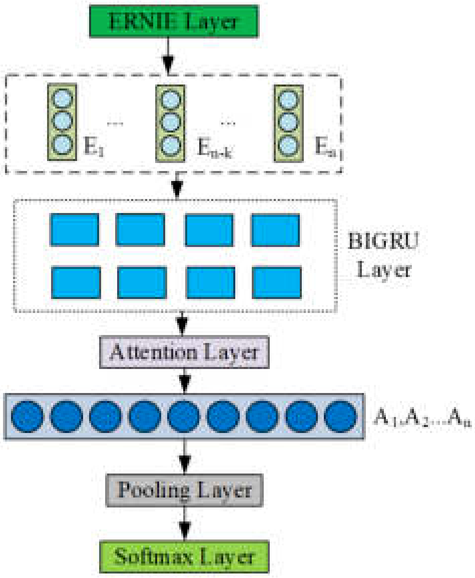


Fig. 3.3: KEECA model structure

In formula (3.10), $P(c)$ is probability that text sample$x$ is included in the category$c$. During training, the class prior probability$P(c)$ and attribute prior probability$P(x_i|c)$ are estimated during training. $P(c)$ is shown in formula (3.11).

$$P(c)=\frac{N_c}{D} \tag{3.11}$$

In formula (3.11), N represents samples number included in category$c$; $D$is samples sum. The calculation method$P(x_i|c)$ is as follows.

$$P(x_i|c)=\frac{N_{cx_i}}{N_c} \tag{3.12}$$

In formula (3.12), N_{cx_i} represents the number of sample $x_i$attribute values contained in the category c and N_c is samples number of c. Tree structure is used in the classification decision. Its structure is shown in Figure 3.4.

Figure 3.4 is the structure diagram of a general decision tree. The first step in the classification process is to build a root node composed of data sets$D$ and divide it into many sub nodes, that is, subsets; The second step is to map the correctly classified subsetD_i at the sub node, and the subset without the correct classification needs to be re divided according to the optimal feature; Finally, the above two operations are recursive until all subsets are correctly classified or have no optional features, and the complete decision tree is established. The calculation process of decision tree establishment is shown in formula (3.13).

$$H(D)=-\sum_{K=1}^{K}\frac{|R_K|}{|D|}\log_2\frac{|R_K|}{|D|} \tag{3.13}$$

In formula (3.13), $H$is empirical entropy function, $D$is dataset,R is category label, $|D|$is number of samples,$|R_k|$ is number of samples in result categoryR_K. For the feature set$T$, the conditional entropy$H$ calculation formula of the data set$D$ is shown.

$$H(D|A)=\sum_{i=1}^{n}\frac{|D_i|}{|D|}H(D_i) \tag{3.14}$$

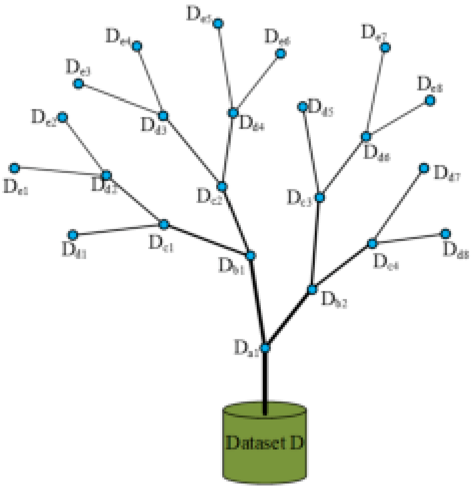


Fig. 3.4: Structure diagram of general decision trees

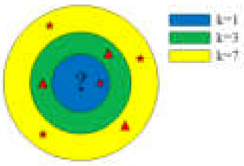


Fig. 3.5: model graph of k-nearest neighbor algorithm

In formula (3.14), A is the feature selection condition andD_i is the subset$i$. According to formulas (3.13) and (3.14), the information gain calculation is obtained in formula (3.15).

$$g(D, A) = H(D) - H(D|A) \tag{3.15}$$

In formula (3.15), g is an information gain function. According to the calculated value$g(D, A)$, a decision tree can be constructed. When finding the sample closest to the given sample in the trained data set, the calculation model is shown in Figure 3.5.

The study utilizes the K-nearest neighbor algorithm in Figure 3.5 to cluster the samples and further classify the text data, laying the foundation for subsequent sentiment recognition classification. The study inputs the training dataset into the K-nearest neighbor algorithm, calculates the distance between each sample and the training sample, and then finds the k nearest neighbors. Next, based on the class labels of these k neighbors, the majority voting principle is adopted to classify the samples. The distance from the processing sample point to each node in the space is calculated by Euclidean distance or cosine similarity, as shown in formula (3.16).

$$\begin{cases} dist(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \\ \cos(x, y) = \frac{x \cdot y}{|x| \times |y|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x^2{}_i \sum_{i=1}^{n} y^2{}_i}} \end{cases} \tag{3.16}$$

In formula (3.16), $x_i, y_i$is point$i$ coordinate in space. After the results are calculated, they are arranged in ascending order by distance. Then, the nearest points$k$ are selected according to the sorting results. The category of most points is the category of pending sample points. When extracting data such as time ordered comment text, it is completed by recurrent neural network (RNN) shown in Figure 3.6.

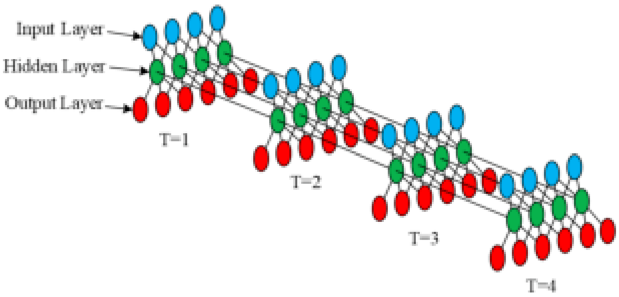


Fig. 3.6: structure diagram of recurrent neural network

Table 4.1: Experimental environment and parameter table for emotional classification of model

Environment		**Hyperparameter list**	
		Parameter	Value
System	CentOS	Epoch_Num	40
		Batch_Size	25
GPU	Titan xp*4	Padding_Size	120
		Hidden_Size	542
Programming language	Python3.6	Filter_Num	425
		Kemel_Size	(3,5,7)
Pre-training model	ERNIE	Dropout	0.2
		LR	1.50E-10

Figure 3.6 is the RNN structure diagram, where the hidden layer result at a certain time T is determined by both the input layer and the previous hidden layer. This study inputs the same dataset into RNN for training. RNN captures long-distance information in text by learning dependency relationships in sequence data, and thus obtains classification results. Hidden state at time$T$ is shown in formula (3.17).

$$h_T = f(W_h h_{T-1} + W_x x_T) \tag{3.17}$$

In formula (3.17), $W_h$is the weight matrix of hidden state,h_{T-1} is output at the time$T-1$, $W_x$is input adjustment weight matrix, $x_T$is the input at the time$T$, $f$is a nonlinear function. After preprocessing with K-nearest neighbor algorithm and RNN, the classification results were obtained. The study combines these two results and uses the weighted average method to integrate the results of the two classifiers to obtain the final classification result. Finally, the classification results are integrated with the emotion classification results of the KEECA model to further improve the accuracy of the model's emotion recognition.

4. Performance evaluation of different text sentiment classification models and empirical analysis of KEECA model. In this study, KEECA, ERNIE, BiGRU, RNN and CNN models were tested in different data sets, and then different models were applied to the analysis of mobile phone sales in an online mall. Finally, the rationality analysis was given according to the performance of different models.

4.1. Performance analysis of Chinese comments based on different text classification models. The environment was CentOS, and data set was a public Chinese comment data set. The emotion labels were divided into positive and negative, in which the positive emotion was marked as 1 and the negative emotion was classified as 0.

Table 4.1 shows the environment and parameters of emotion classification experiment, which aims to achieve the best performance by adjusting the parameters. Epoch_Num indicates the number of iterations, Batch_Size

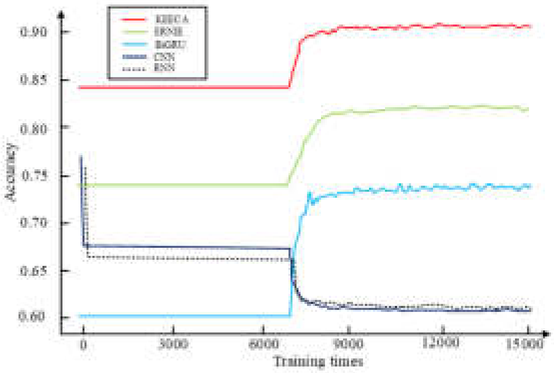
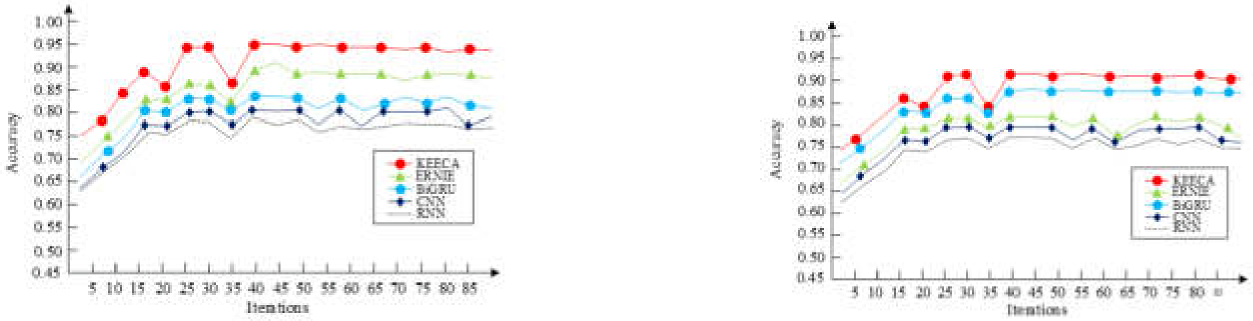


Fig. 4.1: Training performance curves of different models in the training set



(a) Line chart of accuracy rate of each model on M dataset

(b) Line chart of accuracy rate of each model on N dataset

Fig. 4.3: Variation accuracy curve of different model with model training iterations

indicates number of samples input to neural network each time, Padding_Size is the cutting length of the single sentence language in the sample, Hidden_Size is the dimension of the hidden layer, Filter_Num is number of convolution kernels, Kemel_Size is kernel size, Dropout is parameter to prevent over fitting, and LR is the parameter to control the learning rate. The evaluation indexes include accuracy, recall rate and F1 value. The accuracy curve of different models with training times is shown below.

In Figure 4.1, when the training times of all models reached about 7000, the accuracy began to change strongly. The accuracy of KEECA, ERNIE and BiGRU increased as training times increased, while that of CNN and RNN decreased as the increase of training times. When the accuracy was relatively stable, the accuracy of KEECA was 0.947, the accuracy of ERNIE was 0.846, and the accuracy of BiGRU was 0.739. Training performance of fusion model had significant advantages. The accuracies of CNN and RNN were only 0.604 and 0.611 respectively, indicating that the traditional neural network model was not suitable for classifying excessive emotional texts. The change of accuracy of different models in different data sets is shown below.

Fig. 4.3 (a) is the curve of the accuracy of different models on the M data set. All model's accuracy rate tended to be stable after 40 training iterations. The accuracy rate of KEECA model was 0.958, which was 5.1 percentage points higher than ERNIE model, the mainstream text emotion classification model, and 9.7 percentage points higher than BiGRU model. After combining the advantages of different models, the accuracy of KEECA model was greatly improved. The highest accuracy of traditional CNN and RNN model were 0.813 and 0.784, respectively, which are inferior to KEECA, ERNIE and BiGRU models. Fig. 4.3 (b) is the curve of

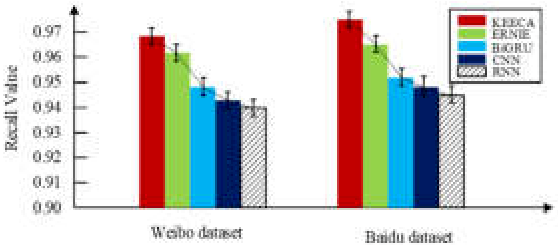


Fig. 4.4: Histograms of recall values for different models

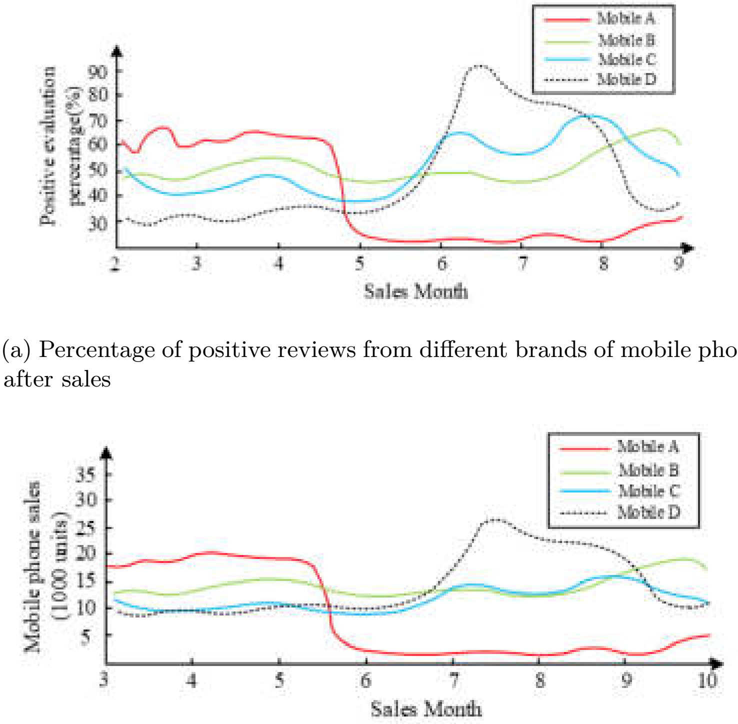
the accuracy of different models on the N data set. The accuracy of KEECA and BiGRU models tended to be stable after 40 training iterations, while the accuracy of ERNIE, CNN and RNN models tended to be relatively stable after 25 training iterations. The highest accuracy of KEECA model was 0.924, the accuracy of BiGRU model was 0.872, and the highest accuracy of ERNIE model was only 0.827. This is because the text of the N dataset was longer than that of the Weibo dataset, and the content was more and more miscellaneous, which made the KEECA model combined with BiGRU and the BiGRU model which had the advantage of long text processing had higher accuracy than other models on the N dataset. The highest accuracy rates of traditional CNN model and RNN model in processing long text were 0.774 and 0.731 respectively, which are also inferior to KEECA, ERNIE and BiGRU models. The recall rates of different models in different data sets are shown in Figure 4.4.

Figure 4.4 shows the recall rates on M dataset and N dataset. KEECA model performed best on the M data set and N data set, with the recall rates of 0.964 and 0.986 respectively, which were 0.007 and 0.013 percentage points higher than ERNIE model, and 0.019 and 0.021 percentage points higher than BiGRU model, indicating that KEECA model had stronger ability to capture emotional information than ERNIE and BiGRU models.

4.2. Empirical analysis of commodity review monitoring based on KEECA model. This study selected the after-sales comment data of four different brands of mobile phones on an e-commerce platform from February to August as the data set, classified the positive and negative comments in the comment text by KEECA model, and then studied the impact of different comments on the daily sales of mobile phones. Figure 4.6 shows the result.

Figure 4.6(a) shows the percentage of after-sales reviews of four mobile phones a, B, C and D in the total reviews from February to August. The positive rate of A mobile phone was around 60% from February to April, with a large decline in May. The percentage of positive comments on Model B mobile phones was generally in the range of 40% to 70%, and its change was relatively gentle. The positive rate curve of model C mobile phones changed mildly before June, increased by about 20% in June, and did not fall until August. The positive rate of model d mobile phone was relatively low before May, and then it continued to rise, reaching a maximum of 90% in mid-June, and then gradually declined. Figure 4.6(b) shows the sales curve of four mobile phones A, B, C and D from March to September. On the whole, sales curve change of mobile phones was consistent with that of the favorable comments in Figure 4.3 (a). The difference is that the change of the sales curve lagged behind that of the favorable comment's percentage curve by one month, which showed that the favorable comments percentage of the mobile phone after-sales in the current month was positively correlated with the sales volume in the following month. The classification experimental data of different models on the mobile phone after-sales comment data set are shown below.

Table 4.2 shows percentage comparison of favorable comments, the accuracy of the classification of favorable comments, and the correlation between sales and the percentage of favorable comments in the data set of mobile phone after-sales reviews by different models. KEECA model had the highest accuracy rate for classification of high praise, and the accuracy rates for a, B, C and D mobile phones were 96%, 97%, 94% and 97%, respectively. The accuracy of BiGRU model and ERNIE model was similar, and the overall accuracy was 90%. However, the accuracy of RNN and CNN models' praise classification was relatively low, and the accuracy of D brand praise classification was very low, only 34% and 37% respectively. From the perspective of the correlation rate between



(a) Percentage of positive reviews from different brands of mobile phones after sales

(b) curve of sales changes for different brands of mobile phones

Fig. 4.6: Percentage of after sales positive reviews and daily sales curve for different brands of mobile phones

Table 4.2: Empirical classification results of mobile phone after sales review datasets using different models

Model name	Positive feedback percentage (%)				Positive classification accuracy (%)				Positive reviews and sales relevance			
Mobile	A	B	C	D	A	B	C	D	A	B	C	D
Keeca	44	53	57	49	96	97	94	97	0.94	0.98	0.97	0.94
RNN	61	52	40	74	70	96	81	34	0.71	0.91	0.76	0.12
Bigru	41	44	54	53	92	84	91	86	0.91	0.74	0.89	0.84
CNN	64	66	71	22	73	76	71	37	0.53	0.61	0.51	0.27
ERNIE	43	56	55	52	92	88	91	87	0.91	0.84	0.93	0.83

sales and positive reviews, KEECA also had obvious advantages over other models, and the correlation rates were above 0.9, indicating that KEECA had a more realistic response to commodity sales than other models.

5. Conclusion. The development of information technology provides technical support for network social media, which provides a platform for free expression for the majority of Internet users. To solve low classification efficiency and poor accuracy in existing emotion classification research and different application fields of emotion classification, KEECA model is proposed. Experiments showed that KEECA had the best performance in the training set, and its accuracy was 0.947. On N data set, KEECA model accuracy was 0.958, which was 5.1% and 9.7% higher than ERNIE model and BiGRU model, respectively. The recall rate of KEECA model was 0.964, which was 0.7% and 1.9% higher than ERNIE and BiGRU models, respectively. On M, KEECA model accuracy was the highest 0.924, which was 9.7% and 5.2% higher than ERNIE model and BiGRU model, respectively. The recall rate of KEECA model was 0.986, which was 1.3% and 2.1% higher than ERNIE and BiGRU models, respectively. In the empirical analysis, KEECA model had the best response to the actual situation of sales. The classification accuracy rates of a, B, C and D mobile phones were 96%, 97%, 94% and 97%, respectively, which were 4%, 9%, 3%, 10% and 4%, 13%, 3% and 11% higher than ERNIE and BiGRU respectively. In terms of correlation rate, KEECA model was 0.03, 0.14, 0.04, 0.11 and 0.03, 0.24, 0.08, 0.1 higher than ERNIE and BiGRU respectively, which means that the description of data set by KEECA model is the closest to the actual situation. In a word, KEECA model can better meet the needs of text sentiment classification and provide important reference value for other application scenarios. However, the model proposed in this study can only classify positive and negative emotions, so we need to improve the richness of emotions.

REFERENCES

[1] Bhuvaneshwari, P. Rao a n. *A Comparative Study On Various Pre-processing Technologies And Deep Learning Algorithms For Text Classification International Journal Of Cloud Computing.* **11**, 61-78 (2022)

[2] Cao, Z. Zhou y, Yang a. *Peng S. Deep Transfer Learning Mechanism For Fine Grained Cross Domain Sentient Classification Con-nection Science.* **33**, 911-928 (2021)

[3] Yang, Z., Li, C. & Zhao, Z. C li. sentient classification based on dependency relationship embedding and attention mechanism Journal of intelligent and fuzzy systems. (0)

[4] Kanan, T. hawashin B, alzubi s, almaita e, Alkhatib a, Maria K. *Elbes M. Improving Arabic Text Classification Using P-stemmer Recent Advances In Computer Science And Communications.* **15**, 404-411 (2022)

[5] Sahoo, R. Mishra b k, DAS B R. *Odia Text Classification Using Naive Bayes Algorithm - An Empirical Study ECS Transactions.* **107**, 8175-8180 (2022)

[6] Li, Q. Peng h, Li J, Xia C, Yang R, sun L, Yu PS. *He L. A Survey On Text Classification: From Traditional To Deep Learning ACM Transactions On Intelligent Systems And Technology.* **13**, 311-351 (2022)

[7] Mehta, P. & And, C. and predictive performance of homogeneous ensemble feature selection in text classifica-tion International Journal of information retrieval research. (0)

[8] Anuratha, K. parvathy M. twitter sentient analysis using social spider Lex feature based synthetic senti rule recurrent neural network classification International Journal of uncertainty. *Fuzziness And Knowledge-based Systems.* **30**, 45-66 (2022)

[9] Wang, K. Zheng y, Fang s. *Liu S. Long Text Aspect Level Sentient Analysis Based On Text Filtering And Improved Bert Journal Of Computer Applications.* **40**, 2838-2844 (2020)

[10] Wang, C. & Wang, T. Yuan C Does applying deep learning in financial sentient analysis lead to better classification performance Economics bulletin. (0)

[11] Arya, V. & Agrawal, R. sentimental classification using feature selection techniques for text data composed of heterogeneous sources Recent advances in computer science and communications.

[12] Singh, S. & Sachan, M. Classification of code-mixed bilingual phonetic text using sentiment analysis. *International Journal On Semantic Web And Information Systems (IJSWIS).* **17**, 59-78 (2021)

[13] Capecchi, I. barbierato e. *Bernetti I. Analyzing TripAdvisor Reviews Of Wine Tours: An Approach Based On Text Mining And Sentient Analysis International Journal Of Wine Business Research.* **34**, 212-236 (2022)

[14] Han, W. Chen h, Gelbukh a, Zadeh a, Morency L, Poria S. *Bi Bimodal Modality Fusion For Correlation Controlled Multimodal Sentient Analysis.* **18** pp. 6-15 (2021)

[15] Kota, V. & Munisamy, S. High accuracy offering attention mechanisms based deep learning approach using CNN/bi-LSTM for sentiment analysis. *International Journal Of Intelligent Computing And Cybernetics.* **15**, 61-74 (2022)

[16] Liu, X., Shi, T., Zhou, G., Liu, M., Yin, Z., Yin, L. & Zheng, W. Emotion classification for short texts: an improved multi-label method. *Humanities And Social Sciences Communications.* **10**, 1-9 (2023)

[17] Hung, L. & Analysis:, A. review of recent trends in text based sentiment analysis and emotion detec-tion. *Journal Of Advanced Computational Intelligence And Intelligent Informatics.* **27**, 84-95 (2023)

[18] Chen, M. & Zhang, L. Application of edge computing combined with deep learning model in the dynamic evolution of network public opinion in emergencies. *The Journal Of Supercomputing.* **79**, 1526-1543 (2023)

[19] Umer, M., Imtiaz, Z., Ahmad, M., Nappi, M., Medaglia, C., Choi, G. & Mehmood, A. Impact of convolutional neural network and FastText embedding on text classification. *Multimedia Tools And Applications.* **82**, 5569-5585 (2023)

[20] Bashir, M., Javed, A., Arshad, M., Gadekallu, T., Shahzad, W. & Beg, M. Context-aware Emotion Detection from Low-resource Urdu Language Using Deep Neural Network. *ACM Transactions On Asian And Low-Resource Language Information Processing.* **22**, 1-30 (2023)

Edited by: Zhengyi Chai
Special issue on: Data-Driven Optimization Algorithms for Sustainable and Smart City
Received: Oct 23, 2023
Accepted: Mar 1, 2024

Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 3, pp. 2088–2102, DOI 10.12694/scpe.v25i3.2770

DESIGN AND APPLICATION OF AUTOMATIC ENGLISH TRANSLATION GRAMMAR ERROR DETECTION SYSTEM BASED ON BERT MACHINE VISION[*]

YU QING[†]

Abstract. Given the traditional handwritten English fonts, the accuracy of grammar error detection is unsatisfactory. This attribute leads to poor grammar error correction. Based on the optimized BERT machine vision model, an automatic English translation grammar error detection system is proposed in this paper. First, the basic architecture of the Transformer model and BERT model is considered, and a mixed attention module is discussed into the Transformer model to capture the features of the target in space and channel dimensions and realize the modeling of the context dependence of the target features. The feature maps are sampled by multiple parallel only if then cavity convolutions with different void rates to obtain multi-scale features and enhance local feature representation. Then, the input words of the BERT model are weighted by TFIDF to improve the feature extraction ability of the BERT model and construct the TF-BERT model. A database query rewriting model based on BERT and Transformer is proposed. The construction details of the model are described from the aspects of encoding processing, table embedding, and decoding processing respectively. Based on the principles of English translation, we extract grammatical features and build a grammar error detection method. TF-BERT model is selected as the basic framework. Combined with the hybrid attention mechanism, an automatic error correction model of English grammar is constructed. Finally, it is found that the loss value of the traditional system is as high as 0.7411, and the accuracy rate is 75%, while the loss value of the English grammar error detection system proposed in this paper is 0.2639, and the accuracy rate is 100%, which is 25% higher than that of the traditional system, and the performance is remarkable.

Key words: BERT; Mixed attention; Transformer model; Empty convolution; Grammar error correction;

1. Introduction.

The quality of English translation is mainly determined by the total number of incorrect texts such as grammatical errors and spelling errors. English translation usually pays more attention to the improvement of translation efficiency, the lightweight upgrade of products and the better user experience, and pays less attention to the quality control of translation [1]. As a result, there are some errors in the translation results. In this case, the reviewer generally needs to implement manual proofreading, which is not only time-consuming and laborious, but also often unable to eliminate all incorrect texts [2]. Therefore, an English translation robotic error detection machine is studied to realize automated proofreading of English translation and enhance the effectiveness and satisfaction of English translation.

At present, Liu Yakui et al. proposed a mechanical function extraction approach primarily based on laptop vision. First, a high-speed image acquisition system is used to collect the motion trajectory of the mechanism, and then Hough transform is used to automatically locate the key corner points in the video image [3]. Wu Yiquan et al. comprehensively reviewed PCB (Printed Circuit Board) defect detection algorithms primarily based on laptop imagination and prescient in the previous 10 years from the three dimensions of usual photo processing, normal computer gaining knowledge of and deep learning, and analyzed their blessings and negative aspects [4]. Zhou Qihong et al. used industrial cameras to capture images of yarn being sucked into the nozzle, improved gray enhancement methods to increase yarn features and background contrast, used the Canny operator for edge detection, and finally obtained yarn image features by dividing the upper and lower regions of interest and optimizing Hof line detection, and extracted the required position information by positioning algorithm [5]. Li Yong et al. solved the problem that the conventional measurement algorithm could not determine the measurement points corresponding to the irregular deformation contour and the region of the higher-order transition curve could not be measured, and realized the intensive measurement width in the axis direction to approximate the true edge width distribution to the greatest extent [6]. Ruan Jie et al.

[*]

[†]The school of general education, ChongQing Industry Polytechnic College, Chongqing 401120, China (yuqingedu@outlook.com)

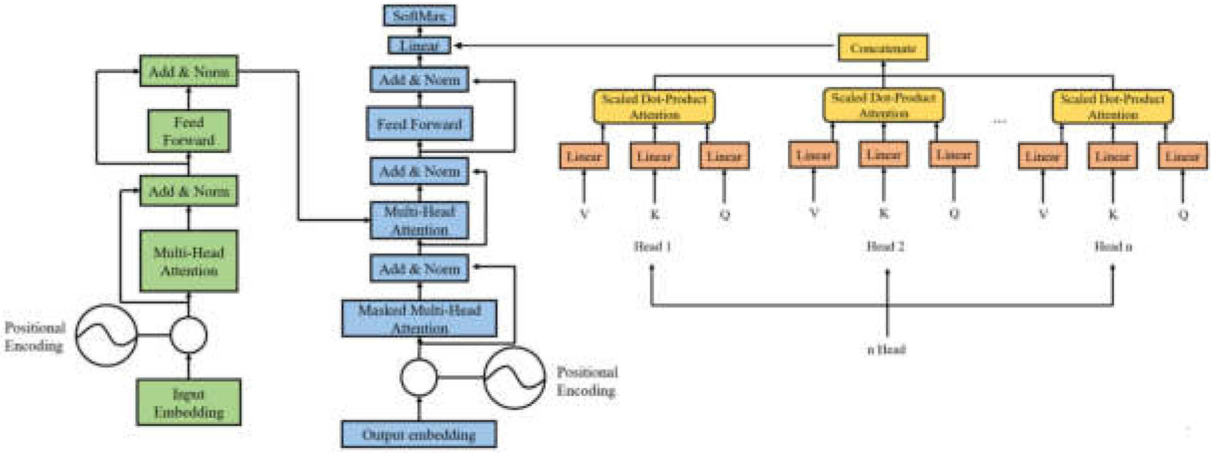


Fig. 2.1: Transformer network structure

converted the information of the template image from the image domain to the frequency domain based on a two-dimensional discrete cosine transform, selected the coefficient matrix represented by a low-frequency signal, converted the matrix into hash value through a hash algorithm, and adopted template matching method for each input frame image to track measurement points and extract pixel coordinates of measurement points [7].

Based on the optimized BERT machine vision model, an automatic English translation grammar error detection system is proposed in this paper. The basic architecture of the Transformer model and BERT (Bidirectional Encoder Representations from Transformers) model is explained, and a mixed attention module is introduced into the Transformer model to capture the features of the target in space and channel dimensions and realize the modeling of the context dependence of the target features. Then, the feature maps are sampled by multiple parallel cavity convolutions with different void rates to obtain multi-scale features and enhance local feature representation. Then, the input words of the BERT model are weighted by TFIDF to improve the feature extraction ability of the BERT model and construct the TF-BERT (Transformer Bidirectional Encoder Representations from Transformers) model. A database query rewriting model based on BERT and Transformer is proposed. The construction details of the model are described from the aspects of encoding processing, table embedding, and decoding processing respectively. Based on the principles of English translation, we extract grammatical features and build a grammar error detection method. The TF-BERT model is selected as the basic framework and combined with the mixed attention mechanism; the automatic error correction model of English grammar is constructed.

2. Optimized BERT model.

2.1. Transformer model. The combination of CNN (Convolutional Neural Network) and LSTM (Long Short-Term Memory) network and the Seq2Seq model has shown good results in many tasks in the field of natural language processing. The use of CNN can effectively extract local information, and the use of the Seq2Seq model based on the LSTM network can well process serialized data [8]. However, limited by the network architecture of LSTM itself, it is still difficult to deal with the problem of long-distance dependence, and the model cannot be parallelized, and the training time is long. To overcome these shortcomings of LSTM, Vaswani et al. 3 proposed a Transformer network model in 2017, which only adopts a self-attention mechanism to build, which not only reduces network parameters and computation amount but also improves parallel computing efficiency. Additionally, it performed better on the majority of tests in the area of natural language processing [9].

The Transformer network design is separated into encoder and decoder portions, just like the Seq2Seg architecture in Figure 2.1.

Among them, an Encoder or Decoder is composed of multiple encoder modules or multiple decoder modules independently stacked together layer by layer. The Encoder part is mainly responsible for extracting the relationship inside the input sequence and modeling the sequence, which can be regarded as the sequence feature extractor. The Decoder section can not only extract features but also build language models and sequence generation models. When the Transformer model receives input vectors. Firstly, the input vector is positionally encoded, relative position information and absolute position information are added, and then the position-encoded vector is input into the Encoder. After one layer of multi-attention layer, sub-residual connection and layer normalization are performed, and then enter into the subsequent layer of thoroughly related feedforward network, and then residual connection and layer normalization are carried out again, the output result of Encoder is obtained, and the result is output to Decoder [10]. In general, every small Encoder layer consists of a self-attentional computing layer and a completely related feedforward community layer for function mapping. The entry of the Decoder phase is additionally a vector encoded by using position, and then it goes via two consecutive multi-sensing layers, residual connection, and layer normalization operations, and then it is entered into the linked feedforward network. Finally, the last output is received through the processing of the linear layer. Decoder is comparable to Encoder; however, Decoder provides a masks multi-perception layer between the self-attention layer and the linked feedforward community layer, that is, it introduces a masks matrix with the values of the top triangle being terrible infinity and the values of the decrease triangle being 0, which serves to masks the future phrases in the sequence. To assist the mannequin, operate higher in the prediction phase.

In essence, the Transformer model is still a stack of multiple network layers, and the main reason why the Transformer model is so outstanding is the introduction of the following four key technologies.

1. Location coding
   LSTM, through its complex gating mechanism and cell state, is able to efficiently process position information in the sequence and transfer important information between different parts of the sequence. However, because Transformer uses a completely different self-attention mechanism to build its model, it has no location information itself. Therefore, it is necessary to introduce position coding to preprocess the input sequence and obtain the relative position and absolute position information of each element in the sequence [11]. The commonly used calculation method of position relation is as follows.

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{model}}\right) \tag{2.1}$$

$$PE_{(pos,2i+1)} = \cos\left(\mathrm{pos}/10000^{2i/d_{model}}\right) \tag{2.2}$$

2. Multi-head perception
   The multi-head perception structure maps Q, K and V in Attention linearly many times, and then scales the mapped matrix to get the value of self-attention. If the above operation is repeated n times, the calculation results of n heads can be obtained, and these calculations can be executed in parallel, where the Attention used is the operation of self-attention, and the calculation method of Multi-head attention is as follows:

$$\text{MultiHead\_Attention}(Q,K,V) = \text{Concat}\left(\text{head}_1,\ldots,\text{head}_i\right) W^o \tag{2.3}$$

3. Residual connection and normalization
   To avoid the problem of gradient disappearing due to the deepening of network layers during the training process, the Transformer model draws on the idea of a residual network and adds a residual module between every two adjacent Encoder layers. That is, the input part x and the output part of the multi-head sensing structure are added directly, and then a layer normalization processing module is connected. The whole process can be expressed in the following formula.

$$LayerOutput = LayerNorm(x + MultiHead\_Attention(x)) \tag{2.4}$$

Table 2.1: Model parameters table

model	N	dmodel	h	Total parameter
BERTbase	12	768	12	1.1
BERTlarge	24	1024	16	3.4

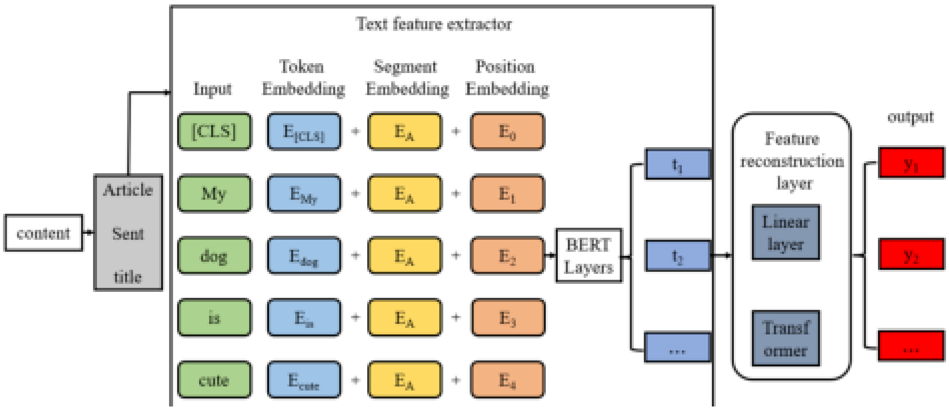


Fig. 2.2: Input structure of the BERT model

4. Position oriented fully connected feedforward network
   To spatially map the outcomes of various senses and improve the model's capacity to extract features, each layer of encoder and decoder, which incorporates a fully linked feedforward network, is applied to each place. The network is made up of a ReLU activation function and two linear transformations as shown below.

$$FFN(x) = max\left(0, W_1 x + b_1\right) W_2 + b_2 \tag{2.5}$$

2.2. BERT model. The basic framework of the BERT language model is essentially to construct a multi-layer bidirectional Encoder network using the Transformer coding part. In other words, the BERT model uses a dual Transformer to learn. Google released two versions of BERT model BERT$_{base}$ and BERT$_{large}$, and their feedforward size is set to four layers [12].

BERT$_{base}$ has 12 coding modules, i.e. N=12, each coding module has 12 headers from the attention operator module, h=12. The entry vector dimension is 768 which is d$_{model}$= 768. In BERT$_{large}$, N=24, h=16. d$_{model}$=1024. These differences bring the total number of parameters for BERT$_{base}$ to 110 million and BERT$_{large}$ to 340 million, as shown in Table 2.1.

There are also some differences between the BERT model and the Transformer model in the processing of input sequences. The BERT model can represent a single sentence or a pair of sentences in an input sequence when facing different tasks. For each Token, its input can be generated by adding segment embedding and positional embedding with text corresponding to each other. As shown in Figure 2.2 below.

To separate two sentences, indicating the end of the previous sentence and the beginning of the last sentence.

In addition to word embedding in the input sequence, BERT also adds position encoding like Transformer but does not use Transformer's sine and cosine encoding. Instead, the location information of each Token in the corpus is randomly initialized with the learned location embedding and then updated gradually with the training of the model. BERT model also adds clause embedding, each sentence uses a sentence entry, mainly to learn the relationship between multiple sentences and text matching pre-training, such as sentence relationship inference.

The purpose of pre-training is to train a model that can handle various downstream tasks as much as possible, and the parameters that need to be adjusted when using this model for transfer learning should be as few as possible [13]. Generally speaking, pre-training is to accumulate experience in various occasions and environments, so that the model can remember each word in a certain context and semantic usage so that the model can handle new tasks according to experience when carrying out transfer learning.

The fine-tuning of BERT includes sequential arbitrary and block tasks, in which there are subpairs and single categories in the column, the sentence beginning of which uses [CLSI], and only needs to take out the encoded vector for the downstream task [14]. The block task also includes the question answering task and sequence labeling task, which do not need to consider the output of [CLSI], but only need the vectorization result of word coding. The question-answering task can judge the position of the answer according to the coding vector, and the sequence-labeling task can predict the word labeling according to the coding vector.

2.3. Mixed attention design. Using self-attention, the location attention module may encode more general context information into local characteristics, improving its capacity to represent spatial information.

The spatial connection between any two feature pixels is modeled by the matrix.

$$T_{Sji} = \frac{\exp\left(T_{Bi} \cdot T_{Cj}\right)}{\sum_{i=1}^{N} \exp\left(T_{Bi} \cdot T_{Cj}\right)} \tag{2.6}$$

At the same time, input feature T$_A$ is input into the convolution layer to generate a new feature map

The convolution layer receives input feature TA at the same time, creating a new feature map T$_D$∈R$^{C\times H\times W}$ and reassembling it into R$_{C\times HW}$. The result is restructured to produce a tensor of dimension C×H×W by performing matrix multiplication between T$_D$ and the spatial attention matrix T$_S$.

Finally, add components to the original feature T$_A$ to produce the final output T$_P$∈R$^{C\times H\times W}$, which should mirror the finished representation of the distant backdrop. This is done by multiplying the result matrix of the aforementioned multiplication by the proportional parameter. The entire computation may be stated as follows:

$$T_{Pj} = \delta \sum_{i=1}^{N} \left(T_{Sji} T_{Di}\right) + T_{Aj} \tag{2.7}$$

The channel interest module makes use of self-attention to mannequin the interdependence between channels. By the usage of the interdependence between the function graphs of exceptional channels [15], the function illustration of unique semantics can be expanded to make the community center of attention on some channels with giant weight values.

Use the softmax layer to get the channel interest graph.

$$T_{Xji} = \frac{\exp\left(T_{Ui} \cdot T_{Vj}\right)}{\sum_{i=1}^{C} \exp\left(T_{Ui} \cdot T_{Vj}\right)} \tag{2.8}$$

In addition, T=A is reorganized again to get T=D∈R$^{C\times H\times W}$, performs matrix multiplication between T$_X$ and T$_W$, and reorganizes its result $^{RC\times HW}$ into R$^{C\times H\times W}$. The result is then multiplied by the scale parameter ε and the element summation is performed with T$_A$ to obtain the final output T$_Q$∈R$^{C\times H\times W}$, as shown below.

$$T_{Qj} = \varepsilon \sum_{i=1}^{C} \left(T_{Xji} T_{Wi}\right) + T_{Aj} \tag{2.9}$$

Formula (2.9) indicates that the remaining function of every channel is the weighted sum of the facets of all channels and the unique features, as a consequence setting up a long-term semantic dependency between function maps.

To further strengthen the feature extraction capability of the BERT model, this paper proposes to carry TF-IDP weight on the words after text preprocessing to represent the importance of different word vectors in

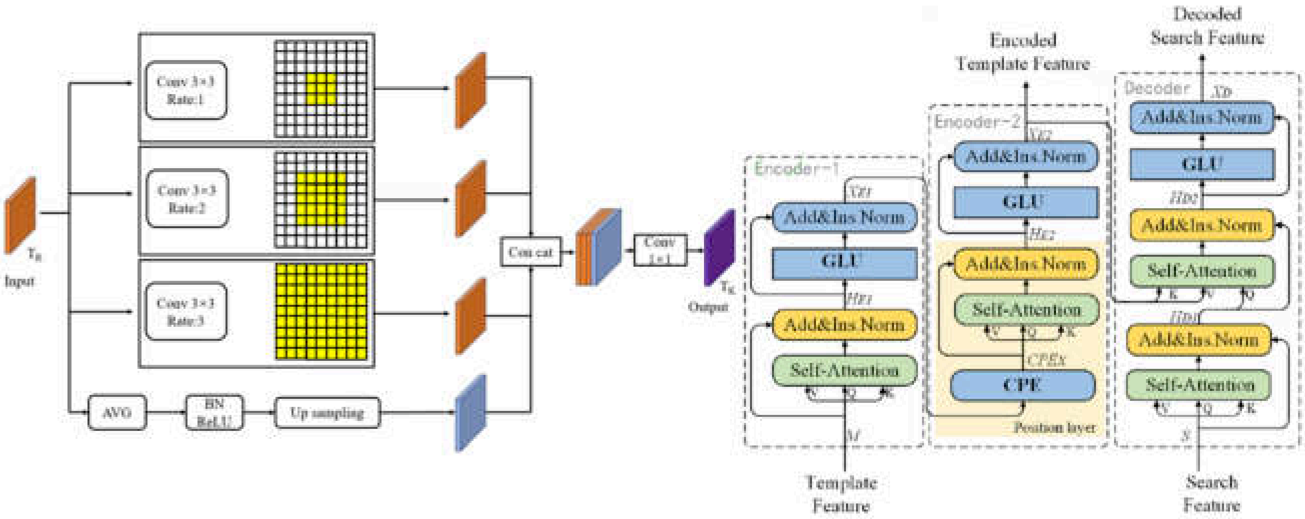


Fig. 2.3: CG-Transformer

the text, to improve the text classification effect. TF-IDF weighted BERT model is here shortened to TF-BERT [16].

TF-IDF is used to indicate the importance of a certain word in an article. If a certain word appears several times in one text and less frequently in other texts, it is considered that the word can be used to distinguish and classify the text. TF is word frequency If a word t appears 5 times in a certain article, the TF value of that word in that article is 5. IDF is the inverse document frequency, its value is determined by the number of occurrences of a certain word in the total document, the fewer the number of articles containing the word t in the total document, that is, the less n, the larger the IDF, the more it indicates that the word t can represent the article. The IDF calculation formula of the word t is shown in 2.9.

$$IDF(t) = \log \frac{N}{n_t}. \tag{2.10}$$

The calculation formula of TF-IDF is shown in 10.

$$K(t, D_i) = \frac{TF(t, D_i) \times IDF(t)}{\sqrt{\sum [TF(t, D_i) \times IDF(t)]^2}} \tag{2.11}$$

Figure 2.3 shows the CG-Transformer structure. This part consists of two layers of encoders and one layer of decoder. To decrease the computational load of the Transformer structure, the encoder is composed of a single-head self-attention module, gated linear unit, and convolutional role coding module. The decoder consists of a single-head self-attention module and a gated linear unit.

The use of self-attention performs a key function in the whole Transformer structure. First, the encoder and decoder enter template characteristic M and search place characteristic S respectively, and then the matrix seriously changes the enter points to achieve question Q, key K, and fee V as the enter of the self-attention section [17]. Then dot product is used to calculate the similarity between Q and K, and a set of interest weights is obtained after passing Softmax. Finally, the weight is dotted with V to obtain a vector with attention weight. The specific calculation process is shown in the following equation.

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.12}$$

3. Design of syntax error correction system based on TF-BERT model.

3.1. Query rewriting model construction based on transformer. Compared with English database query rewriting, the difficulty of Chinese database query rewriting task lies in the word segmentation of the query input sequence. In English natural language processing tasks, word segmentation can be divided according to Spaces. After segmentation, a sequence of sentences is processed into several words with clear meaning. So the result of word segmentation is more accurate. For Chinese tasks, due to Chinese writing habits, punctuation marks are usually used at the end of a sentence to distinguish the text. There is no clear separator between each word, which will inevitably lead to ambiguity or wrong distinction in the process of word segmentation [18]. Taking a piece of data in the TableQA data set as an example, when the sentence ”the highest second-hand real estate price” is divided into three words, ”second-hand”, ”real estate” and ”price”, and ”second-hand real estate price” is a column name in the database table, should be used as a proper term as a whole and not be divided. To reduce the influence of Chinese word segmentation on the model effect. You can use a third-party thesaurus such as jieba HanNLP to further ”load” a custom database column noun library on top of its excellent segmentation capabilities. To obtain the specific lexicon, we need to write a script to collect the column name information of all tables in the database and set the weight of the word according to the data amount corresponding to the column name. The larger the amount of data involved, the higher the word weight corresponding to the column name, and the less easily it is split during word segmentation, to obtain a more accurate encoding. However, from the experimental results, the improvement of the rewriting effect of the third-party lexical segmentation is not obvious.

To solve the problem caused by Chinese word segmentation, BERT’s approach is to use character-level encoding to divide each sentence into word-level granularity. Then it relies on its super-large training corpus to make the relationship between words be represented in the first few layers of parameters of BERT. From this perspective, the task of encoding with BERT does not require additional word segmentation processing. In this way, even if the database column names and other proper nouns are divided, it will not affect the encoding effect. For the English words that may appear in the input, BERT divides the words into more detailed word fragments by the Subword Tokenization method, and the encoding method is more flexible and effective. The advantage of this is that it can solve the problem of unknown words in traditional coding [19].

Unlike textual data, tabular data is distributed in a two-dimensional (2-D) structure. The encoding method requires first converting 2-D table data into linearized 1-D sequence inputs, and then feeding the table data into the downstream language model. The method adopted in this paper is to extract the database table name and column name for each query data, and then flatten these data and concatenate them. In the process of concatenation, the database table name is fixed as the starting information, and all column names are concatenated successively. Wrap, as in formula 3.1.

$$Input_E = [CLS, T, SEP, CLS, C_i, SEP, ..., CLS, C_n, SEP] \tag{3.1}$$

For the encoder to combine table information and query information at the same time, to capture the corresponding information between the table and query, the two parts need to be encoded at the same time. So this article concatenates the query input sequence and the table input sequence, and the final input for the encoder will look like this.

$Input_E = [Input_t, Input_q]$ (3.2)

The standard Transformer decoder structure is also used as the skeleton for the decoder.

After the calculation of 6 layers of stacked attention submodules, the output of the encoder is obtained and then input to the decoder. The decoder will combine the attention calculation results of real SQL and the encoder’s hidden state to obtain multiple calculation results with different probabilities at each position of the result sequence. To obtain the optimal SOL result, it is necessary to combine the bunched search algorithm. Specifically, the algorithm builds a search tree from the starting point of the sequence. Based on the breadth-first strategy, the nodes in the first few probabilities (determined by the width of the cluster) of each layer of the tree are selected as the parent nodes of the next layer of search, and the optimal decoding path is finally selected [20].

During the training phase, the other input to the encoder is a real SQL sequence. In the education phase, the actual label is used as the entry of the subsequent kingdom in the decoding process, that is, it performs the position of Teacher Forcing. The advantages of this method: 1. Prevent the incorrect prediction of the preceding kingdom from inflicting all the subsequent countries to bias the incorrect decoding result, right the prediction of the model, and keep away from similar amplification of blunders in the system of sequence generation. two It radically speeds up the convergence pace of the mannequin and makes the coaching manner of the mannequin extra fast and stable. In this chapter, the BERT pre-training mannequin is used as the enter phrase embedding module at the encoder end. Therefore, to make sure that the exceptional inputs of the mannequin map to the equal phrase vector house and keep away from unknown phrase errors, the equal embedding processing wishes to be carried out on the entry of the decoder side, that is, the enter processing is carried out on the actual SQL first.

$$Input_D = [CLS, SQL, SEP] \tag{3.2}$$

3.2. Design of automatic error correction system.. Aiming at English text content, this paper designs a text feature extraction method based on English translation principles, and then compares and analyzes it with typical English parallel corpora to form a grammar error detection method. Considering that the mutual translation process between the source language and the target language is the same, the word vector parameters between the two can remain uniform [21]. In this paper, the encoder and decoder are used to construct the integrated translation structure. In the actual translation process, the probability calculation formula of the integrated translation bar is.

$$P(BA:\theta) = P\left(\eta v, \eta_1, \eta_2, \cdots, \eta_{m-1} : \theta\right) \tag{3.3}$$

The translation network built based on the encoder can be represented by the mathematical formula.

$$\begin{aligned} &\varpi_\tau = sigmoid\left(H_1 e + \lambda_1 + H_2 s_{\tau-1} + L_1\right) \\ &\sigma_\tau = sigmoid\left(H_1 e + \lambda_1 + H_2 s_{\tau-1} + L_2\right) \\ &\xi_\tau = tans\left(H_1 e + \lambda_1 + H_2 s_{\tau-1} + I\right) \\ &s_\tau = \left(1-\sigma_\tau\right)\xi_\tau + \sigma_\tau \xi_{\tau-1} \end{aligned} \tag{3.4}$$

Using the above formula to constrain the coding process will lead to errors in English grammatical translation results. To resolve this issue, English grammatical features are extracted using a neural network, and the outputs of the feature extraction are then normalized using the softmax function. Create an algorithm that can automatically detect grammar mistakes. The label for a syntax mistake can be represented as follows in the feature space of results for English translation:

$$F = \arg\max \sum_{r=1}^{Z} 1_\psi \tag{3.5}$$

According to the result of English grammar judgment, the probability of English grammar error is calculated.

$$Q = \frac{\sum_{r=1}^{Z} 1_{\beta(r.0)=1}}{E} \tag{3.6}$$

In the TF-BERT model, the Encoder usually encodes all the data of the supply sentence into a fixed-length context vector c, and then the vector c is steady at some point in the decoding method of the Decoder. With Attention, a mechanism for focusing the model's Attention on the currently translated word, the input to the Decoder is no longer a fixed context vector. Instead, the current context vector is calculated based on the currently translated information [22]. Soft Attention considers all inputs but instead of giving equal weight to each input, it pays more attention to certain inputs. The mixed Attention mechanism, Soft Attention, assigns an attention weight, a probability distribution, to each feature. The context vector of its specific region information can be obtained directly by gravity-weighted summation.

To enhance the automatic error correction effect of the system, the horizontal normalization method is used to process the training samples of the grammar automatic error correction model. In simple terms, it is

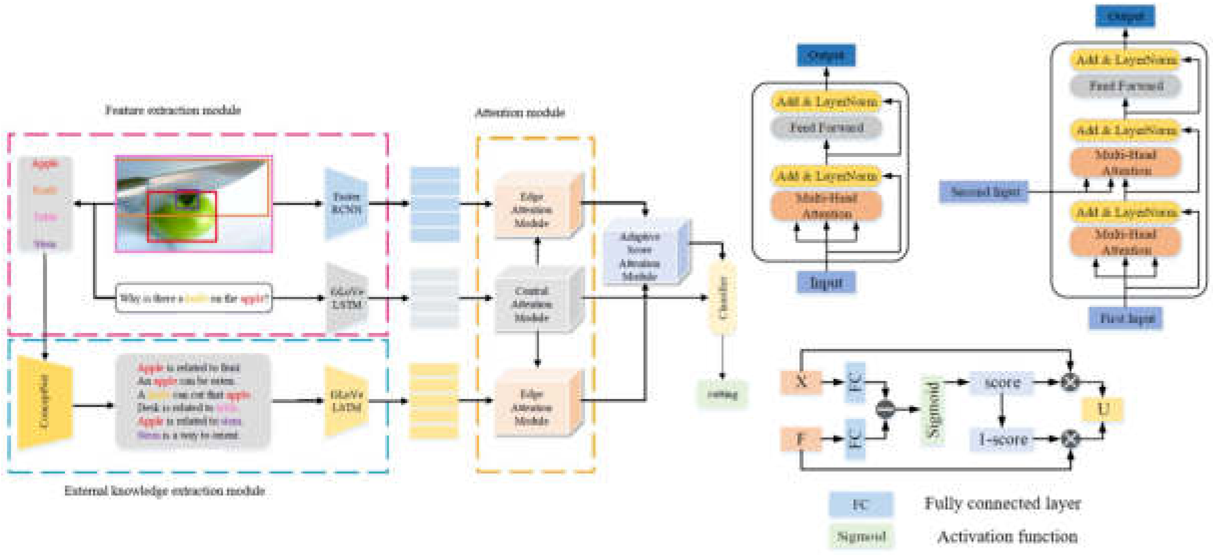


Fig. 3.1: Internal structure based on machine vision

to calculate the input variance of hidden layer neurons, map all neurons into the same space, and calculate the normalized parameters for all hidden layer neurons.

$$\chi = \frac{1}{D}\sum_{u=1}^{D} p_{\mathrm{u}} \tag{3.7}$$

$$\iota = \sqrt{\frac{1}{D}\sum_{u=1}^{D} (p_u - \chi)^2} \tag{3.8}$$

In the formula, χ and ı characterize normalized parameters, D represents the range of hidden layer neurons, and u represents hidden layer neurons. The normalized training samples are input into the automatic grammar correction model, and the result of English grammar correction is obtained.

3.3. Attention module design. Specifically, the basic attention unit consists of two major attention modules, namely, the center attention module and the edge attention module. In Figure 3.1, the problem features are first processed by the GloVe word embeddedness and the long/short memory network is obtained through the central attention module. Then, the visual features from the image and the text features from the problem are processed through the edge attention module at the top [23]. In addition, external knowledge is queried through Concept Net, a large public knowledge graph, and the processed knowledge text features are similarly obtained through the bottom-edge attention module. Visual and textual features are given balanced attention through the impact of the adaptive fractional attention module on the problem. Balanced knowledge features, picture features, and semantic features are fed into the classifier to obtain the final answer.

In this paper, the dot product operation between the question and the key is transformed to the product of matrix Q and matrix K, then divided through the scale factor, which is the rectangular root price of dk. The preceding output is then handed to the softmax characteristic and the interest weight is received by multiplying via the matrix V.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.9}$$

Each layer of the central attention module is stacked by the basic reduced dot-product attention units of Layer G. The problem characteristic formula it inputs is as follows.

$$Y = [h_1, h_2, \ldots, h_N] \, R^{n \times N} \tag{3.10}$$

This paper combines the attention matrix with the problem features to generate the problem feature Y^g of self-attention, as shown in equation 23. In addition, the problem characteristic of self-attention, Y^g, is also seen as an input to the next basic narrowing dot-product attention unit. This process can be expressed by the formula 3.12 as.

$$Y^g = CAM^g \left(Y^{g-1}\right) \tag{3.11}$$

The side interest module consists of a photo facet interest module and a know-how facet interest module. The photograph part interest module and expertise area interest module have an equal mannequin structure, and every mannequin shape includes G stacked layers, which are composed of two fundamental decreased dot-product interest units. Taking the photograph part interest module as an example, the entry of the picture area interest module comes from the special visible function X cited earlier, and the last output of the hassle characteristic Y^G from the central interest module. The first simple scaling dot-product interest unit in every layer of the photo area interest module, which has the same feature as the single-layer central interest module, inputs X's self-attention for calculating visible points [24].

The 2D scaled dot product interest is linked using the softmax feature to research the interest weight matrix. Thus, the interest matrix is utilized from the final layer of the central interest module to the hassle function of concern, Y^G, to output the problem-based photo feature, X^g, which is the entry to the subsequent layer of scaled dot-product attention. The technique is as follows:

$$X^g = IEAM^g \left(X^{g-1}, Y^G\right) \tag{3.12}$$

Where, the preliminary enter X^o=X, Y^G is the output of the last layer of the central attention module. The output of the final G layer of the image edge attention module is X^G, which represents the picture facets processed via the interest module.

4. Experimental analysis. To verify the effectiveness of this system in detecting and recognizing different combinations of handwritten English grammar errors, the experiment will select the hardware and software parts of the system and set the parameters. Among them, the MU3E200M/C camera with 2 million pixels and a frame rate of 60fps was selected to capture handwritten English character images. All experiments were conducted on a CPU/GPU compute node configured with a xx GHz Intel-Xeon Skylake processor and yy GB RAM. The lens model selected VS-2518VM, equipped with a C-port of 2 million pixels to match the camera. The light source is blue light, bar light source: the sensor is E3CVS1G, and the detection distance is set to 10 mm.

The English version of Bert-base-uncased is selected as the pretraining model. The encoder of the mannequin consists of eight layers, 768 hidden units, and 12 interest heads. The complete reference quantity is a hundred and ten M. Set the most sentence size of this mannequin to 128, batch measurement to 32, optimization characteristic to Adam, getting to know fee to 0.00001, dropout layer parameter to 0.5. The experimental facts had been amassed through the usage of the CoLA dataset and annotated NUCLE and FCE gaining knowledge of corpora.

To improve the quality of data collection, it is proposed that the data should be cleaned before model training. The specific cleaning steps mainly include reprocessing, eliminating empty lines, special symbol processing, length control, text segmentation and numerical operation, and full half-angle conversion.

To deepen the understanding of each model, the hyperparameters of the above models are listed one by one. The L2 regularization parameter of each model is set to 0.00005, and the activation function of the output layer adopts the softmax function. The hyperparameters of each model are shown in Table 4.1 below.

Using the final data set obtained in this paper, the above models are trained and tested respectively, and the loss curves of each model are obtained.

Table 4.1: Hyperparameters of each model

Hyper-parametric Model	Learning rate	Batch-size	Training rounds	Fill in	Optimi-zation function	Network layer number	Loss function	Hidden layer activation function
Model 1	0.00001	8	200	Y	SGD	21	Cross entropy	ReLU
Model 2	0.00001	8	200	Y	Adam	33	Cross entropy	Leaky ReLU
Model 3	0.00001	8	200	Y	Adam	32	Cross entropy	Leaky ReLU
Model 4	0.00001	8	200	Y	Adam	31	Cross entropy	Leaky ReLU
Model 5	0.00001	8	200	Y	Adam	31	Cross entropy	Leaky ReLU
Model 6	0.00001	8	200	Y	Adam	24	Cross entropy	Leaky ReLU
Model 7	0.00001	8	200	Y	Adam	25	Cross entropy	Leaky ReLU
Model 8	0.00001	8	200	Y	Adam	26	Cross entropy	Leaky ReLU
Model 9	0.00001	8	200	Y	Adam	27	Cross entropy	Leaky ReLU
Model 10	0.00001	8	200	Y	Adam	18	Cross entropy	Leaky ReLU
Model 11	0.00001	8	200	Y	Adam	19	Cross entropy	Leaky ReLU

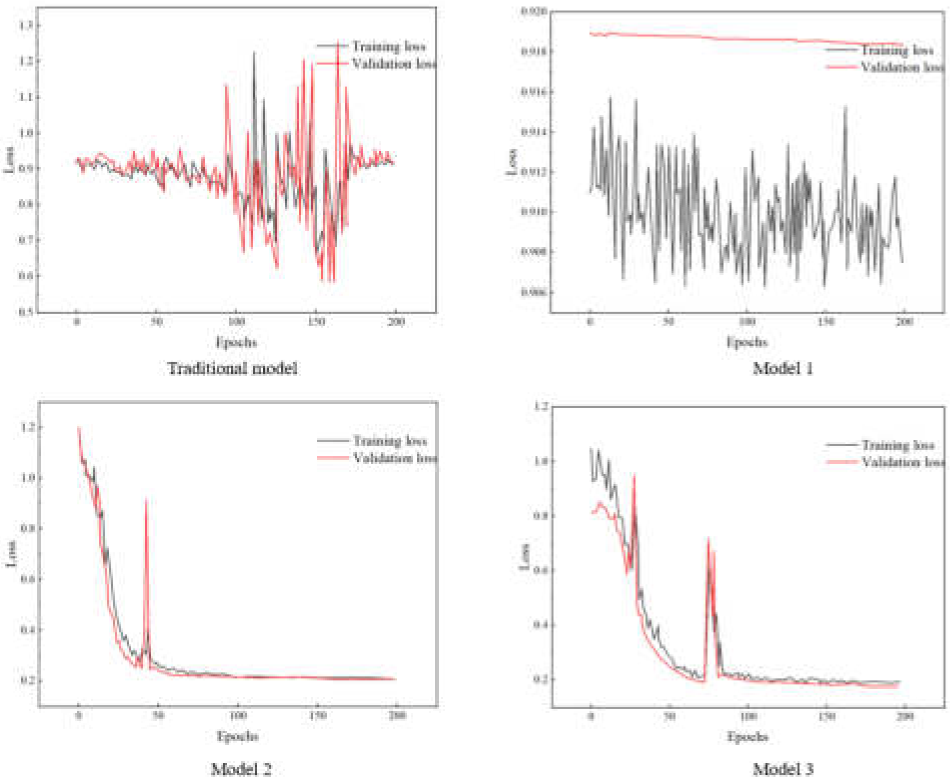
It can be considered from Figure 4.2 that the usual mannequin has apparent oscillations of loss price and accuracy on the coaching set and verification set in this paper, which is no longer steady enough. The loss cost on the take-a-look-at set is 0.7411 with an accuracy of 81.25%. The performance of the traditional model on the data set in this paper is not ideal, and it needs to be further optimized to achieve effective recognition and classification.

As can be seen from Figure 4.1, after reducing the learning rate, the loss value and accuracy of model 1 on the training set still oscillates significantly, while the loss on the verification set gradually decreases, but the change is slow. The loss value on the test set is 0.9157, the classification accuracy is only 43.75%, and the recognition and classification cannot be realized.

Model 2 is the embodiment of the pumpability identification method of mortar based on the 3DCNN+ConvLSTM2D network structure proposed in this paper. It has 3 layers of ConvLSTM2D. Model 2 has optimized and improved model 1, mainly in the following aspects:

1. the optimization function is adjusted to Adam.
2. Adjust the activation function of the hidden layer from ReLU to Leaky ReLU.
3. The networks behind the ConvLSTM2D layer of the third layer are adjusted to :1 GlobalAveragePooling3D layer, 1 Dropout layer, 2 Dense layers, 1 LeakyReLU layer, and 1 Activation layer, respectively. That is, the Global Averagepooling3d-dense-dropout Dense-Activation network structure is used.
4. The learning rate is reduced from 0.005 to 0.00001. The loss of model 3 gradually decreases with the increase of epochs and converges to around 0.17. The loss value of model 3 on the test set is 0.6077, all samples are correctly identified, and the classification accuracy is up to 100%.

As shown in Figure 4.2, Model 4 continues to reduce one ConvLSTM2D layer based on model 3, that is, model 4 has a total of one ConvLSTM2D layer. In addition, the network structure and hyperparameters are consistent with model 3. As can be seen from Figure 4.2, the loss of model 4 gradually decreases with the increase of epochs and converges to around 0.10. The loss value of Model 4 on the test set is 0.4691, and all



Traditional model

Model 1

Model 2

Model 3

Fig. 4.1: Comparison between traditional system and models 1, 2, and 3

samples are correctly identified.

Model 5 tries to improve the learning rate from 0.00001 to 0.0001 based on Model 4. In addition, the network structure and hyperparameters are consistent with model 4. The loss value and accuracy of model 5 on the training set and verification set are oscillated, and cannot converge. All samples, regardless of Pumpability, were identified as Poor Pumpability and could not be effectively classified. The loss value on the test set was 0.8042, and the accuracy rate was only 56.25%.

The loss of model 6 gradually decreases with the increase of epochs and converges around 0.07. The loss value of model 6 on the test set is 0.4153, and the classification accuracy is as high as 93.75%. Compared with model 4, model 6 reduces 7 layers of network, the number of parameters is increased from 1,891,874 to 1,949218, the number of parameters is increased, although the recognition accuracy is slightly decreased, the loss value is reduced, and the convergence speed is greatly improved.

The loss value of model 7 on the test set is 0.3768, and the classification accuracy is as high as 93.75%. Compared with model 6, the learning rate of model 7 is still 0.00001, the number of network layers is increased by one ConvLSTM2D layer, the number of parameters is increased from 1949218 to 2244386, the accuracy rate remains unchanged at 93.75%, the convergence speed is barely improved but the loss value is decreased by 9.30%.

As can be seen from Figure 4.3, the loss of mannequin nine step by step decreases with the amplification of epochs and converges to round 0.05. The loss price of mannequin nine on the check set is 0.4646, and the classification accuracy is up to 93.75%. Compared with mannequin 6, the mastering fee of mannequin nine is nevertheless 0.00001, the range of community layers is decreased by using 7 layers, and the quantity of parameters is expanded from 1949218 to 66725410. The wide variety of parameters is extensively increased, and the accuracy stays unchanged at 93.75%. Although the accuracy on the take-a-look-at set is no longer

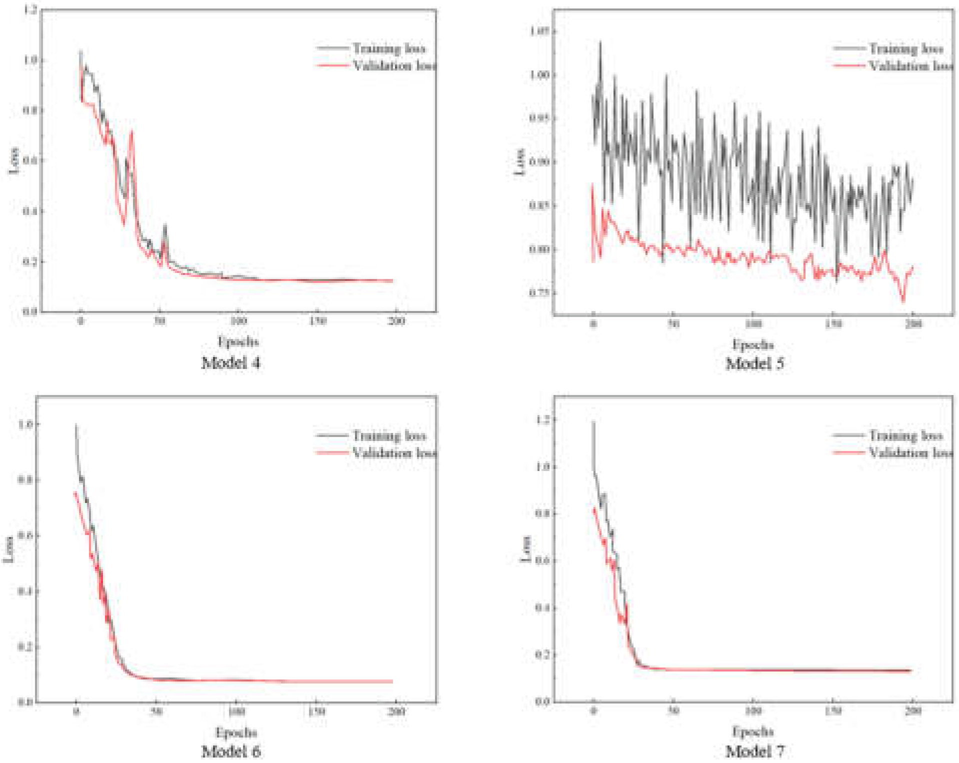


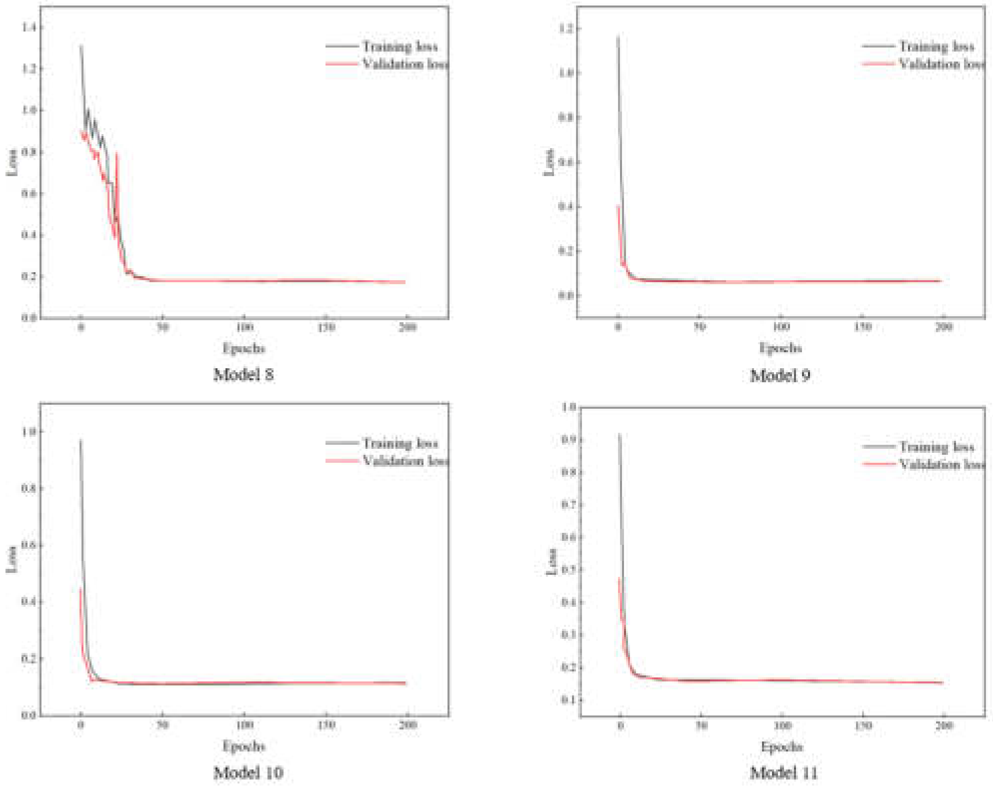
Fig. 4.2: Results of models 4, 5, 6, 7

improved, and the loss cost is barely increased, the convergence velocity is notably accelerated.

Model 11 attempts to add another layer of ConvLSTM2D layer based on model 10, that is, model 11 has 3 layers of ConvLSTM2D layers, in addition, the network structure and hyperparameters are consistent with model 10. In Figure 4.3, the loss of model 11 gradually decreases with the increase of epochs and converges to around 0.15. Model 11 has a loss value of 0.2639 on the test set, and the classification accuracy is up to 100%. Compared with model 10, the learning rate of model 11 is still 0.00001, the number of network layers is increased by one layer, the number of parameters is increased from 67020578 to 67315746, and the accuracy rate is up to 100%. The accuracy of model 11 on the test set is still as high as 100%, and it can realize the correct identification and classification of all samples on the test set. The convergence speed is almost as fast as that of Model 10, and the loss value is reduced by 25.69% compared with Model 10.

5. Conclusion. To detect English translation errors intelligently and efficiently, this paper designs an automatic English translation grammar error detection system based on the optimized BERT machine vision model. A hybrid attention module is introduced into the Transformer model to capture target features in both spatial and channel dimensions and to model the context dependency of target features. Then, the feature maps are sampled by multiple parallel cavity convolutions with different void rates to obtain multi-scale features and enhance local feature representation. Then the input words of the BERT model are weighted by TFIDF to increase the feature extraction capability of the BERT model and construct the TF-BERT model. Specific conclusions are as follows.

1. A visual English automatic translation grammar error detection system based on a mixed attention Transformer is designed. Hybrid attention is embedded in the middle layer of the backbone web to extract more robust features. Image multi-scale feature extraction is realized by using multiple cavity convolution with a smaller cavity rate. The Transformer codec is used to transfer information between



Model 8

Model 9

Model 10

Model 11

Fig. 4.3: Results of models 8, 9, 10 and 11

the template branch and search branch features of the twin network, which improves the accuracy of grammar-translation. The 0-filled convolution method is used to realize more flexible and effective position coding, improve the capability of the web to distinguish similarity, and further increase the efficiency of the automatic translation syntax error detection system.

2. Experiments show that the loss value of the traditional system is as high as 0.7411, the accuracy rate is 81.25%, and the accuracy rate is 75%. The loss of the TF-BERT model increase in this paper decreases with the increase of epochs and converges to about 0.15, with a loss value of 0.2639. When the model in this paper is case 11, the classification accuracy is as high as 100%. The learning rate is still 0.00001, the number of network layers is increased by 1 layer, the number of parameters is increased from 67020578 to 67315746, and the accuracy rate is up to 100%.

First, in the warm-up phase, the Transformmer model has become stable, and there is not much room for optimization at the beginning of confrontation training. Second, the construction of the policy gradient may be unstable, resulting in the generator cannot effectively update parameters according to the results of the discriminator. Therefore, it is necessary to explore more reasonable strategy gradient calculation methods in the future, and select appropriate regular constraints to control the amplitude of each gradient update within a reasonable range, so as to obtain better optimization results.

REFERENCES

[1] Golnabi, H. & Asadpour, A. Design and application of industrial machine vision systems[J]. *Robotics And Computer-Integrated Manufacturing.* **23**, 630-637 (2007)

[2] Robie, A., Seagraves, K., Egnor, S. & Others Machine vision methods for analyzing social interactions[J]. *Journal Of Exper-*

imental Biology. **220**, 25-34 (2017)

[3] Yakui, L., Hongyun, L., Tianzi, L. & Others Research on Mechanical Characteristics Measurement Method of High Voltage Circuit Breaker Based on Machine Vision [J]. *Transactions Of China Electrotechnical Society.* **202432** pp. 35432-20235

[4] Yiquan, W., Lanyue, Z., Yubin, Y. & Others Research status and prospect of PCB defect detection algorithm based on machine vision [J]. *Chinese Journal Of Scientific Instrument.* **43**, 1-17 (2022)

[5] Qihong, Z., Yi, P., Junhao, C. & Others Yarn break location method for yarn joint Robot based on Machine vision [J]. *Acta Textile Sinica.* **43**, 163-169 (2022)

[6] Yong, L., Peijian, S., Qijian, O. & Others Measurement method of irregular shape ultra-thin heat pipe width based on machine vision [J]. *Journal Of South China University Of Technology (Natural Science Edition).* **50** pp. 4 (2022)

[7] Jie, R., Peng, L., Yitong, X. & Others Measurement Method of dynamic lift of hanging string based on Machine vision [J]. (China Railway Science,2022)

[8] Chen, Y., Chao, K. & Kim, M. Machine vision technology for agricultural applications[J]. *Computers And Electronics In Agriculture.* **36**, 173-191 (2002)

[9] Ren, Z., Fang, F., Yan, N. & Others State of the art in defect detection based on machine vision[J]. *International Journal Of Precision Engineering And Manufacturing-Green Technology.* **9**, 661-691 (2022)

[10] Oren, M. & Nayar, S. Generalization of the Lambertian model and implications for machine vision[J]. *International Journal Of Computer Vision.* **14** pp. 227-251 (1995)

[11] Sonka, M., Hlavac, V. & Boyle, R. Image processing, analysis, and machine vision[M]. (Cengage Learning,2014)

[12] Devlin, J., Chang, M., Lee, K. & Others Bert: Pre-training of deep bidirectional transformers for language understanding[J]. (2018), arXiv preprint

[13] Dickmanns, E. & Graefe, V. Dynamic monocular machine vision[J]. *Machine Vision And Applications.* **1**, 223-240 (1988)

[14] Marcel, S. & Rodriguez, Y. Torchvision the machine-vision package of torch[C]//Proceedings of the 18th ACM international conference on Multimedia. *1485-1488.* (2010)

[15] Vision, V. Automated visual inspection and robot vision[M]. (Prentice-Hall,1991)

[16] Penumuru, D., Muthuswamy, S. & Karumbu, P. Identification and classification of materials using machine vision and machine learning in the context of industry 4.0[J]. *Journal Of Intelligent Manufacturing.* **31**, 1229-1241 (2020)

[17] Rehman, T., Mahmud, M., Chang, Y. & Others Current and future applications of statistical machine learning algorithms for agricultural machine vision systems[J]. *Computers And Electronics In Agriculture.* **156** pp. 585-605 (2019)

[18] Kataoka, T., Kaneko, T., Okamoto, H. & Others Crop growth estimation system using machine vision[C]//Proceedings 2003 IEEE/ASME international conference on advanced intelligent mechatronics (AIM 2003). *IEEE.* **2** (2003)

[19] Lenz, R. & Tsai, R. Techniques for calibration of the scale factor and image center for high accuracy 3-D machine vision metrology[J]. *IEEE Transactions On Pattern Analysis And Machine Intelligence.* **10**, 713-720 (1988)

[20] Transforms, P. Properties and machine vision applications[J]. *CVGIP: Graphical Models And Image Processing.* **54**, 56-74 (1992)

[21] Ranft, B. & Stiller, C. The role of machine vision for intelligent vehicles[J]. *IEEE Transactions On Intelligent Vehicles.* **1**, 8-19 (2016)

[22] Kopparapu, S. Lighting design for machine vision application[J]. *Image And Vision Computing.* **24**, 720-726 (2006)

[23] Dickmanns, E. & Graefe, V. Applications of dynamic monocular machine vision[J]. *Machine Vision And Applications.* **1**, 241-261 (1988)

[24] Wildes, R. & Asmuth, J. Green G L, et al. A machine-vision system for iris recognition[J]. *Machine Vision And Applications.* **9** pp. 1-8 (1996)

Edited by: Zhengyi Chai
Special issue on: Data-Driven Optimization Algorithms for Sustainable and Smart City
Received: Oct 31, 2023
Accepted: Mar 25, 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 $\LaTeX 2_\varepsilon$ 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.